



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.



THE UNIVERSITY *of* EDINBURGH

AN INVESTIGATION OF SELECTIVE CONSTRAINTS IN COLORECTAL CANCER

By

Nathalie Anne Feeley

Thesis Presented for the Degree of Doctor of Philosophy

The University of Edinburgh

2024

Declaration

I declare that this thesis has been composed solely by myself. Any contribution by another person has been explicitly referenced and acknowledged. This work is entirely my own and has not been submitted in any previous degree application.

Signed,

Nathalie Anne Feeley

Abstract

Next generation sequencing (NGS) has significantly advanced our understanding of colorectal cancer (CRC) and highlighted the inaccuracies of the traditional, linear model of adenoma-carcinoma evolution. NGS has provided evidence for intra-tumour heterogeneity (ITH) and colorectal carcinomas that lack a full complement of driver mutations. Alternatively, the branched model of adenoma-carcinoma evolution proposes that genetically differing clones can coexist and evolve in spatially discrete areas. Further refinements such as the just-right model and mini-driver models are encompassed by the branched model, providing additional insights.

By examining mutations in multi-regional, colorectal adenomatous polyp samples, I have provided supporting evidence for the branched model of CRC evolution. Following an assessment of the presence of parallel evolution in adenomas, I looked specifically at mutations in Wnt pathway genes to determine whether there is evidence for a “just-right” level of pathway activation by mutations. Canonically, the Wnt pathway is central to epithelial cell growth and regulation in the intestine. When mutationally activated, this pathway can lead to uncontrolled neoplastic initiation and progression. Previous evidence from familial adenomatous polyposis (FAP) suggests that a just-right number of *APC* 20-amino acid repeats results in an optimal level of Wnt pathway activation. The applicability of this model to the entire Wnt pathway has been suggested. The mini-driver model proposes that low-frequency mutations of minor selective advantage, compared with major-driver mutations, can finely adjust Wnt activation levels to provide a just-right level of Wnt signalling. Using a large, CRC dataset (100KGP), I explored mutations in both major-driver and mini-driver genes, seeking patterns of mutual exclusivity and co-occurrence. The results supported the just-right model, revealing mutual exclusivity between mutations in major-driver genes and co-occurrence between mutations in major-driver and mini-driver genes. These results suggested that combinations of Wnt pathway mutations are selected to provide a just-right amount of Wnt signalling for a tumour to evolve. In addition, this study also found potential discrepancies between particular combinations of mutations in different tumour genomic instability subtypes (MSS/MSI) and in different anatomical locations, which was suggestive of differing just-right signalling levels depending on the genomic instability/location context of the tumour. Furthermore, the proposed role of *TCF7L2* as a mini-driver, and CRC predisposition gene, was investigated through genome-wide association studies (GWAS) and mouse model studies, showing some supportive results.

This research contributed valuable insights into the complex dynamics of CRC evolution, shedding light on the role of the just-right model involving the Wnt pathway and supporting the branched model of adenoma-carcinoma evolution.

Lay Summary

The traditional linear model of colorectal cancer (CRC) evolution suggests that when a cell experiences a change in DNA (mutation), this can potentially provide this cell with a growth and survival advantage. This cell then replicates. If another DNA change happens in one of these cells and provides it with an additional growth/survival advantage, this cell then replicates in turn and becomes the dominant “clone” (group of cells from a single ancestor) within the tumour (an abnormal growth of cells). This process can be repeated with accumulation of a number of DNA changes. Next generation sequencing, able to show DNA changes within tumours, has revealed that this linear model of evolution is not always accurate. Instead, different clones can exist in a tumour and evolve alongside one another. I have shown this by examining the genes mutated in multiple samples from a large set of colorectal adenomatous polyps (CRC precursors). I then specifically look at whether there is a “just-right” level of activation due to mutations in a specific cell signalling pathway named the Wnt pathway. Usually, this pathway is central to regulating the turnover of inner surface cells in the intestine, but if mutated can cause the uncontrolled growth and replication of cells. I used a large CRC dataset to look at mutations within “driver genes” (that promote tumour growth when mutated) in the Wnt cell signalling pathway. The results show a pattern of mutual exclusivity between major-driver genes (with big effects on tumour growth). Putative “mini-driver” genes, with a smaller effect on the cell when mutated, show complex patterns of co-occurrence with other mini-drivers and major-drivers. My results suggest that combinations of Wnt drivers are selected to provide a “just-right” amount of Wnt signalling for a tumour to evolve. In addition, particular combinations of mutations are found in different tumour types and different anatomical locations, suggesting that the just-right level of signalling is different depending on the overall context of the tumour. I then investigated the potential mini-driver gene *TCF7L2* which also has a likely role in increasing inherited risk of CRC, based on studies in both humans (genome-wide association studies) and mouse models of CRC.

Acknowledgements

I would firstly like to thank Professor Ian Tomlinson for this PhD position and project. Without this incredible opportunity, and Ian's belief in my ability to carry out these projects, I would not be the person I am today. A huge thank you goes to my supervisory team at the IGC, Professor Mark Arends and Professor Susan Farrington, I am so grateful for all of your time, effort and support in getting me to the finish line. Thank you to Miss Rachel Guest, my new supervisor and friend, for all of her help practically, motivationally and emotionally. I will take her care of others and love of science with me throughout my career. I would like to say thank you to everyone in the Tomlinson group, past and present (Sophie, James, Nacho, Mel, Chloe, Steve, Rosie and Xell), for their ongoing support, with special thanks to Connor Woolley, Joe Ward and Nagore De Leon for the time they spent helping me with the projects in this thesis. Throughout my PhD, a whole community of highly skilled and kind academics at the IGC supported the completion of my projects including Kevin Myant, Malcolm Dunlop, Farhat Din, Doug Speake and their research groups, whom I am very grateful to.

To all of the people who have been there with me throughout this journey, thank you for laughing and crying with me on a daily basis. Kitty, Harv, Caitlin, Guler, Fraser and all of my new and amazing IGC friends, I will always be so incredibly grateful to you for sharing this experience with me and for making it even more worthwhile. Thank you to all of my incredible friends from home for all of your constant support and belief in me.

To Mum, Dad, Ellie, my grandparents and whole family, you are such an inspiration to me. From learning how to read my first words, to writing a PhD thesis, your love and encouragement are my constant source of strength, motivation and happiness. Finally, to Kai, my partner and best friend, there aren't enough words. Thank you for being with me every second of this journey. I couldn't have done this without you.

For you, my Nanna, Margaret.

Contents

Chapter 1 - Introduction

1.1. Introduction to cancer	- 21 -
1.2 Genetic control of the gastrointestinal tract	- 22 -
1.3 The Wnt pathway	- 23 -
1.4 Wnt-pathway mutations in CRC	- 26 -
1.5 Modifiable risk: Lifestyle factors, early detection and diagnosis	- 28 -
1.5.1 Heritable risk: CRC predisposition syndromes.....	- 29 -
1.5.2 Familial Adenomatous Polyposis	- 30 -
1.5.3 Lynch Syndrome or Hereditary nonpolyposis colorectal cancer	- 31 -
1.5.4 Other CRC predisposition syndromes: PPAP and PJS	- 32 -
1.6 Colorectal cancer evolution	- 33 -
1.6.1 The linear model of CRC evolution.....	- 33 -
1.6.2 The big bang model of CRC evolution	- 35 -
1.6.3 The branched model of CRC evolution	- 36 -
1.6.4. A sub-division of the branched model: The “just-right” and “mini-driver” models of CRC evolution	- 37 -
1.6.5 A potential example of just-right, mini-driver selection: Atypical <i>BRAF</i> and <i>KRAS</i> mutations- 40 -	
1.7 Aims and objectives	- 42 -
2.1 Background.....	- 44 -
2.2 Chapter Aims	- 48 -
2.3.6 Phylogenetic trees	- 56 -
2.3.7 Just-right model analysis.....	- 56 -
2.4 Results.....	- 58 -
2.4.1 Phylogenetic trees	- 58 -
2.4.2.1 MD20716	- 58 -
2.4.2.2 MD20430	- 59 -
2.4.2.3 MD20730	- 61 -
2.4.2.4 MD21031	- 62 -
2.4.3.5 MD20980	- 63 -
2.4.3.6 MD21032	- 65 -
2.4.3.7 MD21099	- 66 -
2.4.3.8 MD21236	- 68 -
2.4.3.9 MD21239	- 69 -
2.4.3.10 MD21257	- 71 -

2.4.3.11 MD21274	- 72 -
2.4.3.12 MD21296	- 74 -
2.4.3.13 MD21320 Polyp 1	- 75 -
2.4.3.14 MD21320 Polyp 2	- 76 -
2.4.3.15 MD21340	- 78 -
2.4.3.16 MD21353	- 79 -
2.4.3.17 MD21354	- 80 -
2.4.3.18 MD21365	- 81 -
2.4.3.19 MD21386	- 83 -
2.4.3.20 Just-right analysis summary	- 84 -
2.4.4 Tumour purity, ploidy and whole genome duplication	- 85 -
2.4.5 Mutation burden	- 88 -
2.4.6 Polyp size	- 92 -
2.4.7 Copy number	- 94 -
2.4.8 ITH	- 97 -
2.5 Discussion	- 100 -
3.1 Background	- 126 -
3.2 Chapter aims	- 129 -
3.3 Methods	- 130 -
3.3.1 Statistical assessment of mutational relationships between Wnt pathway genes	- 130 -
3.3.2 Further just-right analysis in the context of <i>APC</i> 20AARs	- 131 -
3.3.3 Proximal, distal/rectal analysis	- 132 -
3.3.5 FAP	- 133 -
3.4 Results	- 135 -
3.4.1.1 Wnt pathway mutations in primary, treatment naïve CRCs	- 135 -
3.3.1.2 Wnt-pathway major-driver mutations and <i>APC</i> , in MSS CRCs	- 147 -
3.3.1.3 Wnt-pathway major-driver mutations and <i>APC</i> , in MSI ⁺ CRCs	- 148 -
3.3.1.4 Mutations in mini-driver genes in the context of <i>APC</i> -mutant, MSS CRCs	- 149 -
3.3.1.5 Mutations in mini-driver genes in the context of <i>APC</i> -mutant, MSI ⁺ CRCs	- 150 -
3.3.1.6 Other mini/major-driver relationships	- 150 -
3.3.2 The number of 20AARs in <i>APC</i> in the presence of other Wnt pathway mutations (mini-drivers) -	152 -
3.3.3 The number of 20AARs in <i>APC</i> in the proximal and distal colon/rectum	- 153 -
3.3.4 The position of <i>AMER1</i> mutations with and without a mutation in <i>APC</i>	- 156 -
3.3.5 Age differences in MSS and MSI ⁺ CRCs	- 159 -
3.3.6 Exploring the just-right model in FAP	- 160 -
4.1 Background	- 170 -

4.1.1 GWAS a brief history	- 170 -
4.1.2 GWAS: The process	- 171 -
4.1.3 Conditional analysis.....	- 172 -
4.1.4 Post-GWAS.....	- 173 -
4.1.5 <i>TCF7L2</i> in CRC GWAS	- 173 -
4.2 Chapter aims	- 174 -
4.3 Methods.....	- 175 -
4.3.1 Conditional analysis.....	- 175 -
4.3.2 eQTL analysis	- 177 -
4.3 Results.....	- 178 -
4.3.1 Conditional Analysis.....	- 178 -
4.3.2 eQTL analysis	- 187 -
4.4 Discussion.....	- 197 -
5.1 Background.....	- 203 -
5.2 Aims.....	- 205 -
5.3 Methods.....	- 206 -
5.3.1 <i>Tcf7l2</i> mouse strain generation	- 206 -
5.3.2 Genotyping.....	- 207 -
5.3.3 Temperature optimisation and gel extraction.....	- 208 -
5.3.4 Tissue retrieval, processing and analysis	- 209 -
5.3.5 DNA/RNA extraction	- 210 -
5.3.6 qPCR.....	- 210 -
5.3.7 Western blot.....	- 211 -
5.4 Results.....	- 213 -
5.4.1 Construct confirmation	- 213 -
5.4.2 Intestinal analysis.....	- 219 -
5.4.3 <i>Tcf7l2</i> RNA expression in <i>Tcf7l2</i> ^(+/-) mice.....	- 240 -
5.5 Discussion.....	- 244 -
6. Final discussion and conclusions	- 249 -
Further statements and acknowledgements.....	- 254 -
References.....	- 255 -

Abbreviations

100KGP	Genomic England 100,000 Genomes Project
AAR	Amino acid repeats
AFAP	Attenuated familial adenomatous polyposis
AP-1	Activator protein 1
APC	Adenomatous polyposis coli
AXIN1/2	Axis inhibition protein 1/2
BCL9	BCL9 transcription coactivator
BCL9L	BCL9 like
CI	Confidence interval
CID	β -catenin inhibitory domain
CIMP	CpG island methylator phenotype
CK1	Casein kinase 1
CTNNB1	Catenin beta 1
CNA	Copy number alteration
CN-LOH	Copy neutral loss of heterozygosity
CRA	Colorectal adenoma
CRC	Colorectal cancer
DNA	Deoxyribonucleic acid
Dvl	Dishevelled
EGFR	Epidermal growth factor receptor
eQTL	Expression quantitative trait loci
ERK	Extracellular signal-regulated kinase
FAP	Familial adenomatous polyposis
FBXW7	F-Box and WD repeat domain containing 7
FISH	Fluorescent <i>in situ</i> hybridisation
FRT	Flippase recognition target
Fzd	Frizzled
GCTA	Genome-wide complex trait analysis
GSK3 β	Glycogen synthase kinase 3 beta
GWAS	Genome-wide association studies
HNPCC	Hereditary nonpolyposis colorectal cancer

HRP	Horseradish peroxidase
H&E	Haematoxylin and Eosin
IBD	Inflammatory bowel disease
IGV	Integrative genomics viewer
Indels	Insertions and deletions
ITH	Intra-tumour heterogeneity
JNK	c-Jun NH ₂ terminal kinase
JP	Juvenile polyposis
LI	Large intestines
LD	Linkage disequilibrium
Lgr5	Leucine-rich repeat-containing G protein-coupled receptor 5
LOF	Loss-of-function
LOH	Loss of heterozygosity
LRP	Lipoprotein receptor-related protein
LS	Lynch syndrome
MAF	Minor allele frequency
MAPK	Mitogen-activated protein kinase
MCR	Mutation cluster region
MLH1	MutL homologue 1
MMR	Mismatch repair
mRNA	Messenger ribonucleic acid
MSH2	MutS homologue 2
MSH6	MutS homologue 6
MSI	Microsatellite instability
MCR	Mutation cluster region
MWAS	Methylome-wide association study
NGS	Next generation sequencing
OCT	Optimal cutting temperature compound
OR	Odds ratio
PCR	Polymerase chain reaction
PJS	Peutz-Jeghers syndrome
PMS2	Post-meiotic segregation increased 2
PPAP	Polymerase proofreading-associated polyposis

PYGO	Pygopus family PHD finger 1
qPCR	Quantitative polymerase chain reaction
QC	Quality control
REA	Repeated arginine glutamic acid alanine motifs
RFLP	Restriction fragment length polymorphisms
RIPA	Radioimmunoprecipitation assay
RNA	Ribonucleic acid
RNF43	Ring Finger 43
RSPO	R-spondin
S:CORT	Stratification in Colorectal Cancer
SI	Small intestines
SNA	Single-nucleotide alteration
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOCCS	Study of colorectal cancer in Scotland
SOX9	SRY-box transcription factor 9
STK11	Serine/threonine kinase 11
TBR	Tree bisection reconnection
TCF7L2	Transcription factor 7 like 2
TCF/LEF	T-cell factor/lymphoid enhancer binding factor
TCGA	The Cancer Genome Atlas
TEMS	Transanal endoscopic microsurgery
TP53	Tumour protein 53
TSG	Tumour suppressor gene
TWAS	Transcriptome-wide association study
UV	Ultraviolet light
VAF	Variant allele frequency
VCF	Variant call format
VEP	Variant effect prediction
VTI1A	Vehicle transport through interaction with T-SNAREs 1A
WES	Whole-exome sequencing
WGD	Whole genome duplication
WGS	Whole-genome sequencing

Wnt	Wingless/integrated
WT	Wild-type

Figures

<i>Figure</i>	<i>Title</i>
<i>Chapter 1</i>	
1.1	The development of a tumour via stepwise selection, according to Nowell (1976).
1.2	The anatomical structure of an intestinal crypt.
1.3	An illustration of the Wnt pathway
1.4	The structure and functional domains of APC.
1.5	The position of the mutation cluster region (MCR) and β -catenin inhibitory domain (CID) in <i>APC</i> .
1.6	The distribution of genomic factors contributing to the development of CRCs.
1.7	Fearon and Vogelstein's model of sequential cancer development from adenoma to metastatic disease.
1.8	Opposing models of CRC evolution.
1.9	Mini-driver scenarios as presented by Castro-Giner, Ratcliffe and Tomlinson (2015).
1.10	The <i>MAPK</i> signalling pathway.
<i>Chapter 2</i>	
2.1	The predicted clonality of a tumour according to the branched model of CRC evolution.
2.2	Multiregional sampling polyp dissection method.
2.3	Phylogenetic tree of MD20716.
2.4	Phylogenetic tree of MD20430.
2.5	Phylogenetic tree of MD20730.
2.6	Phylogenetic tree of MD21031.
2.7	Phylogenetic tree of MD20980.
2.8	Phylogenetic tree of MD21032.
2.9	Phylogenetic tree of MD21099.
2.10	Phylogenetic tree of MD21236.
2.11	Phylogenetic tree of MD21239.
2.12	Phylogenetic tree of MD21257.
2.13	Phylogenetic tree of MD21274.
2.14	Phylogenetic tree of MD21296.
2.15	Phylogenetic tree of MD21320 polyp 1.
2.16	Phylogenetic tree of MD21320 polyp 2.
2.17	Phylogenetic tree of MD21340.

- 2.18 | Phylogenetic tree of MD21353.
- 2.19 | Phylogenetic tree of MD21354.
- 2.20 | Phylogenetic tree of MD21365.
- 2.21 | Phylogenetic tree of MD21386.
- 2.22 | A comparison of tumour purity between samples taken from polyps with a focus of cancer and without a focus of cancer.
- 2.23 | A comparison of tumour ploidy between polyps with a focus of cancer and without a focus of cancer.
- 2.24 | A comparison of WGD events in polyps with and without a focus of cancer.
- 2.25 | The ploidy levels of samples showing evidence of whole genome duplication (WGD) is significantly higher than those without WGD.
- 2.26 | Mutation burden of all multiregional samples including SNVs and indels.
- 2.27 | Average mutation burden per polyp.
- 2.28 | A comparison of the average mutation burden of polyps with and without a focus of cancer.
- 2.29 | A comparison of the average driver mutation burden between polyps with and without a focus of cancer.
- 2.30 | A comparison of average Intogen cancer driver mutation burden (SNAs/indels), monosomy and CN-LOH of *APC* and *TP53*, and 18q allelic imbalance in polyps with and without a focus of cancer.
- 2.31 | A comparison of size between polyps with and without a focus of cancer.
- 2.32 | The relationship between polyp size (mm) and average mutation burden.
- 2.33 | The relationship between average Intogen colorectal cancer driver mutation burden, size (mm) and age.
- 2.34 | Average number of regions with CNAs per polyp.
- 2.35 | The average number of regions with CNAs in MSS polyps with and without a focus of cancer.
- 2.36 | An overview of all SNV types, monosomy, CN-LOH and WGD events in all MSS polyp samples.
- 2.37 | The distribution of mutations present in one (unique) or more than one sample in MSS polyps with and without a focus of cancer.
- 2.38 | The proportion of sub-clonal driver mutations compared to total driver mutations in polyps with and without a focus of cancer.
- 2.39 | A comparison of relative trunk, branch and leaf lengths in MSS polyps with and without a focus of cancer.

Chapter 3

- 3.1 | A pairwise comparison of pathogenic mutations in Wnt pathway genes, with additional covariates, in treatment naïve, primary CRCs.

- 3.2 A pairwise comparison of pathogenic mutations in Wnt pathway genes with additional covariates in treatment naïve, MSS, primary CRCs.
- 3.3 A pairwise comparison of pathogenic mutations in Wnt pathway genes with additional covariates in treatment naïve, MSI⁺, primary CRCs.
- 3.4 An assessment of associations between mutations in Wnt pathway genes and covariates utilising multiple, pairwise logistic regression models (Method A) in MSS, primary, treatment naïve, CRC samples.
- 3.5 An assessment of associations between mutations in Wnt pathway genes and covariates utilising multiple, pairwise logistic regression models (Method A) in MSI⁺, primary, treatment naïve, CRC samples.
- 3.6 An assessment of associations between mutations in Wnt pathway genes and covariates utilising multiple, logistic regression models (Method B) in MSS, primary, treatment naïve, CRC samples.
- 3.7 An assessment of associations between mutations in Wnt pathway genes and covariates utilising multiple, logistic regression models (Method B) in MSI⁺, primary, treatment naïve, CRC samples.
- 3.8 Multiple logistic regression models with feature selection and cross validation for the purpose of exploring relationships between pathogenic mutations in Wnt pathway genes and covariates, in MSS, treatment naïve, primary CRCs.
- 3.9 Multiple logistic regression models with feature selection and cross validation for the purpose of exploring relationships between pathogenic mutations in Wnt pathway genes and covariates, in MSI⁺, treatment naïve, primary CRCs.
- 3.10 The number of 20AARs in *APC* in the presence of an *AMER1* mutation in both MSS and MSI⁺ CRCs.
- 3.11 The distribution of 20AARs in *APC* in the proximal and distal colon/rectum.
- 3.12 The presence of an *AMER1* mutation in the proximal colon and distal colon/rectum is associated with an increase in the number of 20AARs in *APC*.
- 3.13 A lollipop plot representing all *AMER1* mutations in primary, treatment naïve CRCs in the 100KGP.
- 3.14 The protein structure of the two isoforms of *AMER1* (S1 and S2). Adapted from Tanneburger *et al.* (2011).
- 3.15 A lollipop plot representing all *AMER1* mutations also harbouring a mutation in *APC*, in primary, treatment naïve CRCs in the 100KGP.
- 3.16 A lollipop plot representing all *AMER1* mutations also harbouring a mutation in *APC*, in primary, MSS, treatment naïve CRCs in the 100KGP.
- 3.17 A lollipop plot representing all *AMER1* mutations also harbouring a mutation in *APC*, in primary, MSI⁺, treatment naïve CRCs in the 100KGP.
- 3.18 MSI⁺ CRCs have a significantly higher age at sampling than MSS CRCs.
- 3.19 *APC* mutations are associated with a younger sampling age in MSI⁺ CRCs.
- 3.20 *RNF43* mutations are associated with a higher sampling age in MSI⁺ CRCs.

Chapter 4

- 4.1 Genome-wide significant SNPs identified at the 10q25.2 locus, present in European, CRC meta-analysis data.
- 4.2 SNPs identified at the *TCF7L2* locus, present in European, CRC meta-analysis data.
- 4.3 Genome-wide significant SNPs at the 10q25.2 locus when conditioned on rs56188434, using European, CRC meta-analysis data.
- 4.4 Genome-wide significant SNPs when conditioned on both rs56188434 and rs2296782.
- 4.5 LD matrix for 53 genome-wide significant SNPs at the 10q25.2 locus.
- 4.6 The association of the risk allele at rs2296782 and rs56188434 and *TCF7L2* expression in the proximal colon, distal colon and rectum of 107 individuals in the INTERMPHEN study.
- 4.7 The association of the risk allele at rs2296782 and rs56188434 and *VTI1A* expression in the proximal colon, distal colon and rectum of 107 individuals in the INTERMPHEN study.
- 4.8 The association of risk alleles at two *VTI1A* SNP loci and *VTI1A* expression in the distal colon.
- 4.9 The relationship between *TCF7L2* and *VTI1A* expression in the proximal colon, distal colon and rectum.

Chapter 5

- 5.1 The structure of *TCF7L2*.
- 5.2 The conditional knock-out *Tcf7l2* murine allele with *loxP* sites flanking exon 1.
- 5.3 A schematic of the *Tcf7l2* knock-out allele.
- 5.4 Splenic DNA from three different mice used to determine a sequence map of the *Tcf7l2* allele before and after recombination.
- 5.5 *Tcf7l2*^(fl/fl) allele schematic following sequencing of splenic DNA PCR products.
- 5.6 DNA sequence electrophorograms of A) both *loxP* sites B) a section of the deleted/translated region of exon 1, and C) FRT site of *Tcf7l2*^(fl/fl) splenic DNA.
- 5.7 *Tcf7l2*^(fl/fl) allele DNA sequence from first *loxP* site to the start of exon 2.
- 5.8 A more detailed schematic of the *Tcf7l2*^(fl/fl) allele clarified using sequenced splenic DNA PCR product.
- 5.9 Structure of the *Tcf7l2*^{(+/-);Pgk1-cre^(Pos) recombined allele based on sequencing of splenic DNA PCR products.}
- 5.10 *Tcf7l2* allele genotyping assay for both *Tcf7l2*^(+/+) and *Tcf7l2*^(+/-) mice.
- 5.11 The numbers of macroscopically-identified tumours in both the small and large intestines of all strains of *Tcf7l2*;*Apc* experimental mice.
- 5.12 Genotyping assay results for the two potential outlier mice and controls using DNA from the spleen, large intestine and small intestine.
- 5.13 Box and whisker plots showing the number of tumours in the small and large intestines of *Tcf7l2*;*Apc* test and control mouse strains, and their statistical differences.
- 5.14 Images of intestinal tumours in different experimental strains of mice.
- 5.15 Tumour number by SI section.

5.16	Graphs showing the number of tumours in section 2 of the small intestine according to genotype.
5.17	Graphs showing the number of tumours in section 3 of the small intestine according to genotype.
5.18	Graph showing the number of tumours in section 2 of the small intestine according to genotype.
5.19	Graph showing the number of tumours in section 7 of the small intestine according to genotype.
5.20	Tumour volume by SI section.
5.21	An analysis of tumour volume in section 2 of the small intestine according to genotype.
5.22	Graph showing tumour volume in section 3 of the small intestine according to genotype.
5.23	Graph showing tumour volume in section 6 of the small intestine according to genotype.
5.24	Graph showing tumour volume in section 7 of the small intestine according to genotype.
5.25	Distribution of tumour sizes across different genotypes.
5.26	The distribution of tumours in the large intestines of multiple murine strains.
5.27	Graphs showing the number of tumours in section 5 of the large intestine according to genotype.
5.28	Tumour volume distribution across large intestine (LI) sections in different genotypes.
5.29	Analysis of <i>Tcf7l2</i> RNA expression in splenic tissue.
5.30	Analysis of <i>Tcf7l2</i> RNA expression in large intestinal tissue.
5.31	Analysis of <i>Tcf7l2</i> expression in small intestinal tissue.
5.32	Melt curve plot for a heterozygous <i>Tcf7l2</i> ^(+/+) qPCR liver sample.
<i>Chapter 6</i>	
6.1	A diagrammatic representation of just-right Wnt pathway activation as described by the just-right and mini-driver models of CRC evolution.

Tables

<i>Table</i>	<i>Title</i>
<i>Chapter 1</i>	
1.1	The frequency (%) of mutations found in Wnt pathway-related genes in 8733 CRC samples according to cBioportal.
<i>Chapter 2</i>	
2.1	A description of each large adenomatous colorectal polyp and patient.
2.2	The distribution of pathogenic mutations in <i>APC</i> and the corresponding number of 20AAR repeats retained.
2.3	Genetic profile summary for polyp MD20716.
2.4	Genetic profile summary for polyp MD20430.
2.5	Genetic profile summary for polyp MD20730.
2.6	Genetic profile summary for polyp MD21031.
2.7	Genetic profile summary for polyp MD20980.
2.8	Genetic profile summary for polyp MD21032.
2.9	Genetic profile summary for polyp MD21099.
2.10	Genetic profile summary for polyp MD21236
2.11	Genetic profile summary for polyp MD21239.
2.12	Genetic profile summary for polyp MD21257.
2.13	Genetic profile summary for polyp MD21274.
2.14	Genetic profile summary for polyp MD21296.
2.15	Genetic profile summary for MD21320 polyp 1.
2.16	Genetic profile summary for MD21320 polyp 2.
2.17	Genetic profile summary for polyp MD21340.
2.18	Genetic profile summary for polyp MD21353.
2.19	Genetic profile summary for polyp MD21354.
2.20	Genetic profile summary for polyp MD21365.
2.21	Genetic profile summary for polyp MD21386.
2.22	A summary of remaining 20AARs remaining in <i>APC</i> in polyps with two detectable, clonal APC alterations and presence of mutations in mini-driver genes.
<i>Chapter 3</i>	
3.1	The distribution of pathogenic mutations in <i>APC</i> and the corresponding number of 20AAR repeats retained.
3.2	RSPO2/3 fusions in treatment naïve, primary CRC tumour samples.

3.3	The association between the number of 20AARs in <i>APC</i> and additional Wnt pathway mutations in primary MSS CRCs: Mann-Whitney U results.
3.4	The distribution of 20AARs in <i>APC</i> in the proximal and distal colon, in the presence of additional Wnt pathway mutations, in MSS CRCs: A statistical assessment.
3.5	An investigation of the just-right model in FAP.
<i>Chapter 4</i>	
4.1	A summary of all cohorts utilised in the European GWAS Meta-Analysis.
4.2	A conditional analysis of genome-wide significant SNPs at the 10q25.2 locus, conditioning on lead SNP rs56188434.
4.3	Conditional analysis of genome-wide significant SNPs in the region around <i>TCF7L2</i> region, conditioning on both rs2996782 and rs56188434.
4.4	The effect of risk alleles at lead SNPs at both the <i>TCF7L2</i> locus (rs2296782) and at the <i>VTI1A</i> locus (rs56188434) on <i>TCF7L2</i> expression in the proximal colon, distal colon and rectum.
4.5	The effect of risk alleles at both the lead <i>TCF7L2</i> locus SNP (rs2296782) and the lead <i>VTI1A</i> locus SNP (rs56188434) on <i>VTI1A</i> expression in the proximal colon, distal colon and rectum.
4.6	The effect of risk alleles at ten different <i>VTI1A</i> SNP loci significantly associated with increased <i>VTI1A</i> expression in the distal colon, expression prior to Benjamini-Hochberg correction.
<i>Chapter 5</i>	
5.1	Summary of the mouse strains utilised in the study of <i>Tcf7l2</i> as a mini-driver of intestinal tumourigenesis.
5.2	A description of the PCR primers utilised to genotype the strains utilised in this study.
5.3	A summary of the numbers and ages of mice analysed from each mouse strain.
5.4	Probes used for amplifying and assessing the expression of <i>Tcf7l2</i> via qPCR.

Chapter 1

Introduction

1.1. Introduction to cancer

Nowell (1976), standing on the shoulders of Darwin's "Origin of Species", was the first to propose an association between Darwinian evolution and cancer evolution. His evolutionary model revolves around a Darwinian rationale: diversity emerges from simple beginnings.

Nowell postulated that genetic instability, within a single cell of origin, instigates the development of genetic variability. An induced genetic change, within a singular cell, can provide that cell with a selective growth advantage. Neoplastic proliferation and genetic variability then allow for the sequential selection of fitter and more aggressive sub-clones, resulting in tumour formation and progression (Nowell, 1976) (**Fig.1.1**).

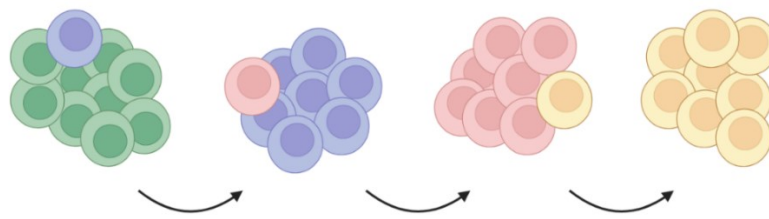


Figure 1.1. The development of a tumour via stepwise selection, according to Nowell (1976). Through an induced genetic modulation, a cell becomes genetically varied from its surrounding clones. This variation results in the acquisition of a selective growth advantage, allowing the clone to proliferate and become dominant. Through repetition of this sequence, sequential selection of fitter and more aggressive sub-clones occurs resulting in tumour formation and progression.

Owing to the prolific development of analytical techniques allowing for the detailed analysis of cancer genetics, it is now understood that the induced genetic changes, referred to by Nowell, that lead to the progression and evolution of a tumour, consist of base-pair substitutions, insertions, deletions, chromosomal rearrangements, copy number alterations (CNA) and epigenetic changes. Together, these deoxyribonucleic acid (DNA) changes may constitutively activate proto-oncogenes to oncogenes and inactivate or suppress the action of tumour suppressor genes (TSG). Maintaining tissue homeostasis relies on the crucial function of proto-oncogenes and TSGs in various molecular cascades, collectively known as signalling pathways. In turn, the mutation of proto-oncogenes/TSGs induces phenotypic cellular changes consisting of sustained proliferative signalling, insensitivity to anti-growth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis and ultimately, tissue invasion and metastasis, commonly referred to as the hallmarks of cancer (Hanahan and Weinberg, 2011).

Understanding the genetic intricacies of CRC evolution is imperative for several reasons. CRC bears a heavy burden on society as a whole. It is the third most common cancer and fourth leading cause

of cancer-related mortality, with over 1.9 million new cases and 930,000 deaths in 2020 (Huyghe *et al.*, 2019a; Morgan *et al.*, 2023). A thorough understanding of CRC genetics and evolution has the potential to improve the prevention, diagnosis and treatment of CRC. For example, the identification and recognition of genetic and molecular variations that occur throughout tumour development can allow for earlier, more accurate diagnoses in addition to enhanced, targeted therapeutic treatment. These same molecular markers also have the ability to provide insight into tumour behaviour in regard to likelihood of progression, response to treatment and prognosis (Guastadisegni *et al.*, 2010). For example, studies have shown *KRAS* mutations to be a predictor of response to CRC treatment with cetuximab. Patients carrying a *KRAS* mutation are unlikely to benefit from its administration (Bachet *et al.*, 2006). It is also believed that approximately 90% of treatment failure in those with metastatic cancer is as a result of mutation and consequently, acquired therapeutic resistance. For example, patients can become resistant to cetuximab in late-stage CRC. Therefore, cancer progression, evolution and cancer evolution in response to treatment are hugely important fields of study. By conducting these studies, molecular variations that emerge throughout the evolutionary process, and mechanisms of resistance, become apparent and their detection and recognition has the potential to improve diagnosis and treatment (Longley and Johnston, 2005). This thesis aims to enhance the current mechanistic understanding of the complex phenomenon that is CRC evolution.

1.2 Genetic control of the gastrointestinal tract

The structure of the gastrointestinal tract is aptly designed for the task of both digesting and absorbing nutrients. The intestinal lumen is lined by epithelial cells, comprising both crypts and villi. Villi protrude out into the intestinal lumen, creating an increased surface area for efficient nutrient absorption. At their base, crypts house leucine-rich repeat-containing G protein-coupled receptor 5 (*Lgr5*)-expressing intestinal stem cells. When intestinal stem cells proliferate, they form transit-amplifying progenitor cells, which in turn differentiate into a range of different cell types including enterocytes, Paneth cells, goblet cells, enteroendocrine cells, tuft cells and M cells (Novellasedumunt, Antas and Li, 2015; Nelson, 2021) (**Fig.1.2**). This process is controlled by various cell signalling pathways allowing for controlled rates of proliferation, differentiation and apoptosis. Signalling pathways such as *Wnt*, *BMP*, *EGF* and *Notch* pathways play a role in intestinal tissue homeostasis. Dysregulation of these tightly regulated processes, via genetic mutation, has a direct effect on the initiation and progression of CRC (Novellasedumunt, Antas and Li, 2015). This thesis focuses on selective constraints within the *Wnt* signalling pathway and its involvement in CRC.

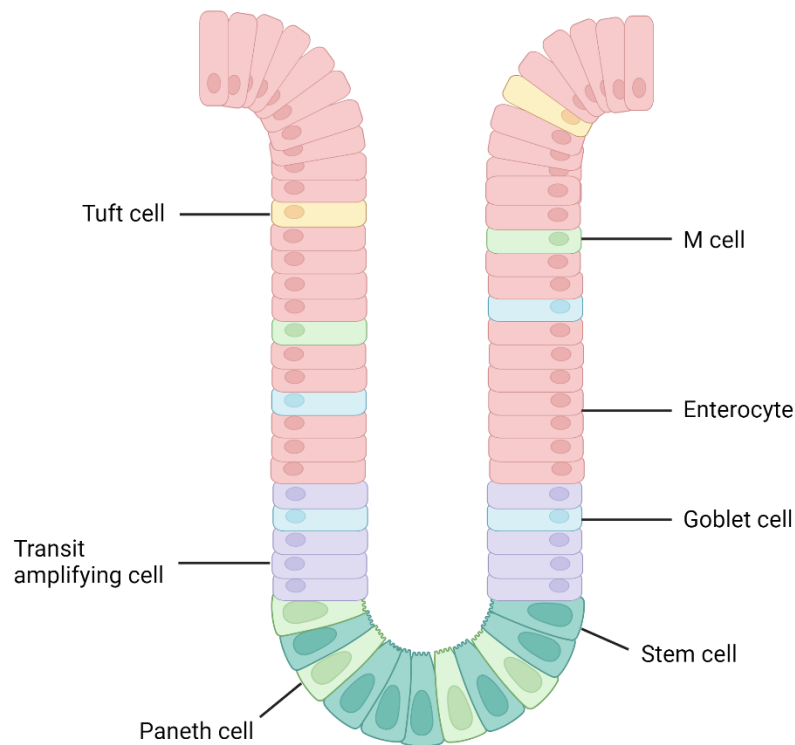


Figure 1.2 The anatomical structure of an intestinal crypt. Adapted from Nelson (2021). Crypts house intestinal stem cells at their base. Intestinal stem cells proliferate and form transit-amplifying cells which in turn differentiate into Paneth cells, goblet cells, enterocytes, tuft cells and M cells (Nicoletti, 2000; Novellasademunt, Antas and Li, 2015; Nelson, 2021).

1.3 The Wnt pathway

The Wnt (wingless/integrated) pathway, conserved in all metazoans, has an ancient and vital role in tissue homeostasis and regeneration (Rim, Clevers and Nusse, 2022). The Wnt pathway is divided into two subsections, the canonical β -catenin dependent pathway and the non-canonical β -catenin independent pathway. This thesis focusses mainly on the canonical, β -catenin dependent Wnt signalling pathway, which has an important role in regulating cell proliferation and tissue development, maintenance and repair (Liu *et al.*, 2022; Rim, Clevers and Nusse, 2022).

The Wnt pathway is heavily implicated in CRC, with mutations in the adenomatous polyposis coli (*APC*) TSG being an early, if not initiating event in 80-85% of CRCs (Zhang and Shay, 2017). *APC*, as well as adenomatous polyposis coli membrane recruitment 1 (*AMER1*) and *AXIN1/2* act as key components of a β -catenin (*CTNNB1*) destruction complex in which they tightly control levels of cytoplasmic β -catenin. The destruction complex along with glycogen synthase kinase (*GSK3*) and casein kinase 1 (*CK1*) prevent the translocation of β -catenin into the nucleus where it functions to promote the transcription of target genes. Wnt signalling is therefore tightly controlled to prevent aberrant cell proliferation (Rim, Clevers and Nusse, 2022; Farnhammer, Colozza and Kim, 2023; Lei *et*

al., 2023). Conversely, When Wnt1 class extracellular Wnt ligands such as WNT2, WNT3, WNT3a and WNT8a bind to and engage receptor Frizzled (Fzd) and co-receptor lipoprotein receptor-related protein 5/6 (LRP5/6), a subsequent cascade of events result in the docking of the destruction complex to the plasma membrane, inhibiting its ability to target β -catenin for degradation. This leads to the accumulation of β -catenin in the cytoplasm and subsequent translocation into the nucleus (Komiya and Habas, 2008; Farnhammer, Colozza and Kim, 2023). Once inside the nucleus, β -catenin can promote the transcription of target genes via the transcriptional co-activation of members of the T-cell factor/lymphoid enhancer binding factor (TCF/LEF) DNA-binding transcription factors (Farnhammer, Colozza and Kim, 2023) (**Fig. 1.3**).

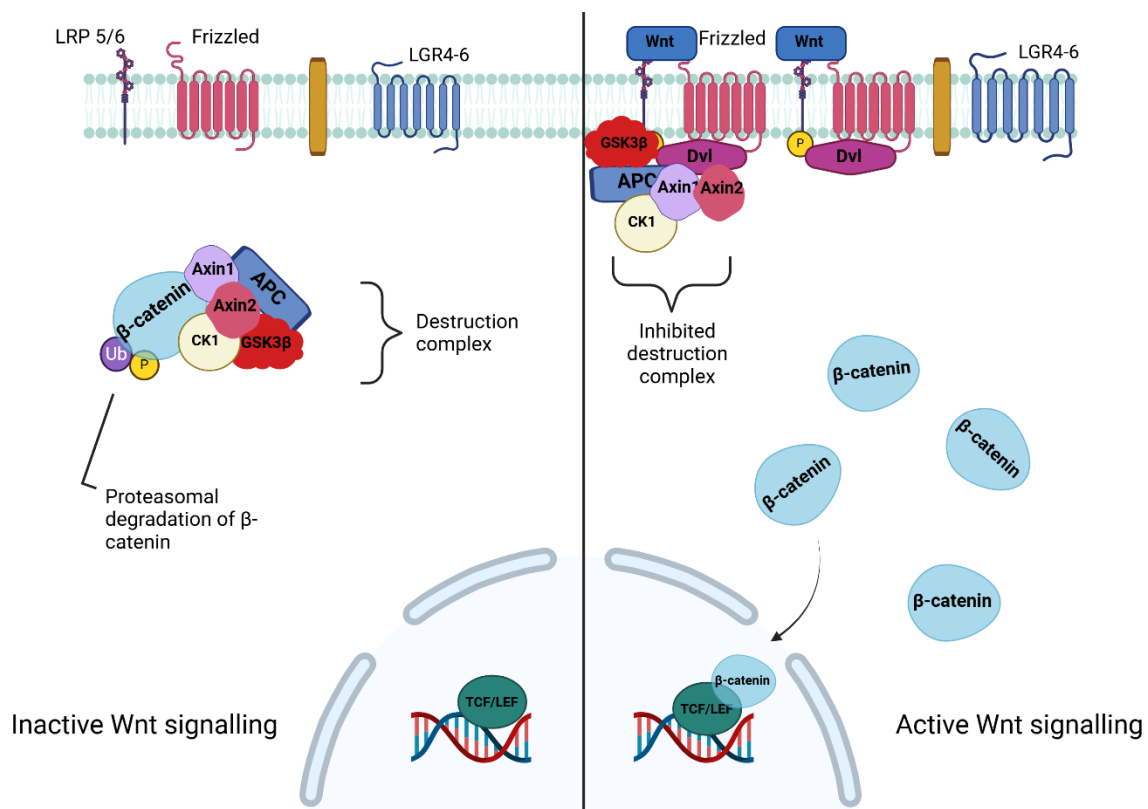


Figure 1.3. An illustration of the Wnt pathway. In the absence of Wnt, APC and AXIN1/2 act as scaffold to the β -catenin destruction complex and bind β -catenin. GSK3 and CK1 phosphorylate β -catenin and target it for proteasomal degradation. When Wnt ligands engage the Fzd receptor and LRP5/6 co-receptors, Dvl is recruited to the membrane. A subsequent cascade of events result in the docking of the destruction complex to the plasma membrane. As a result, the destruction complex can no longer target β -catenin for degradation, leading to an accumulation in the cytoplasm and translocation in to the nucleus where β -catenin can interact with members of the TCF/LEF family of transcription factors and promote the transcription of target genes (Rim, Clevers and Nusse, 2022).

In 2024, a study of approximately 2,000 CRC cases identified over 200 candidate driver genes. Of these 200, there were 13 Wnt pathway-related genes (Cornish *et al.*, 2024). This thesis explores the role of these 13 Wnt genes in colorectal cancer evolution, they include: *AMER1*, *APC*, *AXIN1*, *AXIN2*,

BCL9, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2* and *ZNRF3*. The canonical function of these genes is to further control the level of β -catenin signalling and thus target gene transcription.

Acting alongside APC in the β -catenin destruction complex, APC membrane recruitment protein 1 (AMER1) acts as a scaffold, thus aiding in the ubiquitination and degradation of β -catenin (Tanneberger, Pfister, Kris *et al.*, 2011). Conversely, another function of AMER1 is to promote the phosphorylation of co-receptor LRP6 and therefore aid in the inhibition of the destruction complex and increase β -catenin signalling (Tanneberger, Pfister, Brauburger *et al.*, 2011).

Interestingly, the E3 ubiquitin ligase Ring finger 43 (RNF43) and its homologue Zinc and ring finger 3 (ZNRF3) act as negative regulators of the Wnt signalling pathway. Their function is to target both the Fzd and LRP receptors for lysosomal degradation. Fzd and LRP are therefore no longer able to inhibit the destruction complex allowing for uninhibited targeting and degradation of β -catenin. RNF43 and ZNRF3 are themselves regulated by the binding of R-spondin (RSPO) to the leucine-rich repeat containing G-protein receptor (LGR). When R-spondin binds to an LGR receptor, the R-spondin-LGR-RNF43/ZNRF3 complex is internalised, inhibiting degradation of Fzd. If no R-spondin binds to the receptor, RNF43 is once again free to target Fzd (**Fig.1.3**) (Hao *et al.*, 2012; Koo *et al.*, 2012; Rim, Clevers and Nusse, 2022).

As stated above, once inside the nucleus, β -catenin promotes the transcription of target genes via the transcriptional co-activation of members of the TCF/LEF DNA-binding transcription factors (Farnhammer, Colozza and Kim, 2023). The TCF/LEF family of transcription factors includes LEF1, TCF7L1 and TCF7L2. LEFs/TCFs harbour a domain which functions to interact with armadillo repeats and thus β -catenin (Lien and Fuchs, 2014). TCF7L1 acts as a transcriptional repressor that is inactivated by β -catenin (Shy *et al.*, 2013). TCF7L2 acts as both an activator and repressor dependent on context (Guo *et al.*, 2021). *TCF7L2* will be discussed further in Chapters 4 and 5.

Nuclear proteins Pygopus family PHD finger 1 (PYGO) and BCL9 transcription coactivator (BCL9) play a role in transcriptional regulation of β -catenin and Armadillo. BCL9 assists in the binding of PYGO to β -catenin. The PYGO- β -catenin complex then forms a chromatin reader. Therefore BCL9, by providing key binding sites for transcription, is a transcriptional coactivator in the Wnt signalling pathway (Wu *et al.*, 2023). BCL9-like (BCL9L) is a homologous coactivator of the β -catenin transcription factor complex (Huge *et al.*, 2020).

SRY-box transcription factor 9 (SOX9) is an additional transcription factor that acts within the Wnt pathway. Topol *et al.* (2009) showed evidence for the ability of SOX9 to inhibit Wnt signalling via promotion of β -catenin degradation and inhibiting its transcriptional activity (Topol *et al.*, 2009a). A more recent study exemplifies the ability of TCF-SOX9 interactions to regulate gene transcription and increase the rate of CRC growth showing that the action of SOX9 is context-dependent (Ramakrishnan *et al.*, 2023).

The non-canonical Wnt signalling pathway is divided into two types, the Wnt/calcium pathway and the Wnt/c-Jun NH₂ terminal kinase (JNK) pathways. Previous studies have shown the ability of the non-canonical pathways to inhibit the canonical pathway. Activator protein 1 (AP-1) is a transcription factor composed of JUN and FOS family proteins. Activation of the JNK pathway results in the phosphorylation of JUN. Via addition of Wnt-3a to chondrocytes, Hwang *et al.* (2005) showed evidence of the phosphorylation of c-JUN by JNK resulting in the activation of AP-1. AP-1 activation then resulted in the suppressed expression of transcription factor SOX9 (Weber, Paricio and Mlodzik, 2000; Suto *et al.*, 2004; Liao *et al.*, 2006). Jun expression is controlled by F-Box and WD repeat domain containing 7 (FBXW7)-mediated ubiquitination and proteasomal degradation of c-Jun. FBXW7-mutated mice present elevated levels c-Jun expression and develop adenomas at 9-10 months of age (Yeh, Bellon and Nicot, 2018) (**Fig.1.3**).

1.4 Wnt-pathway mutations in CRC

As stated above, *APC* mutations play a crucial role in the initiation and progression of a large proportion of CRCs. Acting as part of the β -catenin destruction complex, its four 15-AARs and seven 20-AARs bind and downregulate β -catenin. Truncating mutations in *APC* mostly occur upstream of the SAMP repeats, resulting in loss of SAMP-AXIN binding (Smits *et al.*, 1999; Albuquerque *et al.*, 2002; Rim, Clevers and Nusse, 2022) (**Fig. 1.4**).

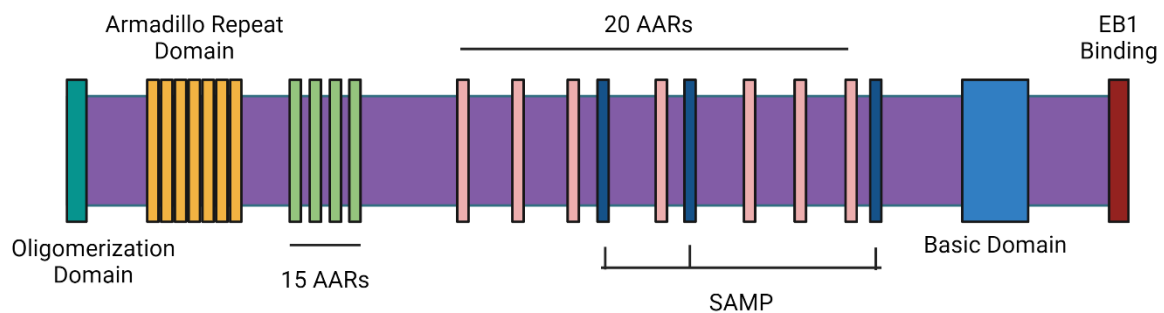


Figure 1.4 The structure and functional domains of APC. APC acts as part of the β -catenin destruction complex in the Wnt signalling pathway. It has four 15-AARs and seven 20-AARs that bind and downregulate β -catenin. The three SAMP repeats of APC bind to other proteins in the destruction complex.

Most *APC* mutations in CRC occur in the mutation cluster region (MCR) between 20AAR1 (R1) and 20AAR3 (R3). Kohler *et al.* (2009) evidenced the existence of a β -catenin inhibitory domain (CID) that exists between 20AAR2 (R2) and R3. The CID region downregulates the level of β -catenin and therefore its transcriptional activity (Kohler *et al.*, 2009). SAMP domains are responsible for interaction with AXIN as part of the β -catenin destruction complex (Matsui *et al.*, 2008) (**Fig. 1.5**).

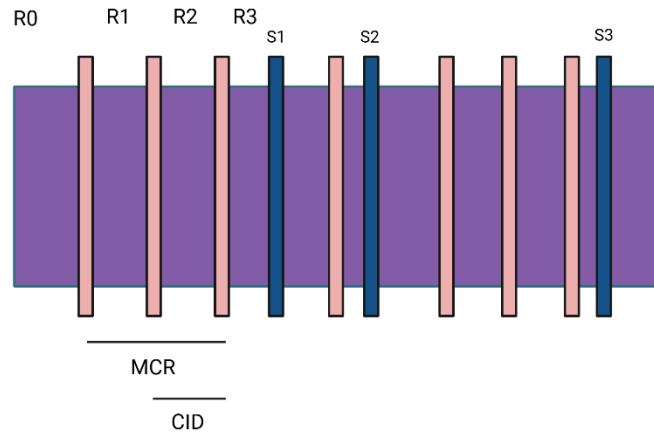


Figure 1.5 The position of the mutation cluster region (MCR) and β-catenin inhibitory domain (CID) in *APC*. The mutation cluster region is positioned between 20-amino-acid-repeat 1 (R1) and 20AAR3 (R3). The majority of *APC* mutations occur within the MCR. The β-catenin inhibitory domain (CID) is positioned between R2 and R3. This region downregulates β-catenin levels and therefore transcriptional activity. SAMP repeats are also shown as S1, S2 and S3.

As well as *APC* mutations, both *KRAS* mutations and chromosome deletions of 18q are highly common in CRC. *KRAS* mutations have been identified in 35-45% of CRCs, whilst 18q deletions occur in over 70% of CRCs. This chromosome harbours SMAD family members *SMAD2* and *SMAD4* (Nguyen and Duong, 2018). *TP53* is also an important gene in the context of CRC, with approximately 40-50% of cases harbouring a *TP53* mutation (Russo *et al.*, 2005).

The frequency of mutations in other Wnt pathway genes of interest, as recorded by cBioportal, an online platform comprising of an aggregation of mutational CRC data from multiple studies, encompassing 8733 samples, can be found in Table 1.1 (Cerami *et al.*, 2012).

Gene	Frequency (%)
<i>FBXW7</i>	14
<i>SOX9</i>	10
<i>TCF7L2</i>	10
<i>RNF43</i>	8
<i>AMER1</i>	7
<i>CTNNB1</i>	6
<i>AXIN2</i>	5
<i>BCL9</i>	3
<i>BCL9L</i>	3
<i>AXIN1</i>	2
<i>ZNRF3</i>	2
<i>JUN</i>	1

Table 1.1 The frequency (%) of mutations found in Wnt pathway-related genes in 8733 CRC samples according to cBioportal.

1.5 Modifiable risk: Lifestyle factors, early detection and diagnosis

The initiation and evolution of sporadic CRC, as opposed to diseases of Mendelian inheritance, is complex. Whereas a Mendelian trait or disease can be traced back to a single gene mutation, cancer often evolves through an intricate interplay between both genetic and environmental influences.

As stated above, cancer is a genetic disease, in which these genetic changes or mutations in DNA can be either germline/constitutional or somatically acquired. Germline mutations are of inherited origin, whereas somatic mutations are sporadic and acquired occurring throughout a person's lifetime (Meyerson *et al.*, 2020). Various lifestyle and other environmental factors increase the likelihood of obtaining pathogenic, somatic mutations in proto-oncogenes or TSGs and influence the development of CRC. Through the identification and modification of various contributing lifestyle factors allowing for protection against genomic error accumulation, CRC and other cancer incidence within the population can be reduced via an increased understanding of preventative measures, implementation of precautionary behaviours and even modification of government policy.

A pivotal longitudinal study by Doll and Hill (1956) showed a significant link between smoking tobacco and subsequent mortality rates in medical professionals. Their four-year study showed

a steady increase in the death rate as a result of lung cancer, at a rate increasing with quantity of smoke intake by each individual (Doll and Hill, 1956). It has now been evidenced that there are over 60 cancer-causing carcinogens present in cigarette smoke able to modify DNA and therefore contribute to cancer initiation (Hecht, 2002).

Geographical variation in CRC incidence has highlighted the influence of economic growth and industrialisation on CRC risk. Approximately 55% of CRC cases occur in more developed countries. This may be attributable to lifestyle factors such as increased consumption of red and processed meat, a sedentary lifestyle and increased obesity levels, which have all been linked to an increased risk of CRC (Johnson *et al.*, 2013; Favoriti *et al.*, 2016).

Early detection is another key modifiable factor that significantly influences patient outcome. Beginning as a benign pre-cancerous growth, adenomatous polyps can often be removed endoscopically (following bowel screening) without recurrence. A patient with Dukes' A (stage I tumour), with their tumour limited to the mucosa, submucosa and muscularis propria layers of the bowel, can undergo surgery resulting in an approximate five-year survival rate of 93%. A patient with Dukes' C (stage III tumour), has a tumour that has spread to at least one lymph node close to the bowel and the survival rate reduces to approximately 47% after five years (National Cancer Research Institute, 2010). It is therefore clear that understanding cancer evolution is of great importance to the prognosis as cancer spread and its evolutionary timeline strongly affects patient outcome.

1.5.1 Heritable risk: CRC predisposition syndromes

Cancer's multifaceted aetiology is emphasised by its evident links between influential lifestyle factors and their effect on the human genome. However, the importance of genetics and its role in cancer development and evolution is emphasised through the study of heritable disease.

Unlike sporadic CRC, caused by the occurrence of somatic mutations, approximately 35% of CRC cases are linked to genetic predisposition (Houlston *et al.*, 2016). 5% of CRCs can be attributed to hereditary cancer syndromes such as familial adenomatous polyposis (FAP) and Lynch Syndrome (LS)/hereditary nonpolyposis colorectal cancer (HNPCC), which are both autosomal dominant, Mendelian inheritance patterns. Inflammatory Bowel Diseases (IBD), Ulcerative Colitis and Crohn's Disease, both confer an increased risk of CRC development (accounting for <1% CRC) and both are associated with single nucleotide polymorphisms that increase their risk. Less than 1% of CRCs are attributed to low-penetrance genetic variations and a larger proportion of CRCs are attributed to currently unknown inherited genomic aberrations or polymorphisms. This highlights a gap in our current understanding of cancer-causing germline mutations and suggests more studies are required to fully understand the heritable risks associated with CRC (Keum and Giovannucci, 2019) (**Fig.1.6**).

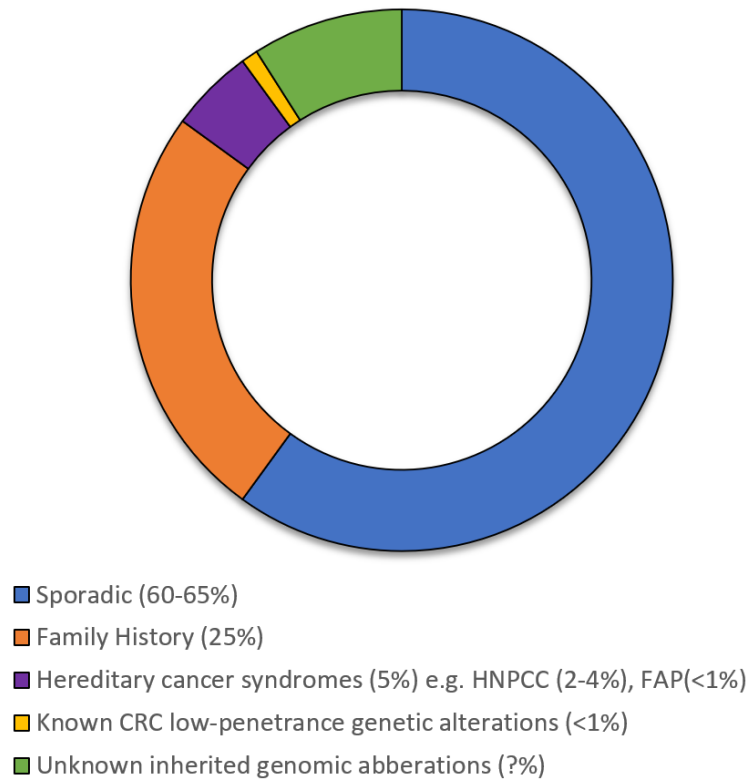


Figure 1.6. The distribution of genomic factors contributing to the development of CRCs. The great majority of CRCs are of sporadic origin and occur without a family history of CRC, through the occurrence of somatic mutations in driver genes such as *APC*, *KRAS*, *p53* etc. Approximately 35% of CRC cases have a hereditary component. This includes those with a family history of CRC with no obvious genetic component, autosomal dominant inherited disease e.g. FAP and LS/HNPCC, CRC attributed to low penetrance genetic variations and unknown genetic aberrations. Adapted from (Keum and Giovannucci, 2019).

This section discusses the genetic and mutational underpinnings of multiple CRC predisposition syndromes highlighting the importance of their identification, and the effect this can have on CRC diagnosis and disease management.

1.5.2 Familial Adenomatous Polyposis

TSGs, as mentioned in section 1.1, play a critical role in the homeostatic regulation of various cellular processes. Their disruption can result in tumour initiation and progression. Knudson, in 1971, was the first to describe TSGs, through the study of retinoblastoma. Knudson noted the existence of heritable, recessive, tumour-initiating mutations in one allele, which in the case of retinoblastoma, are followed by a second hit, or somatic mutation, in the second allele. This is famously known as Knudson's "two-hit hypothesis", which describes the role of TSGs in cancer initiation and progression (Knudson, 1971).

An understanding of the role of TSG, adenomatous polyposis coli (*APC*) has been pivotal in the study of CRC genetic predisposition. Its 16 exons encode for a protein with three 15-amino-acid-repeats (AARs) and seven 20-AARs that enables APC to play a central role in the canonical Wnt signalling pathway. APC binds and regulates β -catenin levels, preventing a build-up of β -catenin in the nucleus. In addition, APC also harbours three SAMP repeats, which have an important role in binding to other proteins in the canonical Wnt signalling multi-protein destruction complex. This destruction complex is able to phosphorylate β -catenin targeting it for subsequent ubiquitination and proteolytic degradation (Komiya and Habas, 2008; Topol *et al.*, 2009b; Tanneberger, Pfister, Kriz, *et al.*, 2011).

Approximately 1% of CRC cases are attributed to FAP, caused by the autosomal dominant inheritance of a germline mutation in the *APC* TSG on chromosome 5q21. Over 300 different mutations (insertions, deletions, nonsense mutations) occurring within *APC* have been linked to FAP, characterised by the development of hundreds/thousands of adenomas within the colorectum. Another commonality of FAP is the absence of symptoms for numerous years until the adenomatous polyps have grown significantly in size resulting in colorectal bleeding or anaemia. If left untreated via prophylactic colectomy surgery, malignancy commonly occurs in one or more of the adenomas, in the majority of patients, at the ages of 40-50 years (Half, Bercovich and Rozen, 2009). Patients with attenuated FAP (AFAP), with mutations at the 5' end of *APC*, develop fewer polyps than classic FAP patients and subsequent cancer progression occurs at an older age (Soravia *et al.*, 1998; Talseth-palmer, 2017).

Through the linkage and identification of *APC* and its role in FAP, mutation testing is utilised to aid in the diagnosis and identification of affected individuals and families. Affected families are offered genetic counselling and surveillance. Mutation testing also prevents the unnecessary surveillance of family members who are not carriers of the mutation (Dunlop, 2002).

1.5.3 Lynch Syndrome or Hereditary nonpolyposis colorectal cancer

An accumulation of DNA damage is crucial for the initiation and evolution of a tumour. Cells employ DNA repair pathways, such as the DNA mismatch repair (MMR) pathway, to maintain genomic integrity and prevent the accumulation of errors. During cell replication, errors may occur either due to base mismatches when incorrectly inserted by the polymerases or due to slippage of the DNA polymerases on repetitive sequences (often termed microsatellites) in the template strand. If these errors remain undetected by polymerase proofreading repair mechanisms, MMR pathway proteins are employed by the cell to identify, excise and correct these mismatched base pairs or expanded/contracted microsatellites. Lynch Syndrome or HNPCC, is another example of an autosomal dominant, genetic tumour syndrome disorder. Unlike FAP, LS/HNPCC does not manifest with an abundance of

adenomatous polyps. LS/HNPCC is characterised by a propensity for early-onset CRC and primarily endometrial cancer amongst other extra-colonic cancers.

LS/HNPCC is known to be due to autosomal dominant, heterozygous, loss-of-function (LOF) mutations in MMR-related genes such as mutL homologue 1 (*MLH1*), mutS homologue 2 (*MSH2*), mutS homologue 6 (*MSH6*) and post-meiotic segregation increased 2 (*PMS2*). These MMR gene mutations inactivate the MMR pathway result in variations in length of repetitive or microsatellite sequences which is easily identifiable as microsatellite instability (MSI), a hallmark of LS/HNPCC (Lynch and Smyrk, 1996; Lynch *et al.*, 2015). LS/HNPCC may be suspected clinically and confirmed by genetic testing. If a patient has three or more family members with CRC or two with CRC and one with endometrial cancer, one of the affected relatives was under 50 years of age at diagnosis and two of the affected relatives are of the first degree, LS/HNPCC is suspected and further investigations are performed, including analysis of the tumours for loss of expression of MMR proteins by immunohistochemistry or presence of MSI by tumour DNA testing leading on to MMR gene testing in constitutional DNA for inherited mutations. However, fulfilment of these criteria is not always necessary as causal MMR mutations can be found in CRC, endometrial and other tumours without the classical family history (Dunlop, 2002).

As with FAP, an understanding of the underlying genetic basis of LS/HNPCC has aided in disease management, including MMR mutation testing, establishment of family registries, and early surveillance (Dunlop, 2002).

1.5.4 Other CRC predisposition syndromes: PPAP and PJS

Additional adenomatous polyposis syndromes are caused by inherited mutations within DNA repair pathways. Polymerase proofreading-associated polyposis (PPAP) occurs due to inherited mutations in two DNA polymerase genes *POLE* and *POLD1*. The role of these two genes is pivotal to the synthesis of DNA, both the leading and lagging strands during DNA replication. There are specific mutations, causing PPAP, that occur in the exonuclease proofreading domain of each protein. The exonuclease domain functions to detect and remove erroneously incorporated bases, a mutation within which allows large numbers of single base mutations to accumulate. The PPAP phenotype involves the development of colorectal adenomas and subsequent progression to carcinoma, although the presence of somatic *POLE* mutation has been associated with better prognosis (Briggs and Tomlinson, 2013; Palles, Latchford and Valle, 2018).

Peutz-Jeghers syndrome (PJS) is another autosomal dominant, inherited intestinal polyp/cancer predisposition syndrome, caused by a germline mutation in serine/threonine kinase 11 (*STK11*). The disease is characterised by gastrointestinal hamartomatous polyposis, mostly restricted to the small bowel (60-90%) and colon (50-64%), with mucocutaneous peri-oral pigmentation and its rate

of occurrence is between one in 50,000 and one in 200,000 live births. PJS polyps are often referred to as hamartomas due to their characteristic histological features. Progression of a PJS polyp into a malignant state has been described as an uncommon event with claims that this evolutionary pathway follows its own unique hamartoma-adenoma-carcinoma sequence. Individuals with PJS have an increased risk of certain cancers, such as gastrointestinal cancers, but not necessarily as a result of the PJS polyps themselves (Dunlop, 2002; Beggs *et al.*, 2010). For instance, in 2011, histology and endoscopy reports were analysed for a total of 2456 polypectomies from 63 PJS patients from 48 families. Six of these polyps contained epithelial atypia or dysplasia and there was no significant correlation between size of the polyp and presence of dysplasia. Three polyps were recorded to be adenomas with no evidence of transformation (Latchford *et al.*, 2011). There is an argument that some of these PJS polyps show some histological features of mucosal prolapse, which may provide a potential explanation for their lack of progression (Beggs *et al.*, 2010).

The identification of these CRC predisposition syndromes and enhanced knowledge of the genetic aberrations that cause them has allowed for more targeted surveillance and the development of early detection strategies. However, there are CRC predisposition genes that are yet to be identified due to the existence of families with clear germline cancer susceptibility transmission where the usual genetic suspects have been ruled out (Dunlop, 2002). Through the identification and study of cancer predisposition genes we can understand the heritable basis of cancer evolution and improve disease management, outcomes and treatment.

1.6 Colorectal cancer evolution

1.6.1 The linear model of CRC evolution

The work of Fearon and Vogelstein on CRC evolution, builds upon Nowell's stepwise model, discussed in section 1.1. Their linear model of CRC evolution suggests that progression from colorectal adenoma (CRA) to an invasive lesion/colorectal carcinoma involves genetic instability, followed by the stepwise accumulation of driver mutations, conventionally initiated via "two hits" to *APC*. This initial mutation is then followed by mutations in other driver genes such as *KRAS* and *TP53*. Through the occurrence of a series of "selective sweeps", clones harbouring advantageous driver mutations competitively expand, resulting in the stepwise progression of an adenoma to malignancy (**Fig. 1.7**) (Fearon and Vogelstein, 1990a).

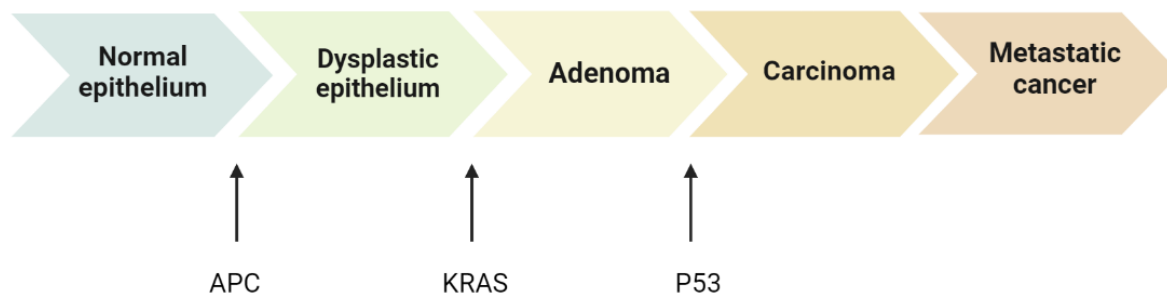


Figure 1.7. Fearon and Vogelstein’s model of sequential cancer development from adenoma to metastatic disease. Black arrows represent the acquisition and accumulation of example driver mutations, in genes such as *APC* and *KRAS* and *TP53*, that result in the competitive expansion of cells progressing towards malignancy then metastatic disease. Adapted from Fearon and Vogelstein (1990).

Following each selective sweep, the previous clones are eliminated and hence the tumour is thought to consist mostly of a dominant clone, or group of cells with a distinct set of driver mutations that Fearon and Vogelstein propose, will likely have been acquired in a specific order e.g. *APC*, *KRAS*, *TP53*.

Through the development of techniques such as G-banding karyotyping, fluorescent *in situ* hybridisation (FISH) and next generation sequencing (NGS), a dramatic advance in the ability to characterize diverse cancer types has been realised. Importantly, NGS techniques have revealed the remarkable complexity of the genetic composition of tumours. Intra-tumour heterogeneity (ITH), the existence of genetic subpopulations within tumours, has now been widely acknowledged (Fisher, Pusztai and Swanton, 2013). This finding is in direct contrast to the linear model of CRC evolution as the linear model predicts tumours to be monoclonal.

The linear model also suggests that, via the chronological accumulation of mutations, a CRC should theoretically harbour a full complement of driver mutations e.g. *APC*, *KRAS* and *TP53*. Therefore, CRAs and CRCs would be genetically distinguishable from one another via this chronological accumulation of driver mutations. However, a study conducted by Smith *et al.* (2002) analysed the mutational landscapes of resection specimens taken from 106 patients undergoing surgery for CRC. They found that only 6.6% of CRCs carried mutations in all three major-driver genes proposed as necessary for progression by the Vogelstein model: *APC*, *KRAS* and *TP53*. 38.7% of CRCs carried mutations in only one of these genes (Smith *et al.*, 2002). These findings highlight the ability of CRCs to arise without a full complement of driver mutations, further highlighting some inaccuracies of the linear model of CRC evolution.

More evidence opposing the linear model has arisen from a multiregional analysis, carried out by Cross and colleagues (Cross *et al.*, 2018), who utilised whole genome sequencing (WGS) and whole exome sequencing (WES) to analyse the evolutionary fitness landscape of nine CRAs and 15 CRCs.

Cross *et al.*, (2018) found ITH in both CRAs and CRCs, albeit to a higher extent within CRAs. This suggests that different sub-clones have the ability to exist and evolve in spatially discrete areas. CRCs were less heterogeneous than CRAs, supported by their phylogenetic analyses, this suggests ongoing selection within CRAs but perhaps negative or stabilising selection in CRCs (Cross *et al.*, 2018). Additional studies have described the evolutionary trajectory of established tumours as “effectively” neutral. If evolution were neutral, this would question the occurrence and existence of selective sweeps, another key premise of the linear model of CRC evolution (Williams *et al.*, 2016).

1.6.2 The big bang model of CRC evolution

The big bang model proposes a non-linear, alternative model of CRC evolution. Sottoriva *et al.* (2015) claim that ITH serves as a highly valuable tool for unravelling the narrative of the initial stages of cancer development. They profiled four adenomas and 11 carcinomas by sampling and sequencing a large number of multiregional tumour glands, to look at the spatial distribution of copy number alterations (CNAs). CNAs can help to identify distinct subpopulations and construct cellular phylogenetic ancestries (Sottoriva *et al.*, 2015).

Based on their findings, Sottoriva *et al.* (2015) argued that, rather than a stepwise accumulation of driver mutations, the mutations required for the transition from adenoma to carcinoma occur very early in tumour development. Therefore, an inherent difference between the big bang model and the linear model of CRC evolution is the theoretical timing of the acquisition of driver mutations. Sottoriva *et al.* (2015) proposed that, following initial transformation, the tumour grows predominantly as a single expansion from a common ancestor. Each sub-clone is predicted to carry the same mutations as the initiating cell. These shared, or public, mutations are capable of driving tumourigenesis. Following initial selection of the fittest sub-clone, tumour evolution will follow a pattern of evolutionary neutrality. This theory suggests a lack of selective sweeps throughout sub-clonal evolution. Over time, the tumour undergoes diversification through the accumulation of additional mutations, producing high levels of ITH. These private mutations, as opposed to public mutations, arise as a result of ongoing replication errors. Private mutations are unique to each sub-clone and are not thought to be dominant, nor provide the sub-clone with a selective advantage. Only those private mutations that arise early enough in development will be pervasive. Late arising private mutations are uniquely present in small sections of the resulting tumour. Therefore, this theory suggests that all mutations capable of driving tumourigenesis are present from tumour initiation, following this, tumour evolution is neutral (Sottoriva *et al.*, 2015).

Cross *et al.* (2018) were able to identify high levels of ITH within CRAs, consistent with the big bang model of CRC evolution. In addition, they also found no significant difference between the frequency of tier 1 (likely pathogenic) and tier 2 (uncertain pathogenic changes in carcinomas or pan-

cancer driver genes) mutations in driver genes between CRAs and CRCs, indicating the acquisition of driver mutations in early development. Interestingly however, through their use of multiregional sequencing analysis, they were able to indicate a divergence of single-nucleotide alterations (SNA) between biopsies; tier 1 driver mutations were not always clonal in CRAs. According to the big bang model, all mutations able to drive tumourigenesis should be present from cancer initiation and are therefore predicted to be clonal. These results of the WGS and WES-based study by Cross *et al.* (2018) are in direct contradiction of this premise.

1.6.3 The branched model of CRC evolution

Through the multiregional analysis of CRAs, Cross *et al.* (2018) underlined the ability of sub-clones to harbour different major-driver mutations and coexist in spatially discrete areas providing supportive evidence for a lack of selective sweeps. As with the big bang model, the branched model of CRC evolution proposes that clones diverge from a common ancestor. Shared, ubiquitous events are represented on the trunk of a tumour's phylogenetic tree. However, distinctive to this model, driver mutations continue to accumulate within different sub-clones forming intermediates, able to coexist alongside more dominant sub-clones. A main premise of this model is that different sub-clones are able to evolve in parallel with one another. Mutations unique to each sub-clone can be mapped onto the branches of a tumour's phylogenetic tree and may each represent a different route to carcinogenesis (Gerlinger *et al.*, 2014).

The premise that different sub-clones may co-exist and evolve in spatially discrete areas within CRAs is therefore supported by the tier 1 driver heterogeneity identified in adenomas by Cross *et al.*, (2018). In addition, as stated above, the finding by Smith *et al.*, (2002) that only 6.6% of tumours carry mutations in *APC*, *KRAS* and *TP53* highlights the alternative pathways to malignancy that must exist for progression to occur in a high percentage of cases.

This model therefore resolves the issue of ITH experienced by the linear model. In addition, it provides an alternative explanation as to why Cross *et al.*, (2018) found evidence of tier 1 driver heterogeneity between biopsies. Unlike the big bang model which states that all mutations capable of driving tumourigenesis are present from the point of tumour initiation; cells are “born to be bad”. A diagrammatic comparison of each of the three models of CRC evolution is shown in Figure 1.8.

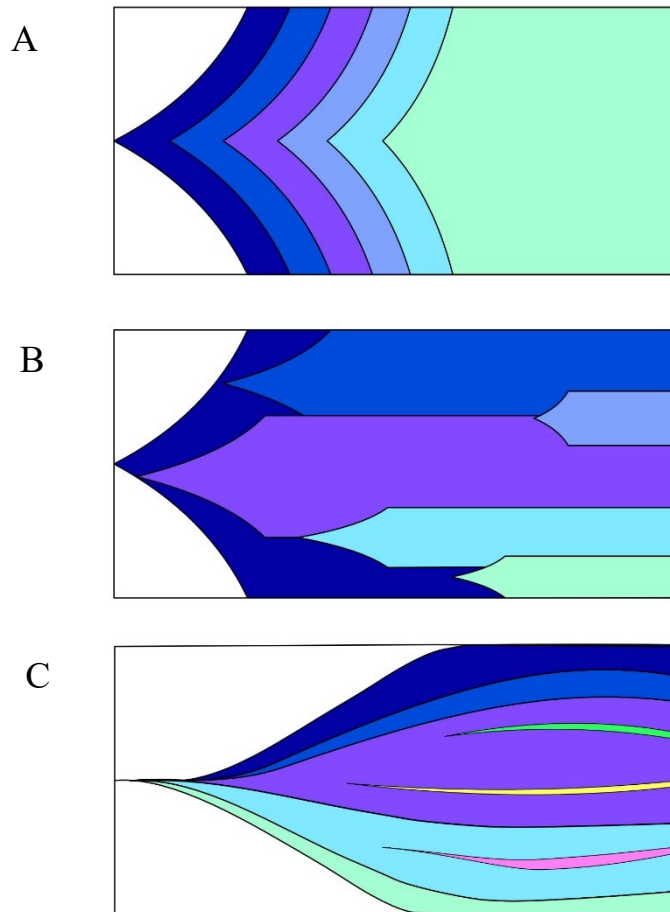


Figure 1.8 Opposing models of CRC evolution. A) Fearon & Vogelstein’s model of linear CRC evolution. The acquisition of driver mutations instigates the stepwise selection of fitter and more aggressive subclones with selective sweeps. There is always a dominant clone. B) The branched model of CRC evolution. Different sub-clones are able to grow and evolve alongside more dominant sub-clones. This results in ITH. C) The big bang model of CRC evolution. Beginning with a large expansion of sub-clones producing high levels of ITH. Each sub-clone harbours the same mutations as the initiating cell. Private mutations occur throughout development but persistence of these mutations is reliant on timing rather than selective advantage.

1.6.4. A sub-division of the branched model: The “just-right” and “mini-driver” models of CRC evolution

According to both the Vogelstein linear model of CRC evolution and through evidence from both FAP and AFAP, the TSG APC plays a crucial role in the progression of CRC. In 1999, a study by Lamlum *et al.* challenged the conventional view of APC as a typical TSG, which generally necessitates two LOF mutations for its classification as such. By examining 210 CRAs from both FAP and AFAP patients, the study focussed on the position and role of each germline and somatic APC mutation. Contrary to the expectation of the “two-hit hypothesis” of TSGs, the findings revealed that germline mutations in APC often influence the selection of site-specific mutations on the other allele. Notably,

mutations near codon 1,300 were found to confer the most significant selective advantage, leading to a pronounced tendency for allelic loss in the corresponding allele (Lamlum *et al.*, 1999).

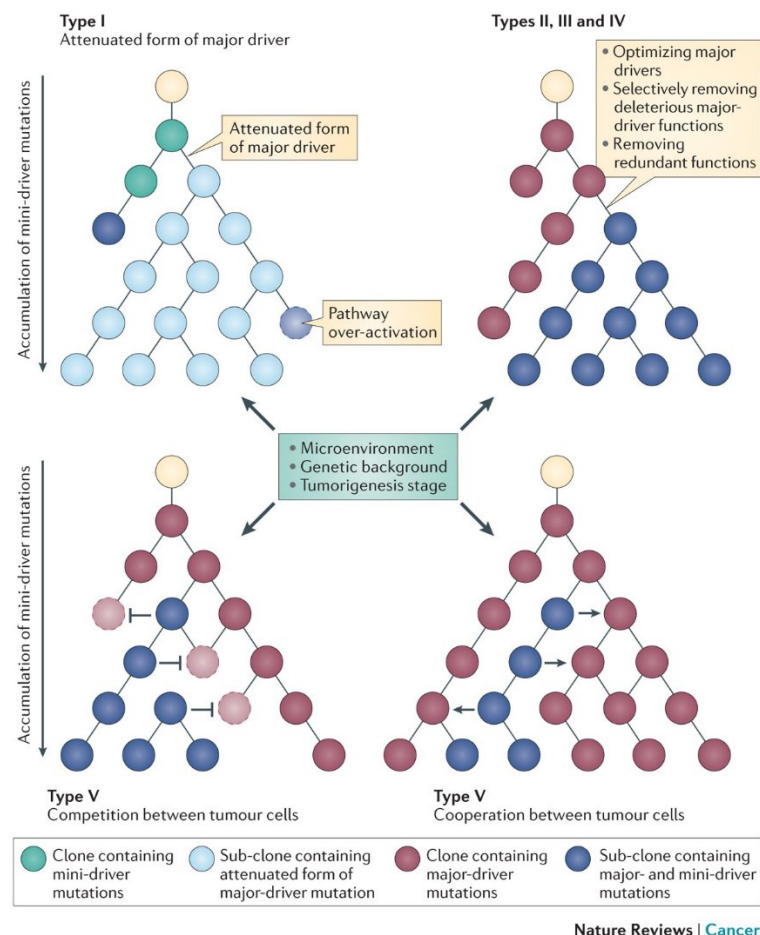
In 2002, the findings of Lamlum *et al.* were supported by a further study of *APC* in FAP patients, this time in the context of *APC*'s seven 20-AARs. Through the analysis of 133 CRAs of six unrelated FAP patients, Albuquerque *et al.* (2002) noted that when the original germline mutation resulted in loss of the seven downregulating 20-AARs, the somatic mutation that would occur in the corresponding WT allele would result in the maintenance of either one or two of the 20-AARs. Conversely, if the germline mutation resulted in a protein that retained one or two 20-AARs, the somatic mutation in the WT allele would remove all 20-AARs by either loss of heterozygosity (LOH) or truncating point mutation (Albuquerque *et al.*, 2002). It was therefore proposed that this association between mutations, and their resulting protein activity, leads to a “just-right” level of β -catenin signalling.

In 2006, Segditsas and Tomlinson suggested that the just-right model could potentially be extended to incorporate the rest of the Wnt pathway genes. According to the just-right model of CRC evolution, a just-right level of Wnt signalling can be achieved via mutations to Wnt pathway genes other than *APC*. A potential example involves activating mutations of *CTNNB1*, which codes for β -catenin, that result in its escape from proteasomal degradation. Interestingly, these activating β -catenin mutations appear to be mutually exclusive with *APC* mutations, suggesting that these mutations in isolation provide a just-right level of β -catenin signalling. An additional *APC* mutation may surpass the level of signalling that is sufficient for the cell and this combination therefore undergoes negative selection during tumour development (Segditsas and Tomlinson, 2006).

This “just-right” model has been expanded upon by Castro-Giner, Ratcliffe and Tomlinson (2015) through their proposal of the “mini-driver” model of CRC evolution. Mutations that occur within cancerous cells are currently categorised according to their contribution to cancer cell fitness, in a conveniently dichotomous fashion. Driver mutations are defined as mutations which confer a selective advantage, whilst passenger mutations have negligible overall effect on cancer cell fitness (Gerlinger *et al.*, 2014). Interestingly, Cross *et al.* (2018) reported relatively weak overall effects of some driver mutations on cancer cell fitness.

A solution to this categorisation issue was proposed by Castro-Giner, Ratcliffe and Tomlinson (2015), who suggested that the functionality of mutations falls along a continuous spectrum, encompassing mutations whose role lies between the traditional “driver or “passenger”. This model suggests that major, canonical driver mutations, such as *BRAF*^{V600E}, continuously and reliably provide the cell with a large fitness advantage. In comparison, there exists “mini-driver” mutations that provide the cell with a relatively small selective advantage. Mini-driver mutations are expected to be present in a small population of the tumour and, due to their relatively weak effects resulting in the absence of

selective sweeps, would be present in sub-clones. Their function would potentially optimise, fine-tune or amplify the effects of a major-driver mutation, but is not essential for tumour growth. Furthermore, a loss of a mutated mini-driver, for example if the mutation was reverted to WT, would not kill the tumour nor substantially reduce growth rate. This model therefore predicts that sub-clones harbouring mini-driver mutations would display parallel or convergent evolution between other sub-clones and cancers of the same type, allowing them to co-exist, evolve and persist alongside a range of sub-clonal major driver mutations (Castro-Giner, Ratcliffe and Tomlinson, 2015). Castro-Giner, Ratcliffe and Tomlinson (2015) also predicted that the presence of multiple mini-driver mutations within a sub-clone could in fact be a substitute for the pathway activation caused by the presence of a major driver mutation. Therefore, a just-right level of pathway activation may be achieved through an accumulation of non-canonical mini-driver mutations. Alternatively, a mini-driver mutation may in fact be able to fine-tune the amount of pathway activation caused by the presence of a major driver mutation (**Fig. 1.9**).



Nature Reviews | Cancer

Figure 1.9 Mini-driver scenarios as presented by Castro-Giner, Ratcliffe and Tomlinson (2015). In a type I scenario, mini-driver mutations provide a selective advantage to cells harbouring an attenuated form of a major-driver mutation. In types II, III and IV scenarios, mini-driver mutations optimise major-driver mutations (II). They remove suppressive side-effects on growth (III) or redundant functions of major-drivers (IV). In a type V scenario mini-driver mutations either result in death of sub-clones through competition or promote the growth of other clones.

An amalgamation of the just-right model and mini-driver model is therefore compatible with the branched model of CRC evolution, as together they postulate that atypical mini-driver mutations provide an alternative route to carcinogenesis. These non-canonical mini-driver mutated sub-clones may coexist and evolve in parallel to more dominant sub-clones.

1.6.5 A potential example of just-right, mini-driver selection: Atypical *BRAF* and *KRAS* mutations

A possible example of mini-driver selection in CRC involves the *MAPK/ERK* signalling pathway. This pathway is initially activated by ligand-binding of the epidermal growth factor receptor (EGFR). Binding of the ligand to the extracellular domain of EGFR results in its phosphorylation and subsequent activation of RAS, B-RAF, MEK and ERK. As a consequence, the activation of this pathway results in overall increased cell proliferation and survival (**Fig. 1.10**).

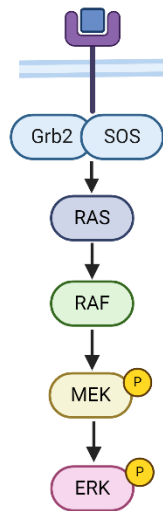


Figure 1.10 The *MAPK* signalling pathway. Activation of the *MAPK* signalling pathway, including the activation of B-RAF, leads to a cascade of events, resulting in increased cell proliferation and survival. Adapted from Ascierto et al. (2012).

B-RAF, coded for by the gene *BRAF*, as part of this cascade, acts as a serine/threonine protein kinase and when activated, phosphorylates MEK. Approximately 8% of CRCs harbour mutations at *BRAF*^{V600E}. This particular mutation is classified as a driver and class I *BRAF* mutation. The presence of *BRAF*^{V600E} causes the B-RAF protein to function as a constitutively active monomer, independent of *RAS* signalling. This in turn causes an increase in cellular proliferation and survival (Yao *et al.*, 2017; Zhang *et al.*, 2018). *BRAF* mutations outside of codon 600, occur less frequently. Mutations outside of codon 600 that do not act as constitutively active monomers are categorised into class II or class III *BRAF* mutations. Class II mutations result in a constitutively active dimer which also acts independently

of RAS. Class III mutations, such as *BRAF*^{D594G/N}, are kinase-dead or have decreased kinase activity compared with wild-type *BRAF*. Their action depends on RAS activation (Yao *et al.*, 2017; Fontana and Valeri, 2019).

A currently unpublished data analysis project carried out by the Tomlinson group has investigated canonical and non-canonical *BRAF* mutations in CRC, utilising various datasets including: Stratification in Colorectal Cancer (S:CORT), The Cancer Genome Atlas (TCGA) and the Genomics England 100,000 Genomes Project (100KGP). This study identified a significant enrichment of additional *RAS* pathway mutations such as *KRAS*, *NRAS*, *NF1* and *PIK3CA*, in tumours also harbouring a class III *BRAF* mutation, a potential mini-driver mutation. Often, in these cases, non-canonical *RAS* pathway mutations were favoured. In addition, a significant enrichment of class III *BRAF* mutations were identified in cancers of the rectum when compared with other locations in the large bowel. This finding may highlight a possible example of an alternative route to carcinogenesis in the rectum driven by an accumulation of mini-driver mutations, rather than the presence of a major driver such as *BRAF*^{V600E}.

Interestingly, clinical studies have shown that whilst response in patients with a class I or class II mutation in *BRAF* is limited, approximately 50% of patients harbouring a class III mutation in *BRAF* respond to treatment with cetuximab, an anti-EGFR based therapy (Yao *et al.*, 2017). The reason as to why only half of these patients respond to this treatment remains elusive. Previous studies into anti-EGFR therapy have validated that *KRAS* mutations are a biomarker for primary resistance to EGFR antibody treatment (Lièvre *et al.*, 2006). If the mini-driver model is therefore an accurate representation of CRC evolution, it is of great importance to study mutational patterns of co-occurrence, particularly in cancers with atypical mutations as additional mutations may alter responses to personalised treatments.

1.7 Aims and objectives

This thesis aims to investigate the selective constraints of mutations in colorectal cancer evolution. Specifically, the objectives of this thesis are to:

- Refine existing evolutionary models and explore evidence for both branched/parallel evolution and the just-right model in large, colorectal adenomatous polyps through WGS/WES of multiregional samples.
- Investigate patterns of co-occurrence and mutual exclusivity between mutations in the Wnt pathway genes, to further explore evidence for the just-right/mini-driver models of CRC evolution using a large CRC dataset (100,000 Genomes Project).
- Test the evidence for *TCF7L2* as a CRC risk gene through both a conditional analysis and eQTL analysis
- Assess the evidence for *TCF7L2* as a mini-driver using a heterozygous *Tcf7l2* knock-out mouse model.

Chapter 2

Assessing branched/parallel evolution and the just-right model through multiregional sampling of large colorectal adenomatous polyps

2.1 Background

As described in the previous chapter, NGS has highlighted the presence of ITH at genome level, in addition, it has shown that CRCs can develop without a full complement of driver mutations. This highlights significant downfalls of the classic linear model of CRC evolution (Cross *et al.*, 2018). Although the big bang model of CRC evolution utilised ITH as a central premise, recent studies have highlighted both the merits and downfalls of the big bang model. Sequencing of CRAs and CRCs in studies by both Niida *et al.*, (2021) and Cross *et al.*, (2018) provide evidence of high levels of ITH in early-stage colorectal tumours, in agreement with the big bang model. In addition, through the analysis of phylogenetic trees, Cross *et al.*, (2018) find evidence for weak or neutral evolution, following tumour initiation, in the branches/leaves of CRCs when compared to CRAs. They also show evidence of CNA and WGD events acting on the tumour in a punctuated fashion. In contrast, these studies also provide contradictory evidence to some of the main premises of the big bang model. Multiple studies have evidenced the presence of sub-clonal driver mutations in CRAs, highlighting the existence of sub-clones with unique driver mutations, able to co-exist and evolve alongside other clones. The occurrence of parallel evolution challenges the big bang model and its premise that all tumours are “born to be bad” as it shows that mutations capable of driving tumorigenesis are not homogenous across the tumour at an early stage. Notably, in addition to the identification of sub-clonal driver mutations in CRAs, the research conducted by Cross *et al.*, (2018) also presents further opposing evidence of the big bang model by demonstrating that in CRCs, large sub-clones continue to occupy spatially discrete areas. This finding contradicts the big bang model’s assertion that, after initial transformation, sub-clones radiate outwards uniformly, without preferential expansion or further selection of beneficial mutations.

As a resolution to this issue, Niida *et al.*, (2021) propose the action of Darwinian selection in early-stage tumours. Their findings are more accurately described by the branched model of CRC evolution. The branched model of CRC evolution adopts a more gradualist or polygenic approach, addressing several critical issues presented by the linear model and big bang model (Gerlinger *et al.*, 2014). The branched model posits that sub-clones have the ability to evolve in parallel to one another in spatially discrete areas. Similar to the big bang model, there is a lack of selective sweeps, however, it is proposed that differing driver mutations can exist and accumulate within separate sub-clones and evolve in coherence with one another. This pattern of evolution also results in a high level of ITH, opposing the linear model of CRC evolution (**Fig 2.1**). Sub-clonal driver mutations evidenced by Cross *et al.*, (2018) is supportive of the branched model.

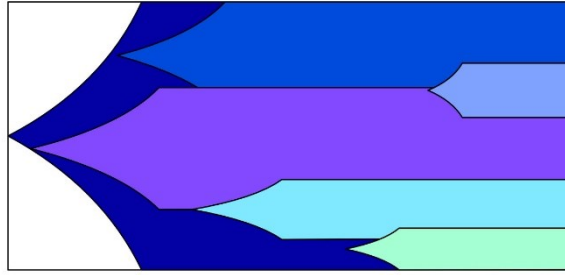


Figure 2.1 The predicted clonality of a tumour according to the branched model of CRC evolution.

Interestingly, through the sequencing of nine, late-stage colorectal tumours, Niida *et al.*, (2018) found driver mutations to be clonal and evidenced a lack of the occurrence of parallel evolution. This suggests that the mutations required to cause tumour progression occurred early in development, followed by transformation and a pattern of evolutionary neutrality. Supporting evidence was identified by Cross *et al.*, (2018) who evidenced that sub-clonal driver mutations were rarer in CRCs than CRAs, although they did not find this difference to be statistically significant. Cross *et al.*, (2018) also found evidence of a neutral evolutionary trajectory in CRCs through the phylogenetic analysis of CRAs and CRCs, in which CRAs evidenced short trunks and long branches as opposed to the long trunks and short branches of CRCs. Williams *et al.*, (2016) also concluded that the neutral power law fit 323 of 904 cancers of 14 different types. Niida *et al.*, (2021) propose that the difference in evolutionary trajectories between CRAs and CRCs could be explained by a transformative event such as a drastic chromosomal event such as copy number alterations (CNAs) and chromosome rearrangements. This punctuated event that happens in a short burst could provide a means of transformation and therefore cause progression of CRA to CRC. In support of this, both Niida *et al.*, (2021) and Cross *et al.*, (2018) evidence a significant increase in the number of CNAs in late-stage tumours compared to early-stage tumours. This suggestion is synonymous with the big bang model which suggests the occurrence of a significant and rapid transformation event, such as WGD, followed by a neutral evolutionary trajectory (Cross *et al.*, 2018; Niida *et al.*, 2021).

Despite the evidence that CRCs harbour more CNAs than CRAs, this does not necessarily equate to chromosomal events occurring in a punctuated fashion. The timing of these chromosomal events may more accurately be described as occurring in a stepwise fashion, further supporting the branched model of CRC evolution. It is therefore of great interest to provide further insight into the mechanism of CRC evolution to further refine and adjust both the branched and big bang models of CRC evolution.

The just-right model offers a compelling explanation for the genetic characteristics of both CRAs and CRCs observed by Niida *et al.*, (2021) and Cross *et al.*, (2018). This model suggests that tumour progression is contingent upon achieving an optimal level of pathway activation, induced by

varying mutations across different signalling pathways. The just-right model and branched model are similar in their premise of lack of selective sweeps and high predicted levels of ITH. The just-right model posits that sub-clones harbouring differing mutations evolve in spatially discrete areas. This therefore explains ITH in CRAs and is consistent with the findings of Cross *et al.*, (2018), that large sub-clones within CRCs can continually occupy spatially discrete areas. A just-right level of signalling could also be achieved via punctuation events. Moreover, the just-right model might elucidate the observed neutrality of evolution in CRCs, positing that once a tumour achieves this just-right activation state, maintaining it through neutral evolution becomes paramount (Segditsas and Tomlinson, 2006).

Though the investigation of FAP/AFAP-related *APC* mutations, Lamlum *et al.* (1999) instigated the formulation of the just-right model of CRC evolution. The “just-right” model of CRC evolution suggests that an advantageous combination of germline and somatic *APC* mutations are selected, resulting in the maintenance of at least one 20-AAR across both alleles. Evidence suggests, in CRC, if the first *APC* mutation occurs before R1, the mutation on the second allele usually occurs either just after R1 or after R2. In the case that the first mutation occurs between R1 and R2, the second allele is usually removed by LOH. Finally, in the case that the first mutation occurs after R2, the second event tends to be a point mutation and the occurrence of LOH decreases. Therefore, R1 is required for tumour development and an optimal number of AARs are selected. R2 cannot bind to β -catenin and R3 is prevented from binding to β -catenin through negative selection. Therefore, CRCs rely on β -catenin binding to R1 to control levels of β -catenin and transcription. This combination of first and second hits in *APC*, specifically in CRC, have been shown to result in intermediate Wnt signalling activity. Complete degradation of β -catenin by *APC* is counter selected in CRC by loss of the CID (Kohler *et al.*, 2009). Therefore, the selection of *APC* mutations has shown a tendency towards a just-right level of Wnt signalling allowing for efficient tumour progression (Lamlum *et al.*, 1999; Albuquerque *et al.*, 2002; Kohler *et al.*, 2009).

NGS has not only highlighted the presence of ITH at genome level, but also an unexpected lack of new drivers mutated at high frequency. This in turn highlights the large potential diversity of tumours that harbour alternative mutations at low or very low frequencies (Castro-Giner, Ratcliffe and Tomlinson, 2015). Further expanding on this, Castro-Giner, Ratcliffe and Tomlinson (2015) proposed the existence of mutations of small selective advantage. They coined the term for these mutations “mini-driver” mutations. Theoretically, a Wnt mini-driver mutation may have the ability to adjust the level of Wnt signalling to a just-right level, that may not be provided by the presence of a major Wnt pathway driver mutation alone. In which case, a pattern of co-occurrence would be visible between known, major Wnt pathway driver mutations in *APC*, *CTNNB1* and *RNF43* and Wnt mini-driver mutations. Similarly, an accumulation of mini-driver mutations in the Wnt pathway may have a similar overall effect on the tumour as a major-driver mutation. Hypothetically, Wnt pathway mini-drivers may have the ability to adjust Wnt signalling to account for a non-optimal selection of 20AARs in *APC*, in which case a co-

occurrence would be seen between *APC* mutations with a sub-optimal number of 20AARs and additional Wnt pathway mini-driver mutations.

Although there are functional models of early tumourigenesis e.g. *Apc*^(Min/+) mice, functional assessments of all mutations that occur in CRC are yet to be carried out. Thus, it is difficult to know if late arising mutations contribute to carcinogenesis or if their contribution is neutral. Evidence for the linear model, big bang model and branched model are therefore mostly based on the study of genetic data through NGS, mathematical modelling and phylogenetic analysis. Further studies of the just-right model and mini-driver models can be assessed functionally as carried out in chapter 5.

When CRAs are compared to CRCs, previous studies have shown an increase in ITH and genetic diversity in CRAs. Cross *et al.*, (2018) suggested that ITH in CRCs could be explained by evolutionary neutrality. In support of this finding, Saito *et al.* (2018), utilised multiregional sequencing to assess differences in ITH between polyps with a focus of cancer (referred to as precancerous lesions) and late-stage CRCs, providing further insight into the development of ITH throughout the course of tumourigenesis. Their results also showed an increase in ITH in these precancerous lesions compared to advanced CRCs and also suggest that the ITH observed in advanced CRCs could be explained by neutral evolution.

This chapter aims to further explore hypothetical mechanisms of CRC evolution through the genetic analysis of 19 large CRAs from 18 different patients. Multiregional samples (85 total) taken from large polyps (≥ 2 cm) were analysed via WES and WGS. Different genetic characteristics were explored such as driver mutation complement, ITH, CNAs, WGD and mutation burden. The methodology utilised in this chapter replicates the approach of Cross *et al.*, (2018) by exploring differences in genetics characteristics such as driver mutation complement, ITH, CNAs, WGD and mutation burden between CRAs and CRCs. The current methodology additionally focuses on the Wnt pathway to further refine current evolutionary models and assess evidence for parallel evolution and the just-right model.

Large polyps are of particular interest due to prior evidence suggesting that they are more likely to have a focus of cancer than smaller polyps (Rembacken *et al.*, 2000). Through pathological analyses, four of the polyps were found to have a focus of cancer. Genetic comparisons could therefore be made between those polyps with and without a focus of cancer, including the relationship of some of these characteristics to polyp size. This provides further insight into the genetic characteristics of tumours throughout early tumour development and refinement of different models of CRC evolution.

2.2 Chapter Aims

- Investigate and, if appropriate, refine existing models of CRC evolution, including the linear, big bang and branched models through an assessment of ITH and driver mutation ITH in large colorectal adenomatous polyps.
- Analyse mutations in Wnt-pathway related genes in large adenomatous polyps, evaluating support for the just-right model in CRC evolution aiding differentiation between functional and non-functional ITH.
- Compare levels of ITH and phylogenetic structure in adenomatous polyps with and without a focus of cancer, providing insight into ITH in the developmental stages of tumourigenesis.
- Assess the mutation burden of adenomatous polyps with and without a focus of cancer.
- Assess the frequency of CNAs and WGD events in adenomatous polyps with and without a focus of cancer and so investigate the potential for punctuated events that could signify the transition to a cancer.
- Assess the relationship of different genetic characteristics to polyp size to explore why large polyps are reported as more likely to transition to cancer.

2.3 Methods

For transparency purposes, Connor Woolley utilised five of these patient sample sets as part of his PhD project. This dataset has now increased to 19 patient sample sets which I utilise as part of this project. For consistency purposes, I have replicated some of the methods that Connor Woolley utilised, particularly DNA/RNA extraction methods, sequencing methods and data-processing methods, as many of these are standard methods described in Connor's previously submitted PhD thesis and some assistance was provided with these methods by Connor Woolley.

2.3.1 Sample collection

Multiregional, biopsies from large colorectal polyps were obtained from individuals undergoing surgical resection or transanal endoscopic microsurgery (TEMS) at the Western General Hospital, Edinburgh. A polyp was classified as large if its diameter was over 1cm. Such surgeries were conducted by Malcolm Dunlop, Doug Speake and Farhat Din. Ethical approval was obtained from the South East Scotland Research Ethics Committee (REC 11/SS/0109). Patient consent and enrolment into the SOCCS3 study was carried out by Research nurse Donna Markie and Doug Speake. Multiregional biopsies of the polyps were taken either by the surgeon or by myself (**Fig. 2.2**). A total of 85 multiregional samples were dissected from 18 patients (19 large adenomatous polyps, samples were taken from one polyp in the rectum and one polyp in the caecum of a single patient). Following dissection, the histopathology reports stated that four of these large adenomatous polyps contained regions of adenocarcinoma, based on routine diagnostic samples. Of those polyps reported to harbour a focus of cancer, it was not possible to determine which of the multiregional samples contained cancerous tissue. The entire polyp was therefore referred to as containing a focus of cancer, rather than a sample. MD20980 had no evidence of adenocarcinoma at initial surgery however presented with a caecal adenocarcinoma at follow up, three months after initial surgery (**Table 2.1**).

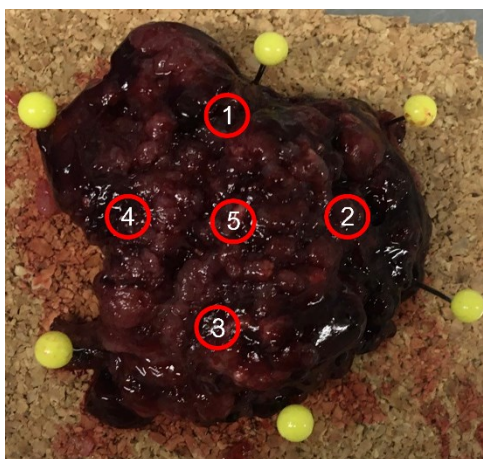


Figure 2.2 Multiregional sampling polyp dissection method. Shown is the dissection map of multiregional samples taken from a large polyp.

Table 2.1 A description of each large adenomatous colorectal polyp and patient. Information is provided for patient characteristics of type of presentation, sex, age at sampling and follow up data. Adenomatous polyp information is provided including size, location, grade of dysplasia, foci of invasion and MSI status. TEMS = Transanal endoscopic microsurgery. MSI = Microsatellite instability. MSS = microsatellite stability.

ID	Sex	Age at sampling	Tumour location	Polyp size (mm)	Surgery	Dysplasia	Type	Focus of invasion	Follow up	# samples	MSI status
MD20716	F	50	Rectosigmoid junction	40	TEMS	Low	Positive bowel screening	No	No cancer, villous adenoma, low grade dysplasia	4	MSS
MD21031	M	75	Rectum	47	TEMS	Low	Symptomatic referral	No	Died 2022 (COVID)	4	MSS
MD21032	M	59	Rectum	38 (biggest of fragments)	TEMS	Low	Symptomatic referral	No	Site check clear (2021)	5	MSS
MD20980	M	67	Rectum	55	TEMS	High	Symptomatic referral	No	CRC at previous TEMS site. Sigmoid tubulovillous adenoma with focal high-grade dysplasia (2021)	5	MSS

MD21099	F	58	Rectum	35	TEMS	Low	Abnormal MRI	No	Nil, remains on waiting list for check colonoscopy	5	MSS
MD20430	M	71	Rectum	45		High	Previous rectal polyp cancer, 2 years prior	No	NA	2	MSS
MD20730	F	57	Small bowel	20	Small bowel resection	NA	FAP	No	No colon remains	4	MSI
MD21236	F	61	Rectum	45	TEMS	High	Recurrence at previous TEMS site	Not definitive	Site check clear	5	MSS
MD21274	M	83	Rectum	46	TEMS	Low	Recurrence at previous TEMS site	No	Died 2021	5	MSS
MD21296	M	78	Rectum	40	TEMS	Low	Polyp surveillance	No	Site check clear	5	MSS
MD21320_1	F	66	Ascending colon	30	Proctocolectomy	Widespread low, focally high	Lynch syndrome (MSH6 germline mutation)	No	No colon remains	5	MSI

MD21320_2	F	66	Rectum	30	Proctocolectomy	Low	Lynch syndrome (MSH6 germline mutation)	“May be non-invasive”	No colon remains	4	MSI
MD21365	M	60	Rectum	49	TEMS	Low	Symptomatic referral	No	Being followed up at the borders	5	MSS
MD21386	F	71	Caecum	27	TEMS	Low	Symptomatic referral	No	Discharged	5	MSS
MD21239	M	64	Rectum	30	TEMS	Adenocarcinoma	Symptomatic referral	Yes	2022 site check – hyperplastic polyp 2023 – metastatic pelvic lymph node	5	MSS
MD21257	F	36	Rectum	90	TEMS	Adenocarcinoma	Symptomatic referral	Yes	2021 – two small polyps, low grade dysplasia	5	MSS
MD21340	M	66	Proximal ascending colon	30		High	Positive bowel screening	No	2021 – right hemicolectomy	2	MSS

									showed adenocarcinoma of caecum		
MD21353	M	84	Rectum	46	TEMS	Adenocarcinoma	Symptomatic referral	Yes	Site check clear	5	MSS
MD21354	M	56	Small bowel (end ileostomy)	65	Small bowel resection	Adenocarcinoma	FAP, previous proctocolectomy and ileostomy	Yes	2022 – further revision of ileostomy, no recurrence	5	MSS

2.3.2 DNA/RNA isolation from polyp tissue

Immediately following dissection, polyp samples were placed on dry ice before being stored at -80°C. The IGC histology service embedded each sample in optimal cutting temperature compound (OCT) following which sections were produced and stained with Haematoxylin and Eosin (H&E) using standard techniques). Samples were returned to -80°C storage before further use. OCT was manually removed from the sample and the sample was placed into a 600 µL mix of β ME and RLT plus according to the Qiagen Allprep DNA/RNA Minikit protocol. A TissueLyser was then utilised at maximum speed for 1-2 minutes to disrupt and homogenise the tissue. DNA and RNA were then extracted from each sample utilising the Qiagen Allprep DNA/RNA Minikit according to the protocol. DNA and RNA were quantified using Qubit Fluorometric Quantification (Qubit™), dsDNA Broad Range Assay Kit. DNA samples were stored at -20°C and RNA was stored at -80°C. Polyp samples had either matched normal tissue samples or blood samples. DNA from blood was extracted by IGC facilities via standard protocols.

2.3.3 Whole genome and whole exome sequencing

In 2019, five sets of multiregional polyp samples, taken from five polyps, were submitted to *Novogene. Co., Ltd* for WGS. This was conducted on the Illumina platform at a depth of 50X with 150bp paired-end reads. $\geq 1\mu\text{g}$ of DNA per sample was prepared for WGS. In 2023, following the collection of additional samples, a second set of polyp samples was sent for both WGS and WES. Multiregional samples corresponding to polyps identified by pathology to have cancerous foci underwent WGS, while the remaining multiregional polyp samples underwent WES. In the 2023 batch, WGS was performed at a depth of 100X on tumour samples and 30X on normal samples. WES was conducted at a depth of 70X on both tumour and normal samples. $\geq 1\mu\text{g}$ of DNA per sample was prepared for WGS and $\geq 500\text{ng}$ of DNA per sample was prepared for WES. Additionally, mRNA sequencing was carried out at 40 million reads, on all samples.

2.3.4 Data preparation

Connor Woolley performed key parts of the data preparation and processing steps replicating the methodology previously published in his PhD thesis. The following processing steps were performed by Connor Woolley:

Down-sampling was performed on all BAM files to address differences in read depth between WGS and WES. Furthermore, reads were restricted to those specified by the Agilent SureSelect V6 Hg38 probeset (covered.bed) to allow for appropriate comparisons to be made between variants called by WGS and WES. Sub-setting and down-sampling were carried out using Samtools -L (v1.16.1) (Danecek *et al.*, 2021). The resulting data was processed utilising the bcbio-nextgen 1.2.4 pipeline. Alignments to hg38 was carried out utilising the Burrows-Wheeler aligner (H.Li & Durbin, 2009).

Somatic variant calling was carried out using Mutect2 (McKenna et al, 2010), Strelka2 (S.Kim et al, 2018), Vardict (Lai et al, 2016). Ensembl was then used to create and annotate consensus VCFs (McLaren *et al.*, 2016). Germline variant calling was carried out using freebayes (Garrison & Marth, 2012), Strelka2 (S. Kim et al, 2018) and gatk-haplotype (Poplin et al, 2018). Structural variants were evaluated via LUMPY (Layer, Chiang, Quinlan & Hall, 2014), DELLY (Rausch et al, 2012) and MANTA ((Chen, Schulz-Trieglaff, Shaw, Barnes, Schlesinger, Källberg, Anthony J. Cox, *et al.*, 2016).

I personally carried out MSI status analysis of each polyp using MANTIS and performed copy number calling on each sample using FACETS (Shen and Seshan, 2016). FACETS produced numerous missing values including those relating to cell fraction, total copy number and lesser copy number as well as information on WGD. According to the documentation, this can happen due to a lack of reads in particular regions or a lack of heterozygous SNPs which FACETS relies upon for assessing copy number. To rectify this, I had firstly opted to exclude any copy number calls displaying a total copy number of “2” which were combined with a missing lesser copy number as it remained unclear whether these samples are diploid or whether they had undergone CN-LOH (n=220). Connor Woolley then attempted using ASCAT for copy number calling which produced much more reliable results which were used to do further analysis. ASCAT also provided data on sample purity, ploidy and WGD.

2.3.5 Data analysis

The Cancer-Related Analysis of Variant Toolkit (Cravat) was utilised to determine overall SNA/indel burden of each sample (Masica *et al.*, 2017).

Mutations in 76 CRC driver genes identified by Intogen, and Wnt genes of interest, present in each polyp sample, were identified. Wnt pathway major-driver genes *APC*, *CTNNB1* and *RNF43* and mini-driver genes *AMER1*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *FBXW7*, *JUN*, *SOX9*, *TCF7L2* and *ZNRF3* were chosen based on their definitions as so in previous literature e.g. Brunet Guasch et al. (* currently unpublished, in which I am a contributing author). A list of Intogen CRC driver genes can be found in Appendix 2 (Martínez-Jiménez *et al.*, 2020). Mutations in both Intogen CRC driver genes and Wnt genes of interest were filtered to moderate or high predicted impact according to Ensembl (Harrison *et al.*, 2024) using Bcftools (Li, 2011). Some mutations in Intogen CRC driver genes were classified as pathogenic driver mutations by Intogen. Each missense mutation present in the polyp samples were assessed by checking both Intogen pathogenicity ratings and REVEL scores (Ioannidis *et al.*, 2016). Truncation mutations including stop gained and frameshift mutations were classed as pathogenic. Where Intogen pathogenicity scores were available, all truncation mutations were checked for pathogenicity. Splice site mutations were assessed using SpliceAI (de Sainte Agathe *et al.*, 2023).

Mutations in *APC* and *CTNNB1* were treated as exceptions to these rules. Only stop mutations gained and frameshift truncating mutations were considered pathogenic in *APC*. Truncation mutations

occurring downstream of the SAMP repeat (AAR position 1569) were excluded. In *CTNNB1*, only missense mutations in the hotspot region of codon 32 to 45 were included (Zaidi *et al.*, 2020). Stop gained mutations were not classified as pathogenic. No frameshift or splice site mutations were identified.

Any Intogen driver mutations that were present in the majority of multiregional samples with an exception of one, were checked again using BAM files in IGV browser.

All statistical tests carried out in this section are performed using Graphpad Prism (Graphpad Software, www.graphpad.com).

2.3.6 Phylogenetic trees

Firstly, Connor Woolley created nexus files enlisting mutations present in each of the multiregional samples. I then inputted the nexus files into PAUP (Phylogenetic analysis using parsimony) to carry out phylogenetic analysis (D.L. Swofford, 2003). The trees were rooted to a pseudonormal sample (pN) harbouring zero mutations to orient the tree topologies appropriately. The robustness of the inferred trees was assessed through bootstrap resampling with 10,000 replicates. A full heuristic search was employed to explore the tree space, utilising tree-bisection-reconnection (TBR) branch swapping algorithm. As a result, phylograms were produced with truncal lengths representing the number of mutations shared by each multiregional sample, branch lengths representing the number of mutations shared by more than one sample and leaf lengths representing the number of mutations present that were unique to each sample. The .tre files were then inputted into affinity designer for visualisation formatting purposes before annotating with Intogen colorectal cancer driver mutations and Wnt pathway mutations (Affinity designer software: affinity.serif.com).

2.3.7 Just-right model analysis

A recent study by Brunet *et al.*, (currently unpublished*), to which I am a contributing author, utilised 2023 CRC samples, from the 100KGP, to assess the number of 20AARs remaining in *APC* in samples harbouring two pathogenic *APC* mutations. The methodology used to assess the number of 20AARs in *APC* is replicated in this chapter. As stated above, only stop mutations gained and frameshift truncation mutations were considered pathogenic. Truncation mutations occurring downstream of the SAMP repeat (AAR position 1569) were excluded. Any samples with only one clonal mutation and no clonal CNA were also excluded. WGD events were assumed to occur after *APC* inactivation. There was one case of CN-LOH in *APC*, where there was also presence of a mutation in region 2, the number of 20AARs remaining on this allele was therefore duplicated. In addition, there was an example of LOH in which the major allele was amplified to three copies. This case the number of 20AARs on this allele was multiplied by 3. The position of each stop gained and frameshift mutation was placed into one of four categories depending on the position of the mutation in regards to *APC*'s amino acids. The number

of 20AARs retained due to presence of a truncation mutation at this amino acid position, was then noted (Table 2.1).

Table 2.2 The distribution of pathogenic mutations in *APC* and the corresponding number of 20AAR repeats retained. Adapted from Brunet *et al.* (*).

Region	Amino acid position	20AARs retained
R ₀	0-1256	0
R ₁	1256-1370	1
R ₂	1370-1486	2
R ₃	1486-1569	3

2.4 Results

2.4.1 Phylogenetic trees

Parsimony phylogenetic trees were produced to assess ITH and further refine different models of colorectal cancer evolution based on the clonality of Intogen colorectal cancer drivers. Wnt pathway mutations are also annotated onto the trees to assess evidence for the just-right/mini-driver models of CRC evolution. Evidence for the just-right model was assessed in polyps with two-hits in *APC*.

2.4.2.1 MD20716

Driver ITH is identified in this polyp in the form of *BCOR* and *FAT4* mutations in sample 6. The *FAT4* mutation was not pathogenic according to REVEL annotation.

Supportive evidence for the just-right model is demonstrated in this polyp through the alteration of both *APC* alleles. A clonal *APC*^{Glu1284Ter} mutation in region 1 results in the maintenance of one 20AAR. A copy number change is also observed, resulting in a 3-0 (major-minor) copy number change. As a result, this polyp contains three 20AARs in *APC*. A clonal *FBXW7* mutation is also present in combination with these clonal alterations to *APC*. (**Fig. 2.3**).

Pathogenic mutations evidenced in this polyp include: *KRAS*^{Gly12Val}, *FBXW7*^{Thr385Ala}, *APC*^{Glu1284Ter} and *BCOR*^{Phe1639Cys}.

Table 2.3 Genetic profile summary for polyp MD20716. This table details the genetic characteristics of MD20716 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	No	No	3	<i>FBXW7</i>	96-97	No

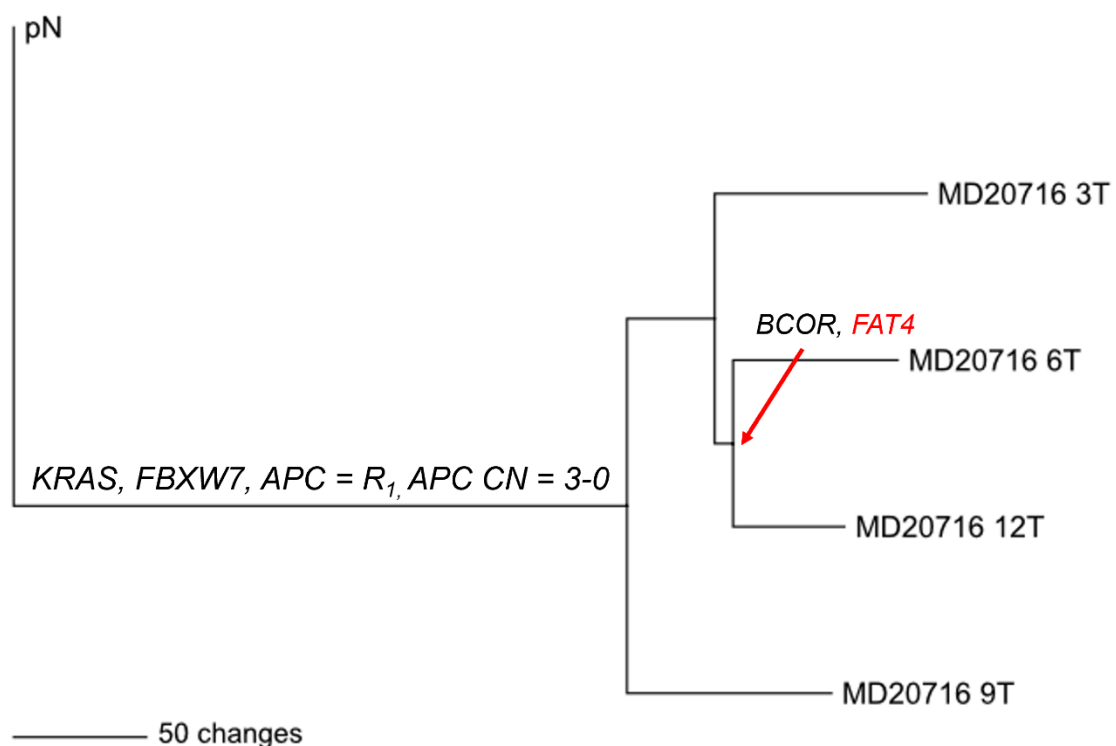


Figure 2.3 Phylogenetic tree of MD20716. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. CN = copy number (major-minor allele frequency). Red font = not pathogenic.

2.4.2.2 MD20430

Driver ITH is demonstrated in this polyp through the presence of sub-clonal mutations in *NCOR2*, *KRAS*, *SOX9* and *FBXW7*^{Ser583Leu}. WGD is present in sample 1 alone.

Only one clonal alteration in *APC* is detectable in this polyp in the form of a mutation in region 0 and so the classic just-right model analysis involving two *APC* hits cannot be applied to this polyp.

Both *APC* alleles are duplicated in sample 1 alone, likely due to WGD. The clonal *APC* mutation is observed in combination with a clonal *FBXW7*^{Arg278Ter} mutation. It is also observed in combination with *SOX9* and *FBXW7*^{Ser583Leu} mutations in sample 2 alone (**Fig. 2.4**).

Pathogenic mutations evidenced in this polyp include: *ERBB3*^{Arg475Trp}, *NCOR2*^{Gly1572Asp}, *PIK3CA*^{Val344Gly}, *FBXW7*^{Arg278Ter}, *SMARCA4*^{Thr910Met}, *APC*^{Ile638TyrfsTer8}, *KRAS*^{Lys117Asn}, *GNAS*^{Arg844His}, *SOX9*^{Thr314HisfsTer264}.

Table 2.4 Genetic profile summary for polyp MD20430. This table details the genetic characteristics of MD20430 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	No	Sample 1 only	NA	<i>FBXW7</i> (x2), <i>SOX9</i>	48-91	No

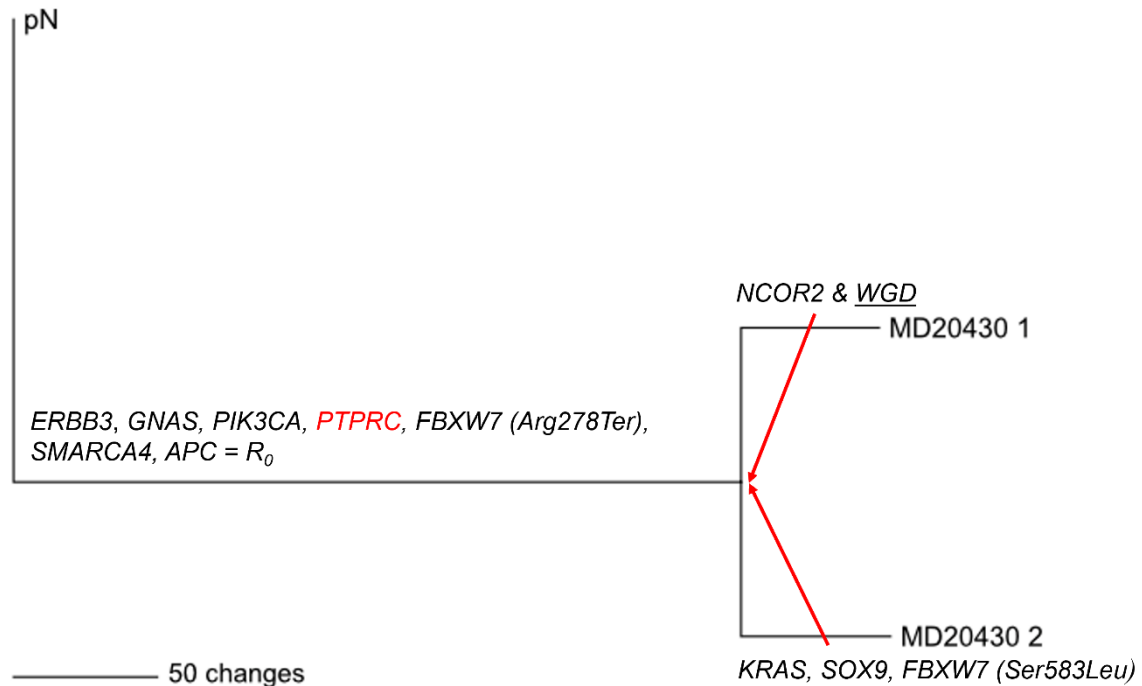


Figure 2.4 Phylogenetic tree of MD20430. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. CN = copy number (major-minor allele frequency). WGD = whole genome duplication. Red font = not pathogenic.

2.4.2.3 MD20730

This patient has a diagnosis of FAP. Interestingly, all samples were identified as MSI, which is unusual in FAP but has been shown to happen in previous literature (Sai Ge *et al.*, 2022). Samples 3 and 4 had lower purity percentages at 20 and 21%. The only driver mutation identified was a *KRAS*^{Gln61His} mutation in samples 3 and 4 (**Fig. 2.5**).

Again, there was no detectable somatic *APC* alteration so the classic just-right model analysis involving two *APC* hits could not be applied to this polyp.

Table 2.5 Genetic profile summary for polyp MD20730. This table details the genetic characteristics of MD20730 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	No	No	NA	NA	20-100	No

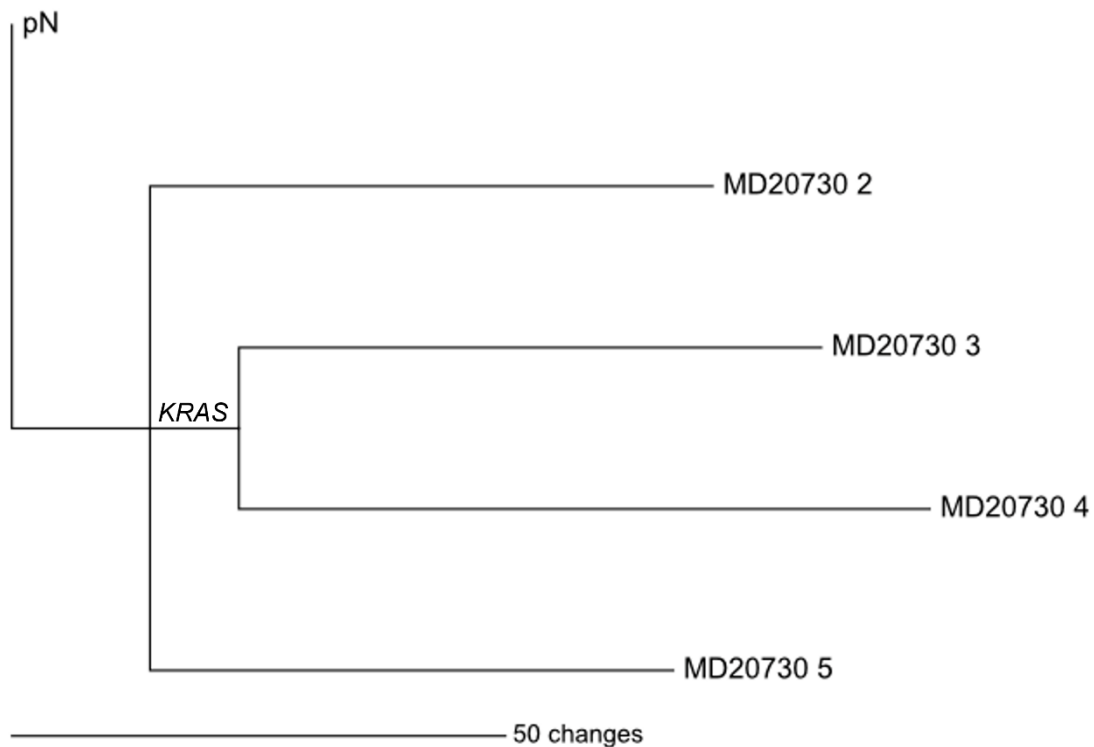


Figure 2.5 Phylogenetic tree of MD20730. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal.

2.4.2.4 MD21031

Driver ITH was shown in this polyp due to the presence of the same *GNAS*^{Arg844His} mutation across three samples, a *PCBP1*^{Leu100Gln} mutation in samples 1 and 2 and a *TCF7L2* mutation in sample 3. Absence of the *GNAS*^{Arg844His} mutation in sample 4 was confirmed via investigation of the .bam files in IGV browser.

Two clonal *APC* mutations in R₀ and R₁ result in the total maintenance of one 20AAR. These *APC* mutations occur in combination with a clonal *SOX9* mutation and a *TCF7L2* mutation in sample 3 (**Fig. 2.6**).

Pathogenic mutations observed in this polyp include: *SOX9*^{Gln411Ter}, *APC*^{Ser1100LeufsTer26}, *APC*^{Ile1287Ter}, *KRAS*^{Gly12Val}, *MAP3K1*^{Pro195Ser}, *PCBP1*^{Leu100Gln}, *GNAS*^{Arg844His} and *TCF7L2*^{Arg420Trp}.

Table 2.6 Genetic profile summary for polyp MD21031. This table details the genetic characteristics of MD21031 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN- LOH)	WGD	<i>APC</i> 20AARs	Wnt mini- drivers	Purity range (%)	Focus of cancer
No	No	No	1	<i>SOX9</i> , <i>TCF7L2</i>	20-100	No

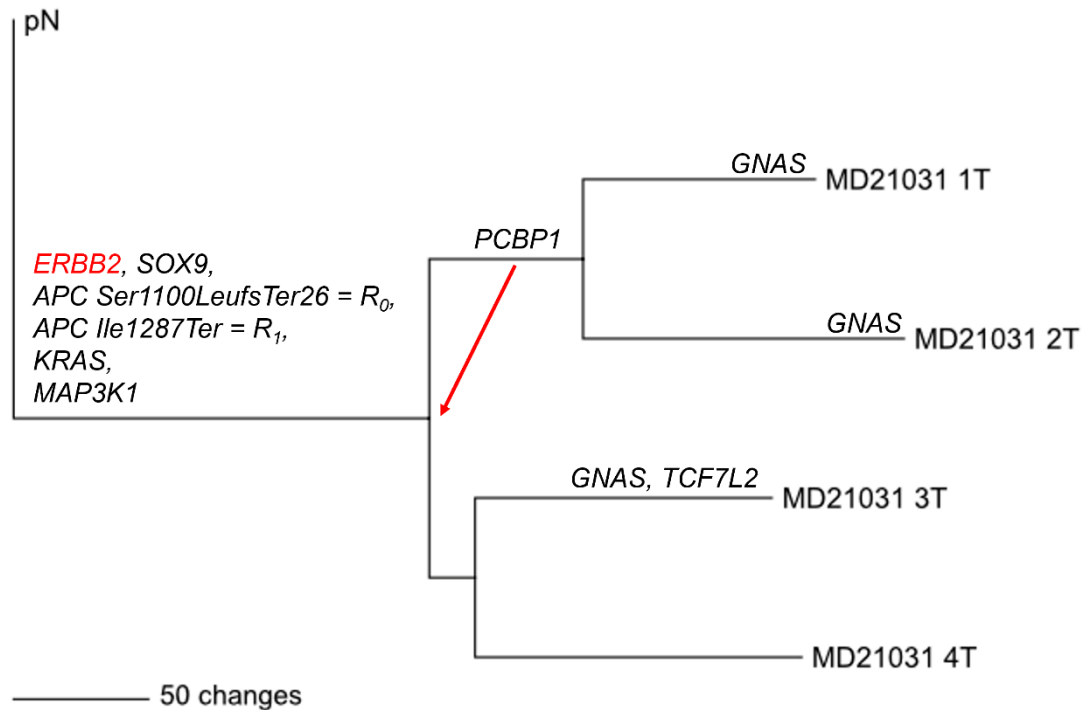


Figure 2.6 Phylogenetic tree of MD21031. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. Red font = not pathogenic.

2.4.3.5 MD20980

Driver ITH was again shown as present in this polyp with both *KMT2C* and *FBXW7* mutations in different samples. An example of convergent evolution is also shown in this polyp as sample 5 harboured a *GNAS*^{Arg844Cys} mutation, whereas sample 3 harboured a *GNAS*^{Arg844His} mutation.

Only one clonal alteration to *APC* was identified, so this polyp so the classic just-right model analysis involving two *APC* hits could not be applied. The *APC* mutation occurred alongside a mutation in Wnt mini-driver *FBXW7*. Notably, a mutation in *CTNNB1* was identified but this was not classed as pathogenic, therefore pathogenic mutations in the two key Wnt major-driver genes *APC* and *CTNNB1* remain mutually exclusive (**Fig. 2.7**).

Pathogenic mutations observed in this polyp include: *APC*^{Gln1291Ter}, *KRAS*^{Gly12Asp}, *GNAS*^{Arg844Cys}, *FBXW7*^{Ala2421Thr}, *KMT2C*^{Tyr1395Ter}.

Table 2.7 Genetic profile summary for polyp MD20980. This table details the genetic characteristics of MD20980 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	No	No	NA	<i>FBXW7</i>	85-95	No

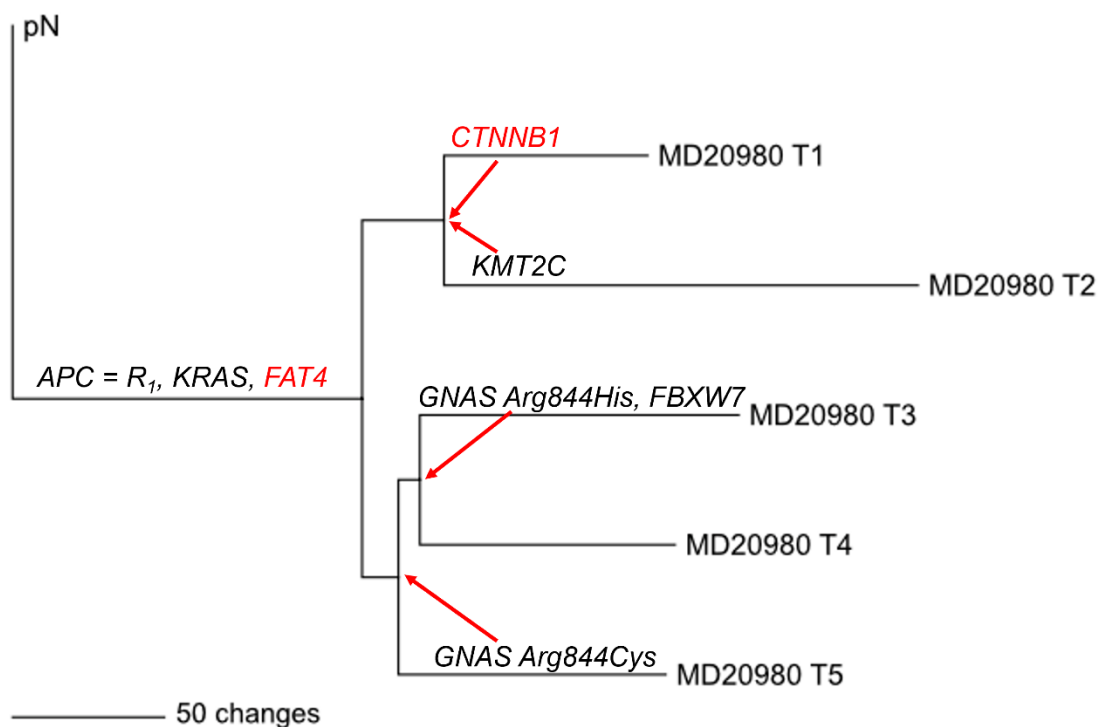


Figure 2.7 Phylogenetic tree of MD20980. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. Red font = not pathogenic.

2.4.3.6 MD21032

Driver ITH is presented in this polyp with the presence of an *ATM* mutation in sample 1, a *FAT4* and an *APC*^{Leu954Ter} mutation in sample 4 and a *TP53* mutation in sample 5. *ITH* is also shown in loss of a *TP53* allele in sample 2, 3, 4 and 5 only.

Only one clonal alteration to *APC* was detectable in this polyp so the classic just-right model analysis involving two *APC* alterations could not be applied. Interestingly, sample 4 harbours an additional *APC* mutation in region 0. As *APC* is a tumour suppressor gene, requiring two-hits for inactivation, this suggests that the polyp should have two clonal alterations in *APC*, further suggesting that the second clonal alteration in *APC* is undetectable through either variant calling errors or through epigenetic silencing of *APC*. If so, this would mean that sample 4 has a third hit in *APC*. However, the presence of another *APC* mutation in sample 4, could also represent a region of the polyp that developed through an alternative mechanism of *APC* inactivation, resulting in the maintenance of three 20AARs. In conclusion there is a lack of clarity regarding *APC* inactivation in this polyp making it difficult to accurately assess the just-right model. No other mutation in a Wnt mini-driver gene is present (**Fig. 2.8**).

Pathogenic mutations present in this polyp include: *KRAS*^{Gly12Asp}, *GNAS*^{Arg844His}, *APC*^{Thr1556AsnfsTer3}, *ATM*^{His2552Arg}, *FAT4*^{Arg126Trp}, *APC*^{Leu954Ter}.

Table 2.8 Genetic profile summary for polyp MD21032. This table details the genetic characteristics of MD21032 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	Monosomy: samples 2, 3, 4 & 5	No	NA	NA	89-94	No

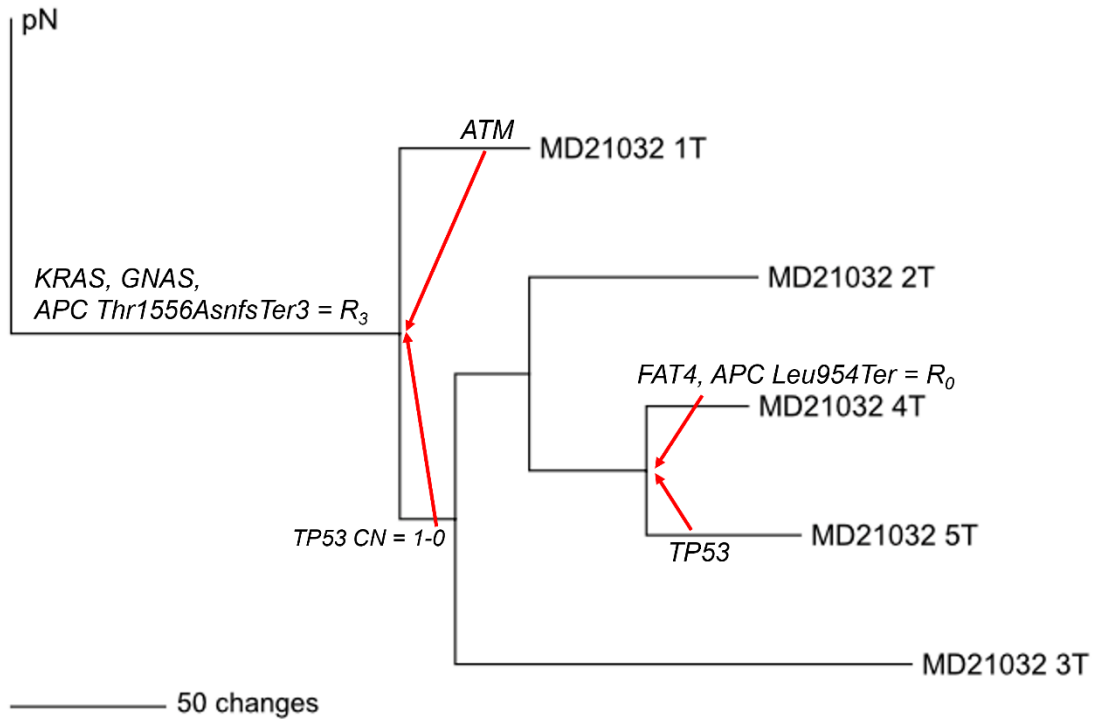


Figure 2.8 Phylogenetic tree of MD21032. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. CN = copy number (major-minor allele frequency). Red font = not pathogenic.

2.4.3.7 MD21099

Driver ITH is shown throughout this polyp. *CARD11* and an *APC* mutation are in all samples except sample 3. A *KRAS* mutation is found in samples 2, 4 and 5. A *GNAS* mutation was found in samples 1 and 5. The same *AXIN2* mutation was found in both samples 4 and 5, and an *FBXW7* mutation in sample 4. The purity of sample 1 was 54%, the purity of all other samples was above 92% (**Fig. 2.9**).

APC has only received one hit in this polyp in R_1 and so the classic just-right model involving 2 hits to *APC* cannot be assessed, although the *APC* mutation is found in occurrence with mutations in mini-drivers *FBXW7*, *TC7L2* and *AXIN2*.

Pathogenic mutations observed in this polyp include: *CARD11*^{Arg331His}, *APC*^{Gln1294Ter}, *KRAS*^{Gly13Asp}, *GNAS*^{Arg844His}, *FAT4*^{Arg2289Ter}, *FBXW7*^{Arg505Gly}, *AXIN2*^{Cys9TrpfsTer11}, *TCF7L2*^{Glu51Ter}.

Table 2.9 Genetic profile summary for polyp MD21099. This table details the genetic characteristics of MD21099 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	No	No	NA	<i>FBXW7</i> , <i>AXIN2</i> , <i>TCF7L2</i>	54-100	No

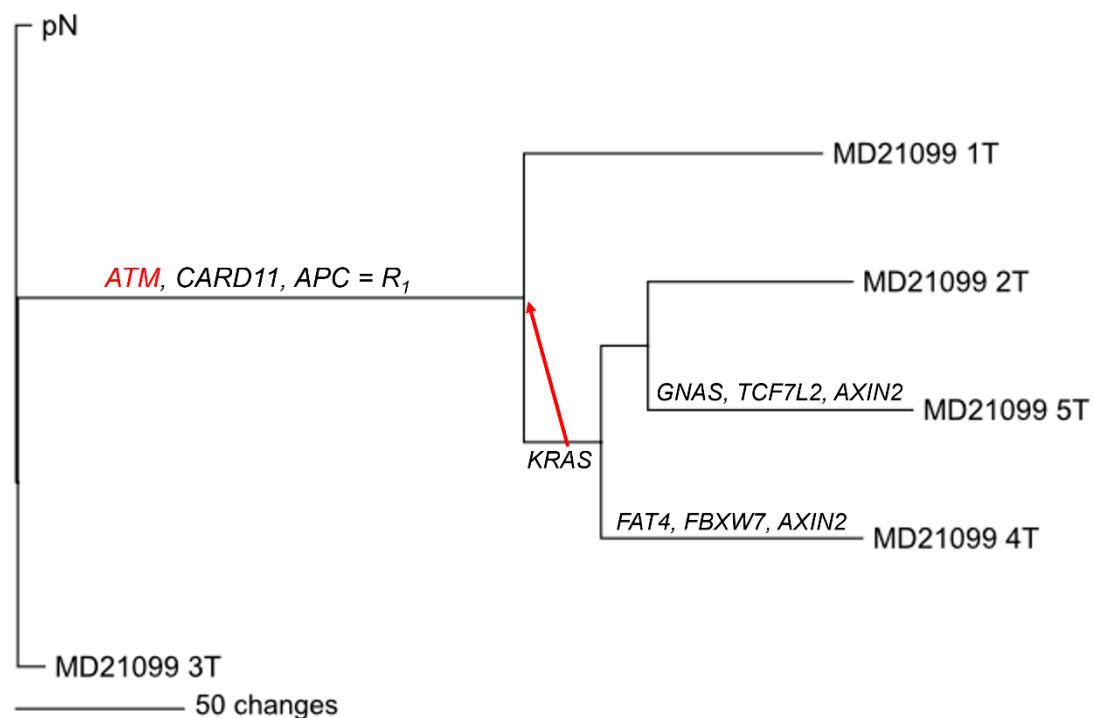


Figure 2.9 Phylogenetic tree of MD21099. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. Red font = not pathogenic.

2.4.3.8 MD21236

Driver ITH was again identified in this polyp. The same *KRAS* mutation was found in all samples except sample 2. A *GRIN2A* mutation was identified in sample 4 and two different *PIK3CA* mutations were identified in sample 1 alone (**Fig. 2.10**).

This polyp had two clonal hits in *APC* in the form of CN-LOH and mutation in R₂. This mutation in R₂ results in the maintenance of two 20AARs however, the presence of CN-LOH suggests this allele was duplicated and so the total number of 20AARs is four. This mutation in *APC* was found in combination with a mutation in Wnt mini-driver gene *TCF7L2*. Notably, a clonal mutation was identified in *CTNNB1*, however showed no evidence of pathogenicity, again maintaining the mutually exclusive relationship between pathogenic mutations in Wnt major-driver genes *APC* and *CTNNB1*.

Pathogenic mutations observed in this polyp include: *TCF7L2*^{Thr282HisfsTer46}, *APC*^{Arg1450Ter}, *KRAS*^{Gly12Asp}, *GRIN2A*^{Asp813Asn}, *PIK3CA*^{His1047Arg}.

Table 2.10 Genetic profile summary for polyp MD21236. This table details the genetic characteristics of MD21236 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN- LOH)	WGD	<i>APC</i> 20AARs	Wnt mini- drivers	Purity range (%)	Focus of cancer
Yes: clonal	No	No	4	<i>TCF7L2</i>	68-82	Not definitive

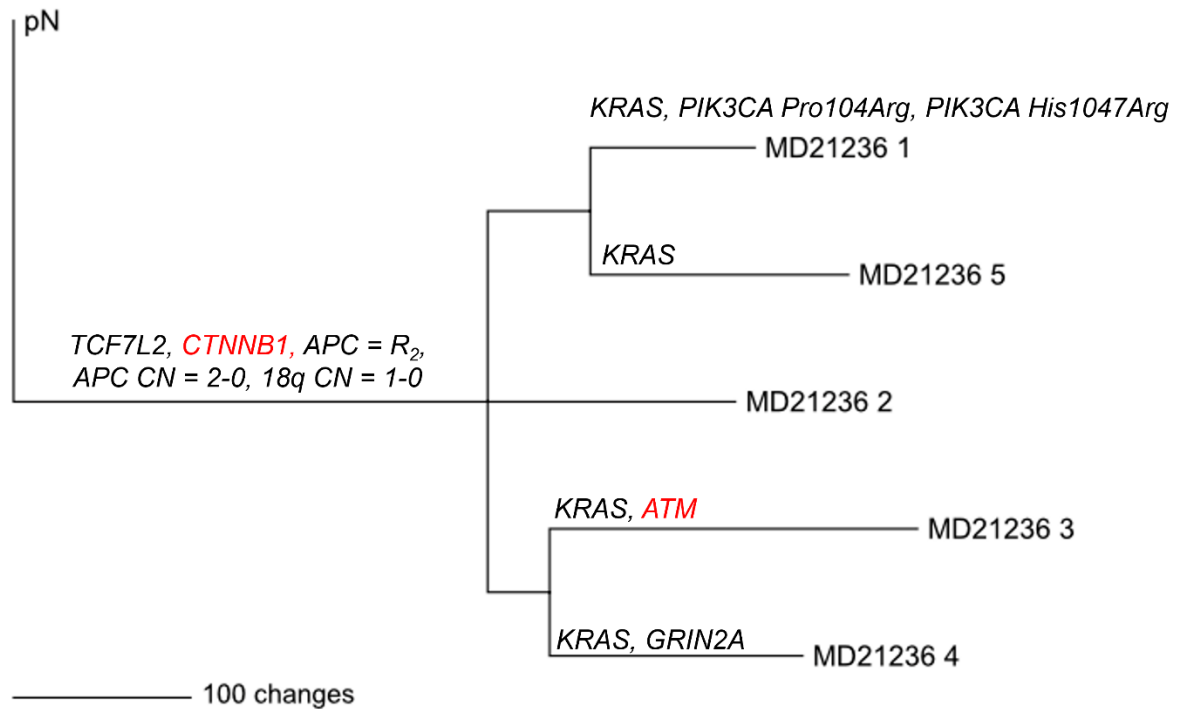


Figure 2.10 Phylogenetic tree of MD21236. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. CN = copy number (major-minor allele frequency). Red font = not pathogenic.

2.4.3.9 MD21239

Driver ITH is evident in this polyp through the presence of mutations in *LRP1B* in sample 2, and the same *EIF3E* mutation in samples 2 and 3 alone. This polyp also shows evidence of convergent evolution through the presence of two different *FAT4* mutations in samples 1 and 4 (**Fig. 11**).

Only one-hit to *APC* is present in this polyp in the form of clonal monosomy, so the classic just-right model involving 2 hits to *APC* cannot be assessed with no evidence of a mutation in a Wnt mini-driver gene.

Pathogenic mutations observed in this polyp include: *KRAS*^{Ala146Thr}, *ERBB2*^{Ser310Phe}, *MTCPI*^{Gln10Ter}, *LRP1B*^{Arg2373His}, *EIF3E*^{Gln247Glu}, *FAT4*^{Pro4902Leu}, *FAT4*^{Gln2775Leu}.

Table 2.11 Genetic profile summary for polyp MD21239. This table details the genetic characteristics of MD21239 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AA Rs	Wnt mini-drivers	Purity range (%)	Focus of cancer
Yes: all except 4	CN-LOH: clonal	Yes: all	NA	NA	42-79	Yes

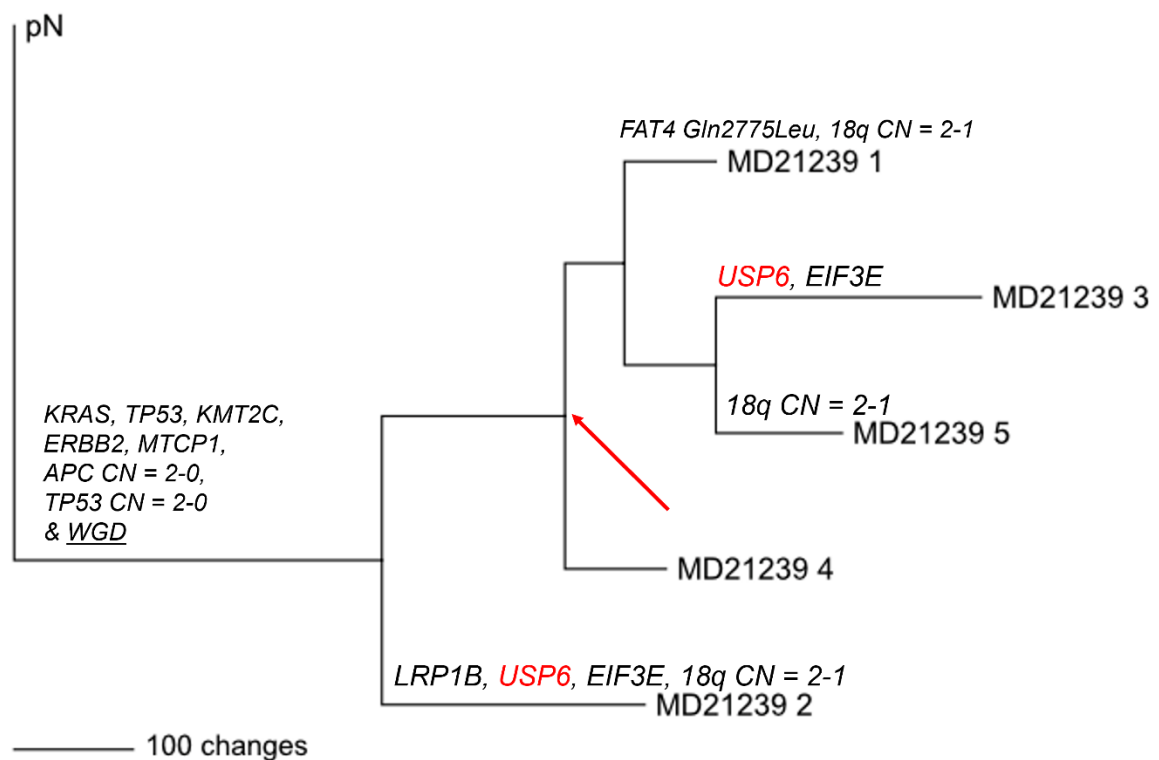


Figure 2.11 Phylogenetic tree of MD21239. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. CN = copy number (major-minor allele frequency). Red font = not pathogenic. Black = trunk. Blue = branches. Green = leaves.

2.4.3.10 MD21257

Driver ITH is evident in this polyp through a mutation in *ATM* in sample 4, *FBXW7* in sample 1 and *SMAD4* and *PIK2CA* in sample 5. Despite the presence of only a clonal *KRAS* mutation, this polyp was not recorded as sessile serrated. ITH is also present through the allelic imbalance of 18q which is only seen in samples 1, 3 and 5. In addition, monosomy of *TP53* is only harboured in sample 1. The only Wnt-related gene that harbours a mutation in this polyp is *FBXW7* in sample 1 (**Fig. 2.12**).

Pathogenic mutations present in this polyp include: *KRAS*^{Gly12Val}, *ATM*^{Leu507Ser}, *FBXW7*^{Arg479Gln}, *SMAD4*^{Pro356Ser}, *PIK3CA*^{Arg108His}.

Table 2.12 Genetic profile summary for polyp MD21257. This table details the genetic characteristics of MD21257 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
Yes: samples 1, 3, 5	Monosomy: sample 1	No	NA	<i>FBXW7</i>	57-82	Yes

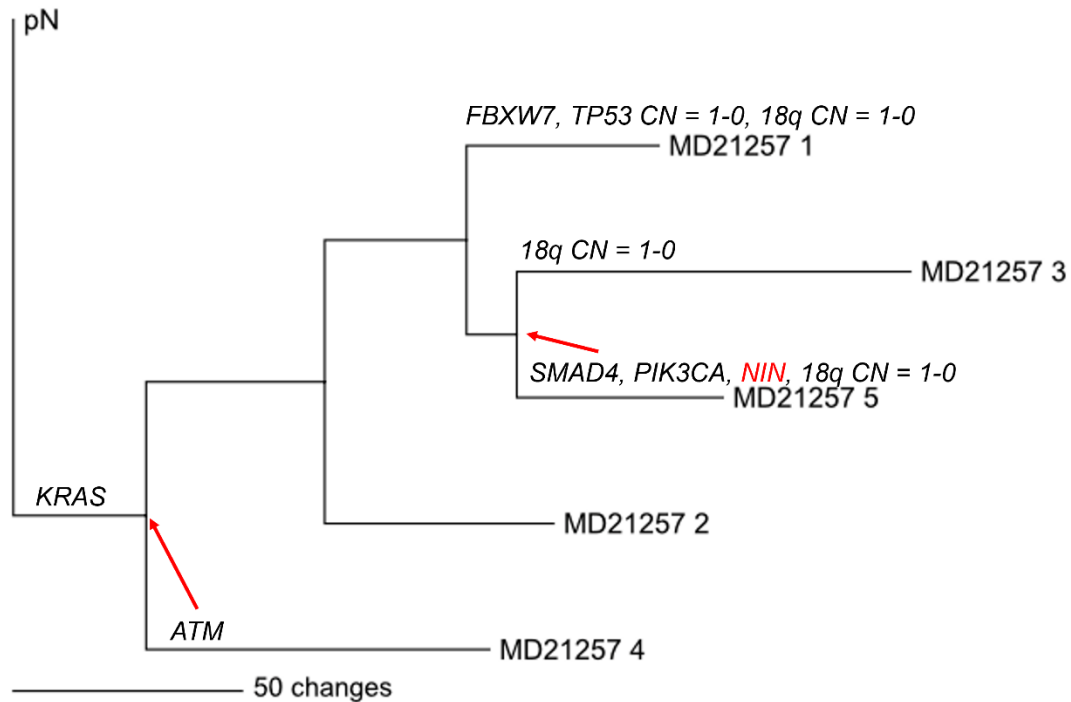


Figure 2.12 Phylogenetic tree of MD21257. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. CN = copy number (major-minor allele frequency). Red font = not pathogenic.

2.4.3.11 MD21274

Driver ITH is presented in this polyp through the same *KRAS* mutation only being present in samples 2, 3 and 5. Sample two had a purity of 20%, all other samples had a purity of above 72%.

APC has received two clonal hits in this polyp in the form of two mutations. These mutations reside in regions R0 and R3 resulting in the maintenance of three 20AARs. These *APC* mutations are present in combination with clonal mutations in Wnt mini-driver genes *BCL9L* and *FBXW7* (**Fig. 13**).

Pathogenic mutations in this polyp include: *GNAS*^{Arg844His}, *MYH9*^{Thr630Ile}, *BCL9*^{Pro517SerfsTer7}, *FBXW7*^{Arg465Cys}, *APC*^{Arg232Ter}, *APC*^{The1493ArgfsTer14}, *KRAS*^{Gly12Ser}.

Table 2.13 Genetic profile summary for polyp MD21274. This table details the genetic characteristics of MD21274 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	No	No	3	<i>FBXW7</i>	20-100	No

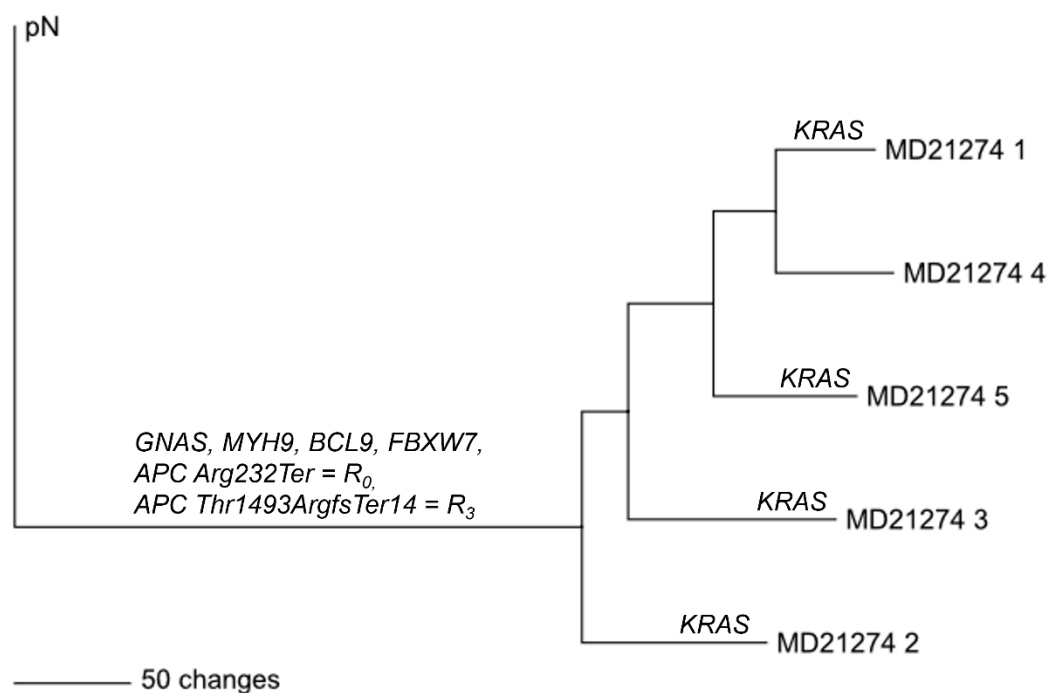


Figure 2.13 Phylogenetic tree of MD21274. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal.

2.4.3.12 MD21296

Driver ITH is evident in this polyp due to the presence of a *BCOR* mutation in sample 4 and the same *FAT3* mutation in samples 1 and 2 only. There is no evidence of *APC* mutations or any mutations in Wnt mini-driver genes (**Fig. 2.14**). After kindly examining the H&Es of this sample, it was indicated that this polyp was likely of inflammatory origin.

Pathogenic mutations in this polyp include: *BCOR*^{Ala4Glu} and *FAT3*^{Arg3877His}.

Table 2.14 Genetic profile summary for polyp MD21296. This table details the genetic characteristics of MD21296 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
Yes: clonal	No	No	NA	NA	87-91	No

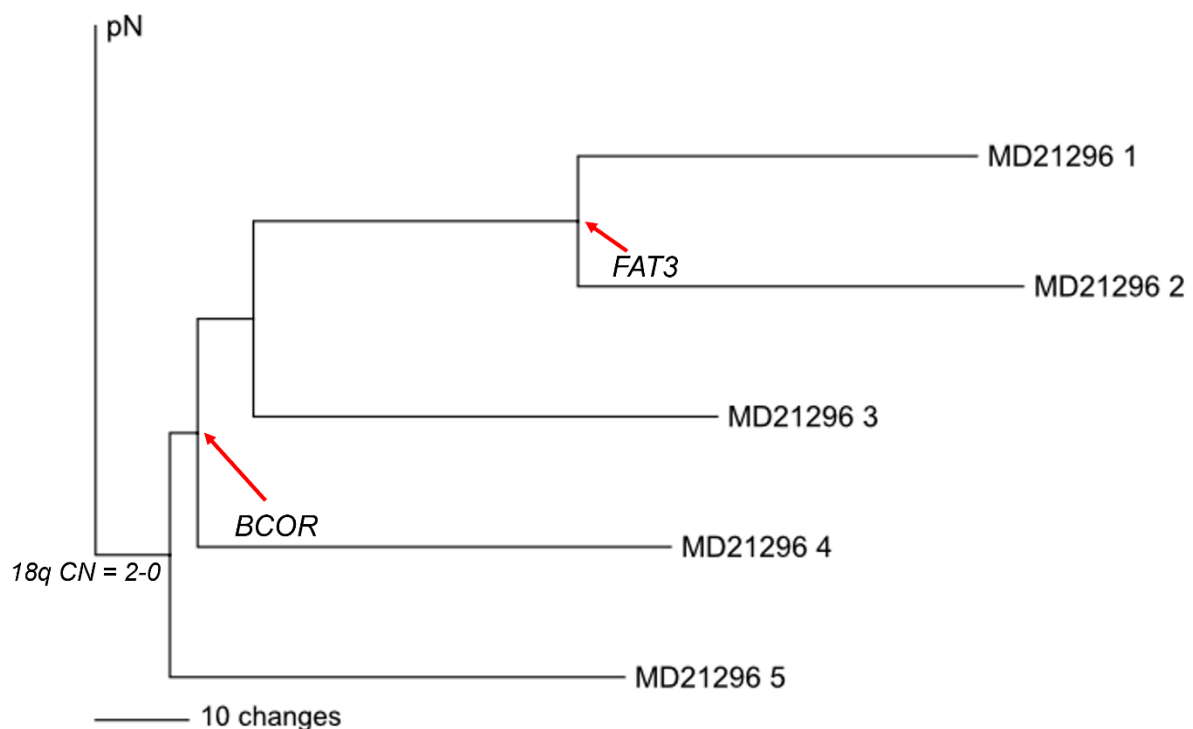


Figure 2.14 Phylogenetic tree of MD21296. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. CN = copy number (major-minor allele frequency).

2.4.3.13 MD21320 Polyp 1

Two polyps were sampled from the same patient. This patient has lynch syndrome and a germline *MSH6* mutation. Sample 1_5 had 21% purity; all other samples from both polyps had above 74% purity. All samples were identified as MSI and so have a hypermutator phenotype.

In polyp 1, multiple examples of driver ITH can be seen. Clonal *APC* mutations in polyp 1 result in the maintenance of two 20AARs. Interestingly, these mutations are also seen in the presence of a mutation in Wnt major-driver *RNF43*^{Gly659ValfsTer41}. There is evidence that this mutation is not pathogenic but this is controversial in previous literature (Tu *et al.*, 2019). If the just-right model is accurate, the fact that it occurs alongside another major-driver mutation in *APC*, might suggest a lack of pathogenicity. Multiple mutations are found in Wnt mini-drivers including *AMER1*, *ZNRF3*, *AXIN1*, *AXIN2* and *BCL9L*. Convergent evolution is also shown in this polyp through the occurrence of differing *CTNNB1* mutations in samples 1 and 2, neither show evidence of pathogenicity (**Fig. 2.15**).

Pathogenic mutations in this polyp include: *APC*^{Arg213Ter}, *APC*^{1450Ter}, *ZNRF3*^{Asp837Tyr}, *RNF43*^{Gly659ValfsTer41}, *AMER1*^{Arg626Ter}, *BCL9*^{Val910CysfsTer41}, *AXIN1*^{Val416Met}, *BCL9L*^{Pro32LeufsTer13}, *KRAS*^{V14Ile}.

Table 2.15 Genetic profile summary for MD21320 polyp 1. This table details the genetic characteristics of MD21320 polyp 1 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence/absence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	No	No	2	<i>RNF43</i> , <i>ZNRF3</i> , <i>AXIN2</i> <i>AMER1</i> <i>BCL9</i> <i>AXIN1</i> <i>BCL9L</i>	21-100	No

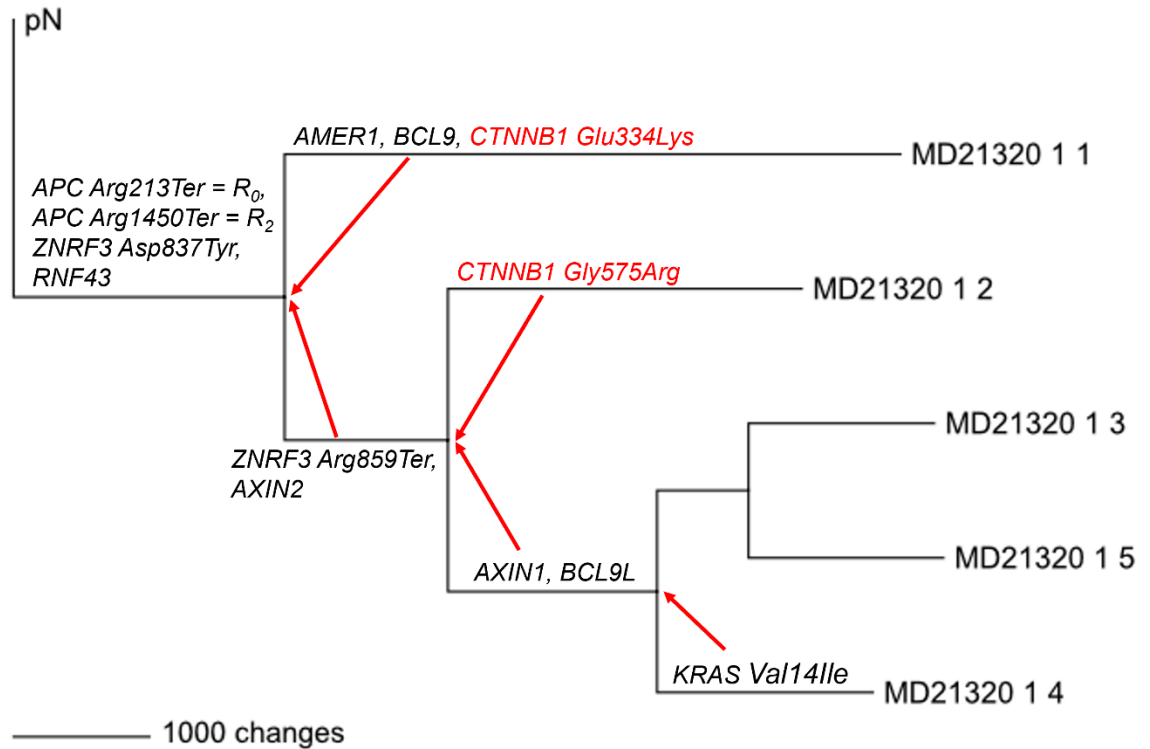


Figure 2.15 Phylogenetic tree of MD21320 polyp 1. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. Red font = not pathogenic.

2.4.3.14 MD21320 Polyp 2

In polyp 2, again driver ITH is shown. A clonal mutation in R₀ is present clonally, there is no other detectable, clonal mutation in *APC*. As with polyp MD21032 there is a second, sub-clonal mutation in *APC* in region 2. This may be representative of a third *APC* alteration or may represent part of the polyp that developed through an alternative route of *APC* inactivation. It is therefore difficult to assess the just-right model in this polyp. The same *RNF43* mutation as polyp 1 is also found in sample two of polyp 2 along with a mutation in *FBXW7*. There are two examples of convergent evolution in this polyp through the development of two different *FBXW7* mutations in samples 3 and 4 along with two different *BCL9L* mutations in samples 1 and 4 (**Fig. 2.16**).

Pathogenic mutations in this polyp include: *APC*^{Arg564Ter}, *APC*^{Ser1465TrpfsTer3}, *KRAS*^{Gly13A}, *RNF43*^{Gly659ValfsTer41}, *FBXW7*^{Arg479Gln}, *FBXW7*^{Arg465His}, *BCL9L*^{Gly554Ala}, *BCL9L*^{Arg409Trp}, *FBXW7*^{Arg505Cys}.

Table 2.16 Genetic profile summary for MD21320 polyp 2. This table details the genetic characteristics of MD21320 polyp 2 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	No	No	NA	<i>FBXW7</i> , <i>BCL9L</i>	21-100	“Suspicious”

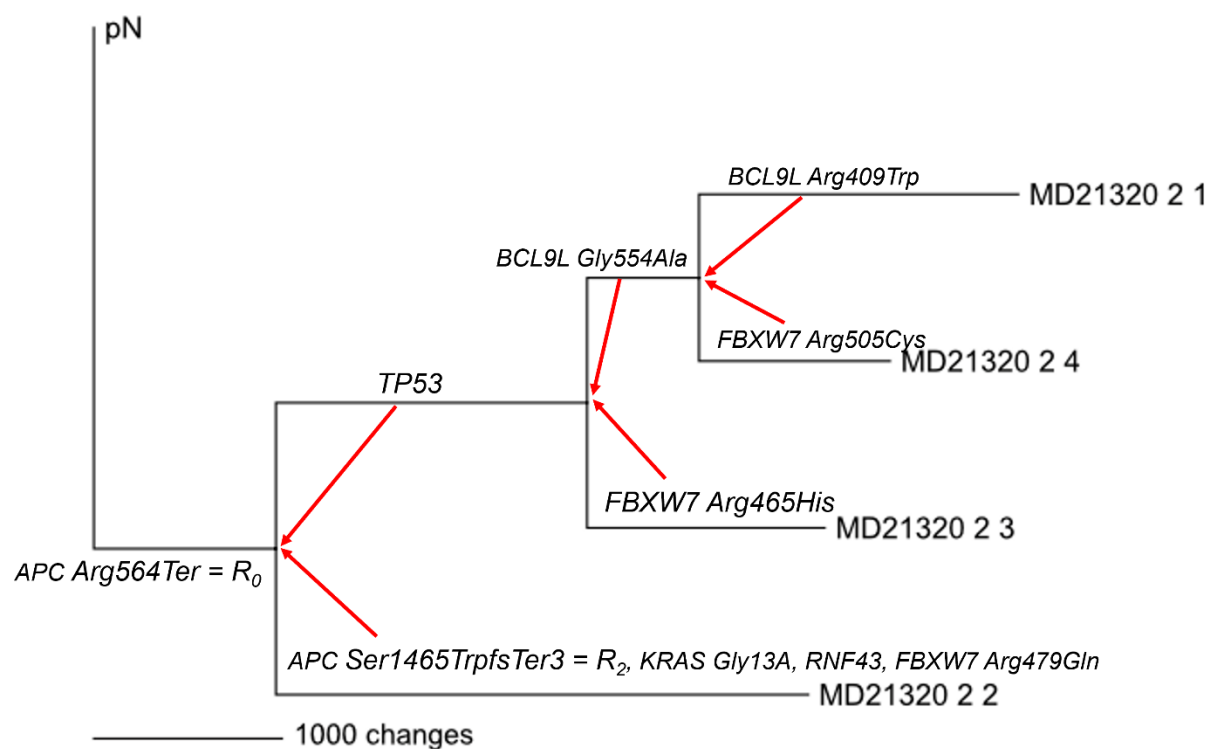


Figure 2.16 Phylogenetic tree of MD21320 polyp 2. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal.

2.4.3.15 MD21340

Driver ITH is evidenced in this polyp through the presence of a *PPP2R1A* mutation in sample 1. *APC* has two clonal hits in *APC* resulting in the total maintenance of two 20AARs. There is no evidence of any other Wnt-related mutations (**Fig. 2.17**).

Pathogenic mutations in this polyp include: *TP53*^{spliceacceptor}, *PPP2R2A*^{Pro130Leu} and *APC*^{Glu941Ter}, *APC*^{Arg1450Ter}.

Table 2.17 Genetic profile summary for polyp MD21340. This table details the genetic characteristics of MD21340 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	Monosomy: clonal	No	2	NA	27-36	No

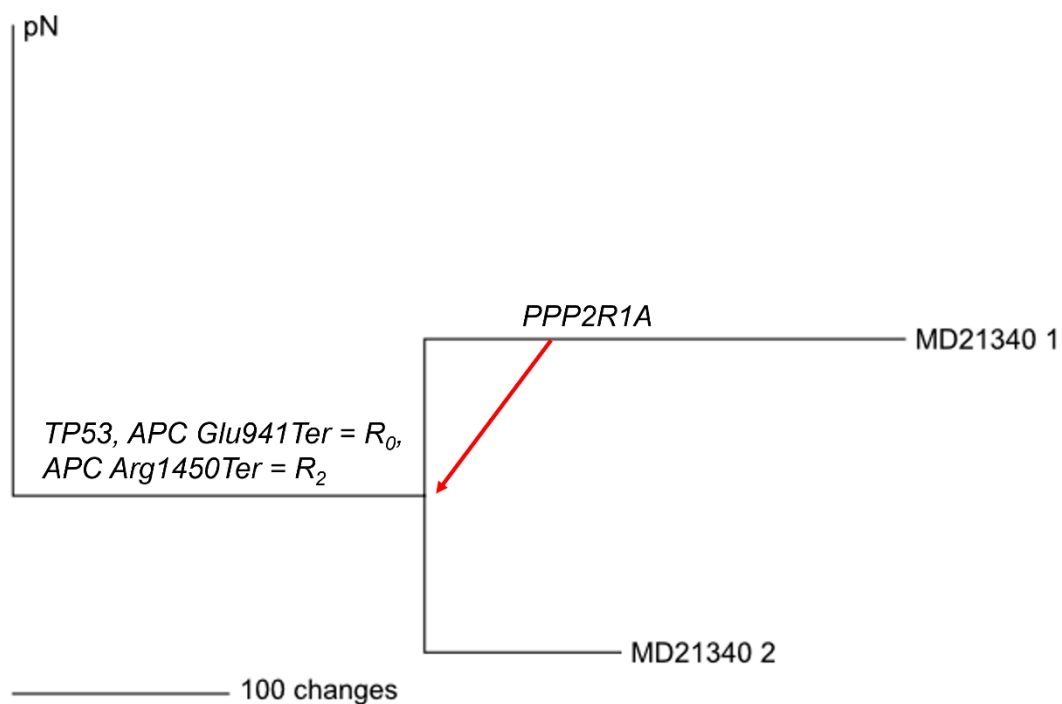


Figure 2.17 Phylogenetic tree of MD21340. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal.

2.4.3.16 MD21353

Driver ITH is evidenced in this polyp via the same *AMER1* mutation in samples 1, 2, 4 and 5 only; an *LRP1B* mutation in sample 3 only; the same *GNAS* mutation in samples 2 and 5 and *SMAD2* and *BCL9* mutations in sample 1. Evidence of convergent evolution is also seen in this polyp through the development of two different *SOX9* mutations in samples 2 and 4 (**Fig. 2.18**). ITH is also seen in the form of 18q imbalance which only occurs in sample 1.

There are two clonal pathogenic hits to *APC* resulting in the total maintenance of three 20AARs. These mutations are found in combination with a clonal mutation in Wnt mini-driver gene *FBXW7*. An *AMER1* mutation then occurs in samples 1, 2, 4 and 5. *SOX9* mutations occur in both samples 4 and 2 and a *BCL9* mutation occurs in sample 1.

Pathogenic mutations in this polyp include: *KRAS*^{Gly12Cys}, *ACVR2A*^{Arg298Ter}, *ACVR2A*^{Asp386CysfsTer6}, *APC*^{Tyr951LeufsTer12}, *APC*^{Ser1539Ter}, *FBXW7*^{Arg465Cys}, *AMER1*^{Arg531Ter}, *LRP1B*^{Ser692Leu}, *SOX9*^{Tyr297Ter}, *SMAD2*^{Ser464Ter}, *BCL9*^{Pro1063Leu}, *GNAS*^{Arg844His}, *SOX9*^{Ser284Glu287delinsTer}.

Table 2.18 Genetic profile summary for polyp MD21353. This table details the genetic characteristics of MD21353 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
Yes: sample 1	No	No	3	<i>AMER1</i> , <i>SOX9</i> , <i>BCL9</i>	100	Yes

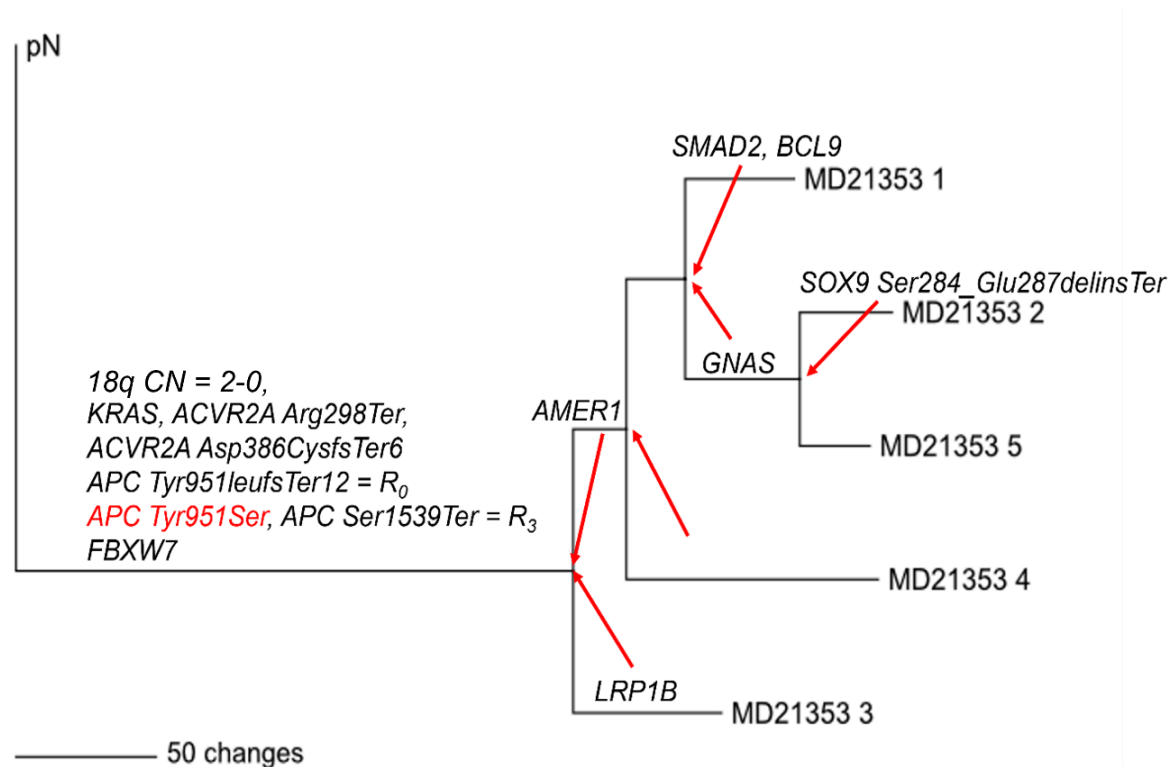


Figure 2.18 Phylogenetic tree of MD21353. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal.). Red font = not pathogenic.

2.4.3.17 MD21354

Polyp MD21354 is sampled from an FAP patient. Driver ITH is evidenced in this polyp through the occurrence of an *LRP1B* mutation in sample 1, a *SOX9* mutation in sample 3 and the same *APC* mutation in samples 2, 3 and 4. Truncal *APC* mutations could not be detected and so action of the just-right model via two hits to *APC* could not be assessed (**Fig. 2.19**). There was no evidence in the clinical report that this is a serrated polyp.

Pathogenic mutations in this polyp include: *KRAS*^{Gly12Asp}, *LRP1B*^{Glu2708Ter}, *APC*^{Thr1556AsnfsTer3}, *SOX9*^{Val486GlyfsTer92}.

Table 2.19 Genetic profile summary for polyp MD21354. This table details the genetic characteristics of MD21354 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	No	No	NA	NA	100	Yes

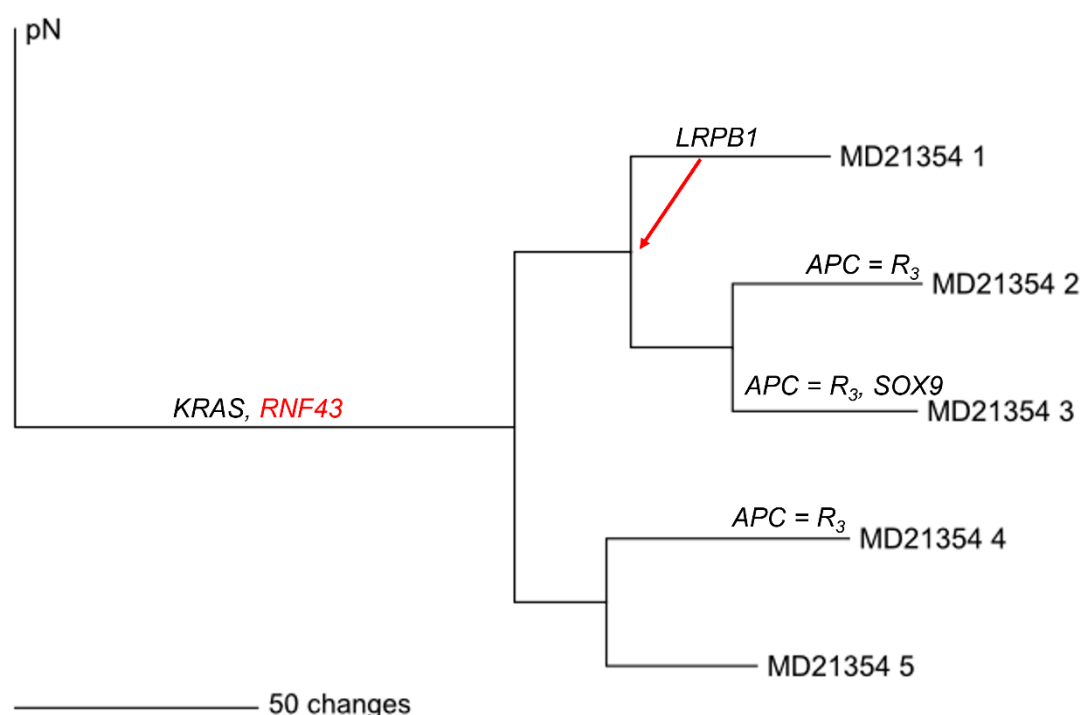


Figure 2.19 Phylogenetic tree of MD21354. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. Red font = not pathogenic.

2.4.3.18 MD21365

Driver ITH is evidenced in this polyp through the occurrence of the same *KRAS* mutation in samples 1, 2, 4 and 5; a *TCF7L2* mutation in sample 5 and an *NBEA* mutation in sample 1 (**Fig. 2.20**). ITH is also shown in the form of WGD which was found in sample 3 alone. Sample 3 had a purity of 28%, all other samples had a purity of above 80%.

There are two clonal hits to *APC* in the form of two mutations, one in R_0 and the other in R_2 , resulting in the maintenance of two 20AARs. This is found in combination with a clonal *FBXW7* mutation.

Pathogenic mutations observed in this polyp include: *FBXW7*^{Arg465His}, *APC*^{Arg1114Ter}, *APC*^{Ser1411ArgfsTer4}, *KRAS*^{Gly12Val}, *TCF7L2*^{Asp16His}.

Table 2.20 Genetic profile summary for polyp MD21365. This table details the genetic characteristics of MD21354 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN-LOH)	WGD	<i>APC</i> 20AARs	Wnt mini-drivers	Purity range (%)	Focus of cancer
No	No	Yes: sample 3	2	<i>FBXW7</i> , <i>TCF7L2</i>	28-85	No

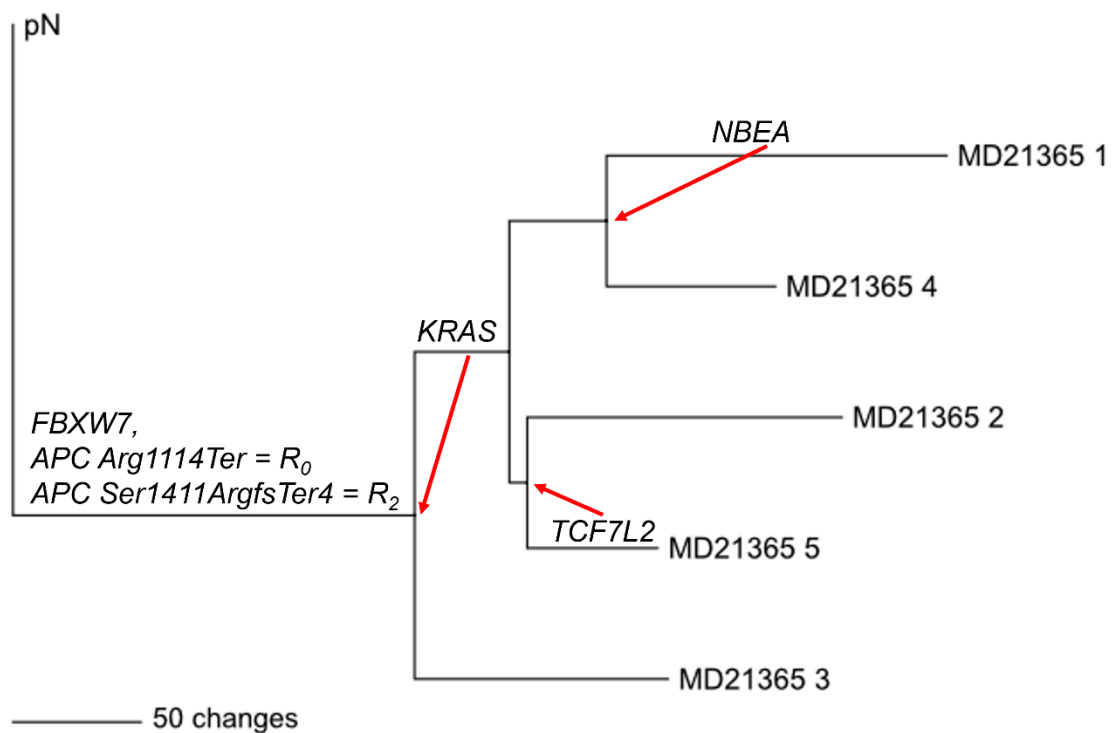


Figure 2.20 Phylogenetic tree of MD21365. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal.

2.4.3.19 MD21386

Driver ITH is evidenced in this polyp through the occurrence of a *BIRC6* mutation in sample 4 and the occurrence of *KRAS Gly13Asp* in both samples 3 and 5. Convergent evolution can be seen due to the occurrence of *KRAS Gly12Asp* in sample 2 (**Fig. 2.21**).

There are two clonal hits in *APC*, one in R₁ and the other in R₂, resulting in the maintenance of three 20AARs. There is no evidence of mutation in any other Wnt mini-driver genes.

Pathogenic mutations observed in this polyp include: *BIRC6*^{Thr1700LysfsTer6}, *APC*^{Met1211AsnfsTer6}, *APC*^{Glu1408Ter}, *KRAS*^{Gly12Asp}, *BIRC6*^{Arg4725Ter}.

Table 2.21 Genetic profile summary for polyp MD21386. This table details the genetic characteristics of MD21386 including: 18q allelic imbalance; detection of monosomy or copy-neutral loss of heterozygosity (CN-LOH) in *TP53*; presence/absence of whole genome duplication (WGD); the number of 20AARs in *APC* following two hits; the presence of Wnt mini-driver mutations; sample purity percentage range; cancer focus.

18q allelic imbalance	<i>TP53</i> (monosomy/CN- LOH)	WGD	<i>APC</i> 20AARs	Wnt mini- drivers	Purity range (%)	Focus of cancer
No	No	No	3	NA	No	No

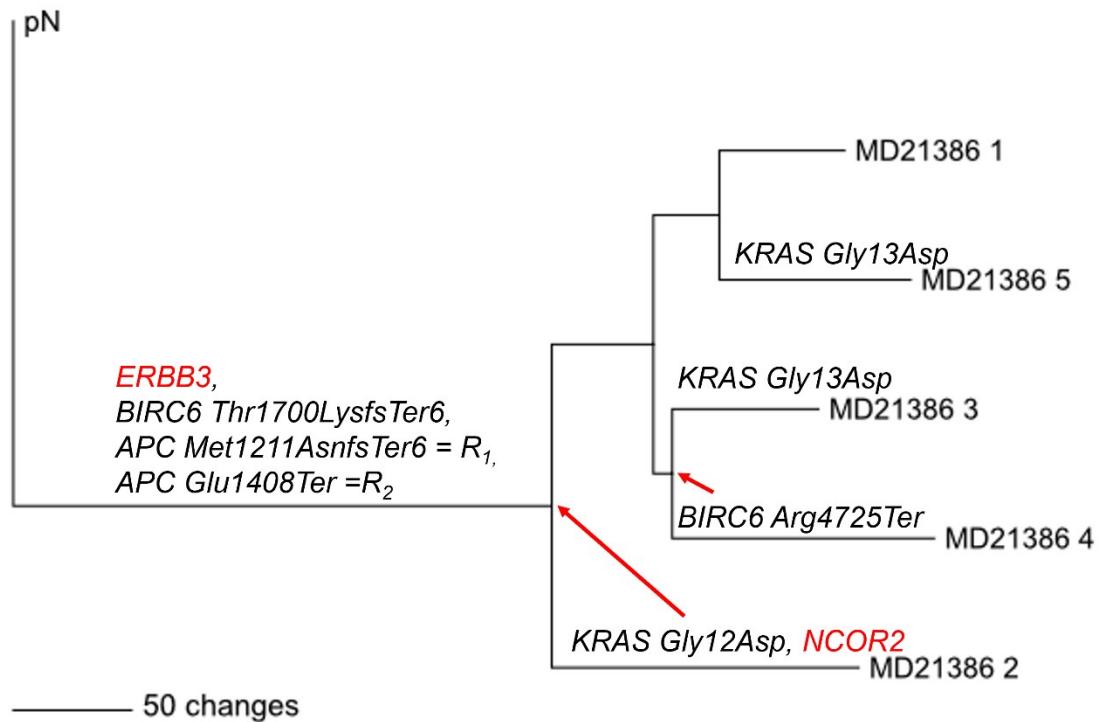


Figure 2.21 Phylogenetic tree of MD21386. Annotated are both mutations in Intogen colorectal cancer driver genes and Wnt pathway mutations of interest. pN = pseudonormal. CN = copy number (major-minor allele frequency). Red font = not pathogenic. Black = trunk. Blue = branches. Green = leaves.

2.4.3.20 Just-right analysis summary

Analysis of the just-right model showed that there were between one and four 20AARs in *APC* remaining in nine polyps that had two detectable, clonal alterations in *APC* (Table 2.22).

Table 2.22 A summary of remaining 20AARs remaining in *APC* in polyps with two detectable, clonal *APC* alterations and presence of mutations in mini-driver genes. * = MSI⁺.

Polyp	20AARs	Clonal mini-drivers	Sub-clonal mini-drivers
MD20716	3	<i>FBXW7</i>	NA
MD21031	1	<i>SOX9</i>	<i>TCF7L2</i>
MD21236	4	<i>TCF7L2</i>	NA
MD21274	3	<i>FBXW7</i>	NA
MD21320_1*	2	<i>RNF43, ZNRF3</i>	<i>AMER1, BCL9, AXIN1, AXIN2</i>
MD21340	2	NA	NA
MD21353	3	<i>FBXW7</i>	<i>AMER1</i>
MD21365	2	<i>FBXW7</i>	NA
MD21385	3	NA	NA

Clonal *FBXW7* mutations were found in three polyps with three 20AARs remaining and one polyp with two 20AARs remaining. There was one polyp with two 20AARs and another with three 20AARs with no additional Wnt pathway mutations. A clonal *SOX9* mutation was found in the context of a polyp with one 20AAR in *APC*. A clonal *TCF7L2* mutation was found in a polyp with four remaining 20AARs in *APC*. *RNF43* and *ZNRF3* mutations were found in an MSI⁺ polyp with three 20AARs. Additional sub-clonal mini-driver mutations are presented in Table 2.20.

2.4.4 Tumour purity, ploidy and whole genome duplication

A Mann-Whitney U test was conducted to determine if there is a difference in tumour purity between samples taken from polyps with a focus of cancer (CRC) and polyps without a focus of cancer (CRA). Establishing equivalent tumour purity is important to ensure any downstream differences found between the groups are not attributable to differences in purity. The results indicated no significant difference in tumour purity between the two groups (U = 574.5, p = 0.36) (**Fig. 2.22**).

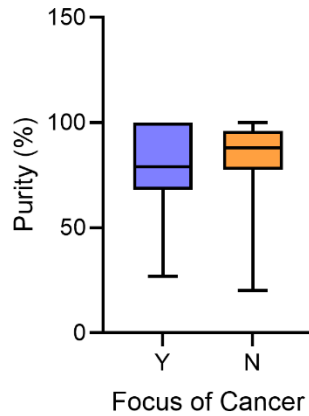


Figure 2.22 A comparison of tumour purity between samples taken from polyps with a focus of cancer and without a focus of cancer. Y = Yes, sample taken from polyp with a focus of cancer. N = No, samples not taken from a polyp with a focus of cancer.

Differences in tumour ploidy were also analysed between polyps with a focus of cancer (CRC) and those without (CRA). To address potential correlations among multiregional samples from the same polyp, average tumour ploidy values were calculated for each polyp. A Mann-Whitney U test was then performed to assess the difference in tumour ploidy between groups. No significant difference in tumour ploidy was found between polyps with a focus of cancer and those without a focus of cancer (**Fig. 2.23**).

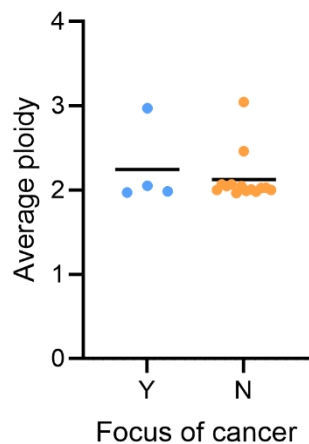


Figure 2.23 A comparison of tumour ploidy between polyps with a focus of cancer and without a focus of cancer. Horizontal bar = mean. Y = Yes, sample taken from polyp with a focus of cancer. N = No, samples not taken from a polyp with a focus of cancer.

Whole genome duplication (WGD) was identified in a total of seven samples. Five samples originated from the same polyp, MD21239, indicating a correlation among these multiregional samples. Polyp MD21239 harboured a focus of cancer. The other two samples exhibiting WGD were from two separate polyps MD21365 and MD20430, neither polyp harboured a focus of cancer. Since only one

multiregional sample from two separate polyps indicated WGD, this is an example of ITH and branched evolution.

To determine if there is a significant difference in the incidence of WGD between polyps with and without a focus of cancer, the proportion of each polyp exhibiting WGD was calculated. A Mann-Whitney U test was then utilised to determine if there is a significant difference in the proportion of polyps exhibiting WGD in samples with and without a focus of cancer. This showed no significant difference between the two groups ($U = 25.5$, $p = 0.711$) (**Fig. 2.24**).

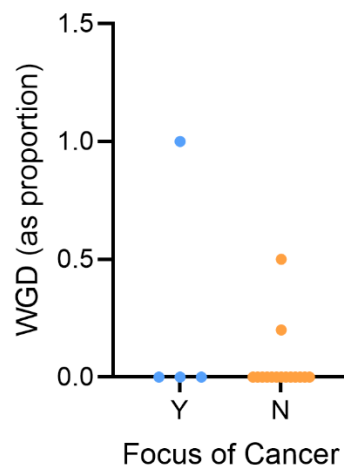


Figure 2.24 A comparison of WGD events in polyps with and without a focus of cancer. The proportion of each polyp exhibiting WGD was calculated and compared between groups. Y = Yes, sample taken from polyp with a focus of cancer. N = No, samples not taken from a polyp with a focus of cancer.

A Mann-Whitney U test was used to determine if there is a significant difference in the ploidy levels of polyp samples either exhibiting WGD compared with those not exhibiting WGD. This showed a significant difference between the two groups ($U = 0$, $p < 0.0001$) (**Fig. 2.25**).

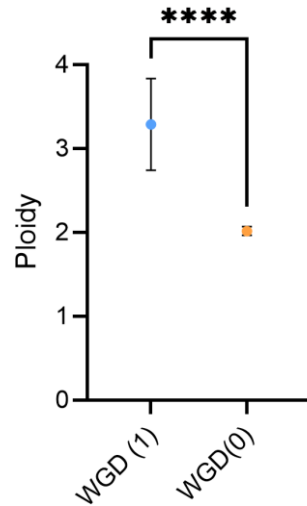


Figure 2.25 The ploidy levels of samples showing evidence of whole genome duplication (WGD) is significantly higher than those without WGD. Total number of WGD samples = 7. Number of non-WGD samples = 78

2.4.5 Mutation burden

The distribution of mutation burden across all samples can be seen in Figure 2.26. Following this comparison, the analysis in section 2.4.5 was carried out using MSS samples only. MSI⁺ samples were excluded due to their hypermutable phenotype as a result of defective mismatch repair (Boland and Goel, 2010). None of the MSI⁺ polyps had a focus of cancer.

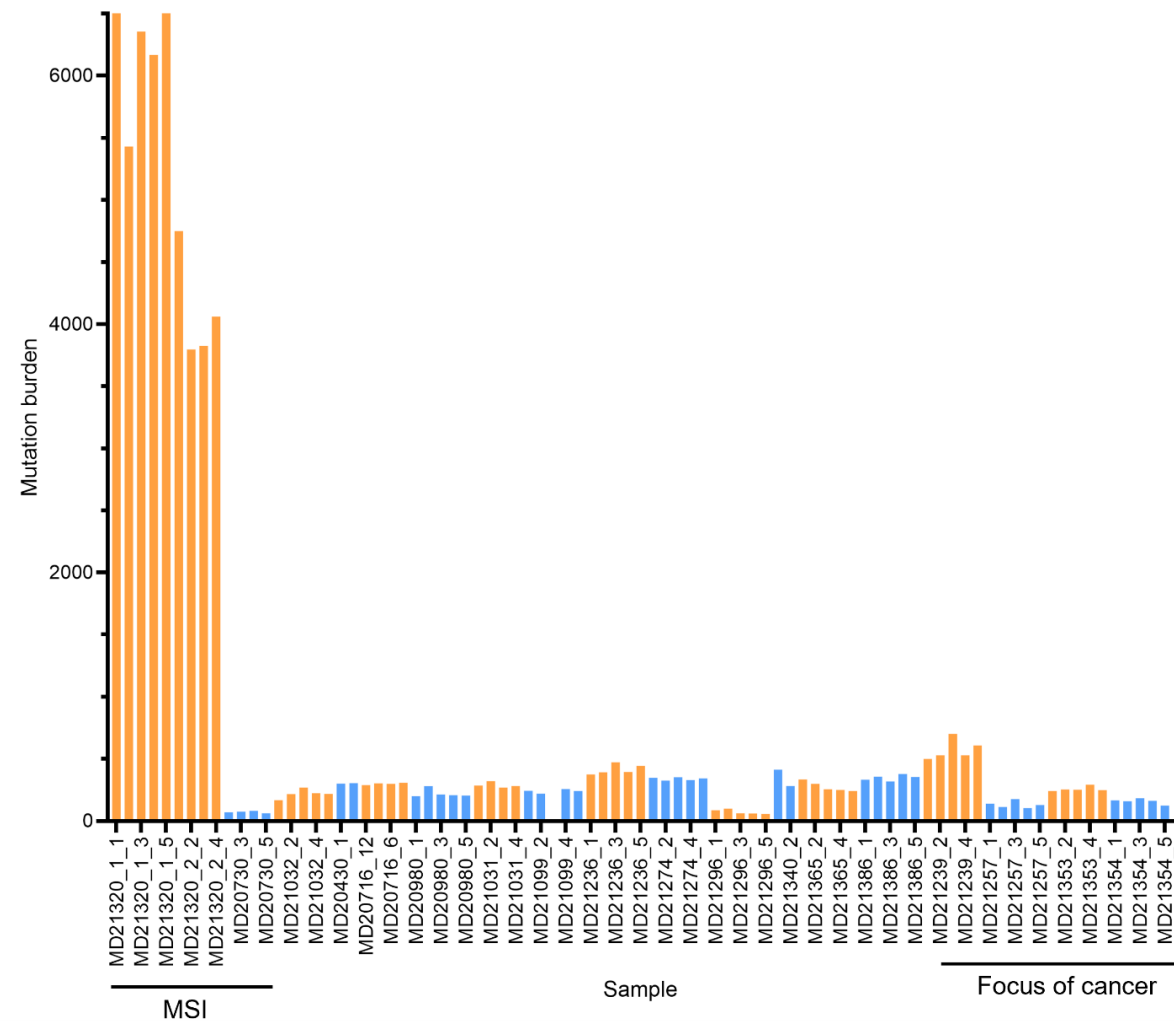


Figure 2.26 Mutation burden of all multiregional samples including SNVs and indels. MSI = Microsatellite instability. Orange and blue coloured bars are utilised to cluster multiregional samples from each polyp.

Although to an extent, mutation burden appears to vary between multiregional samples from the same polyp (**Fig. 2.26**), multiple factors, such as age, can influence the molecular characteristics of polyps. This shows similarities in mutation burden across multiregional samples from the same polyp (**Fig. 2.27**).

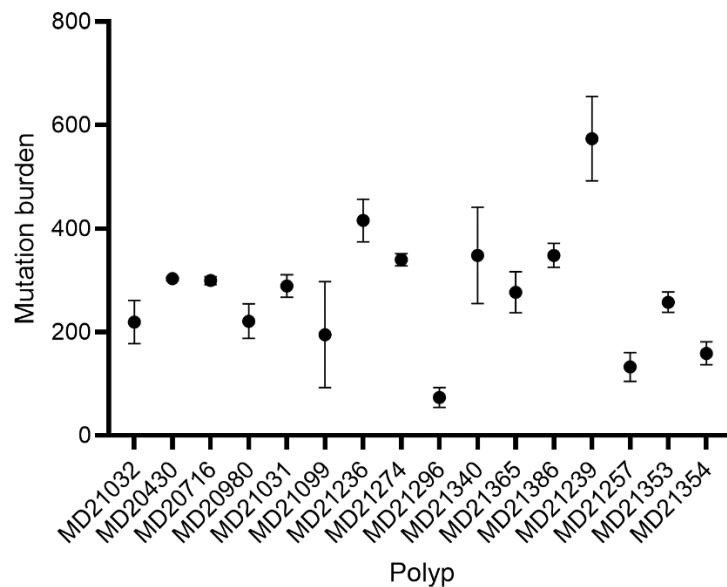


Figure 2.27 Average mutation burden per polyp. Displayed are the mean and standard deviations of mutation burdens calculated from multiregional samples within each polyp.

Due to this similarity in mutation burden between samples from the same polyp, the average mutation burden across each polyp was calculated and the difference in mutation burden between polyps with (CRCs) and without (CRAs) a focus of cancer, rather than samples, was analysed.

A Shapiro-Wilk test indicated that polyps with a focus of cancer (Y) and those without a focus of cancer (N) groups exhibited normal distributions ($p < 0.05$). A T-test was therefore carried out, which determined that although slightly higher in polyps without a focus of cancer, there is no significant difference in the mutation burden of polyps with and without a focus of cancer (**Fig. 2.28**). The same result was found by Cross *et al.*, (2018).

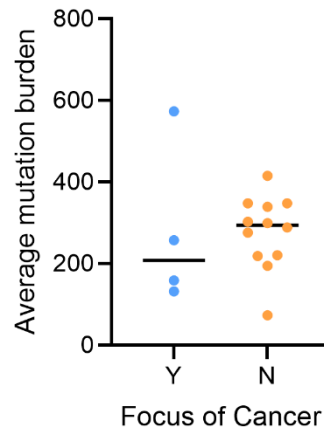


Figure 2.28 A comparison of the average mutation burden of polyps with and without a focus of cancer. Horizontal line = median. Y = Yes, polyp has a focus of cancer. N = No, polyp does not have a focus of cancer.

Driver mutation burden was assessed using the Intogen colorectal cancer driver genes. The number of pathogenic SNVs and indels within the 76 Intogen colorectal cancer driver genes in MSS samples was established. Again, due to potential correlation between samples, an average driver mutation burden was calculated for each MSS polyp. Both those polyps with a focus of cancer (Y) and those without (N) displayed a normal distribution via Shapiro-Wilk test ($p > 0.05$). No significant difference in driver mutation burden was found between polyps with and without a focus of cancer (**Fig 2.29**).

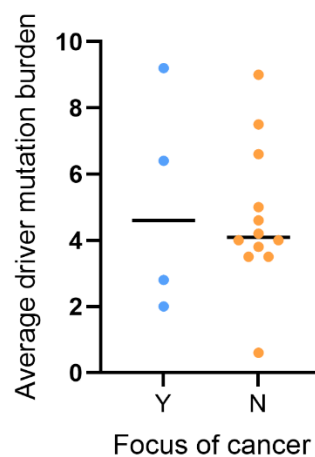


Figure 2.29 A comparison of the average driver mutation burden between polyps with and without a focus of cancer. Horizontal line = median. Y = Yes, polyp has a focus of cancer. N = No, polyp does not have a focus of cancer.

In addition to driver mutation burden in the form of SNAs and indels, copy-neutral loss of heterozygosity (CN-LOH), monosomy events in *TP53* and *APC* and 18q allelic imbalance were incorporated into the analysis. CN-LOH and monosomy act as second hits in major tumour suppressor genes (TSG) *TP53* and *APC*. 18q deletion is frequent in CRC (Tanaka *et al.*, 2006).

Total driver mutation burden, plus CN-LOH and monosomy events in *APC* and *TP53*, as well as 18q imbalance events, were averaged across polyps. A Shapiro-Wilk test identified the group of polyps with a focus of cancer (Y) showed data that was not normally distributed ($p < 0.05$), therefore a Mann-Whitney U test was carried out to determine if there was a significant difference between polyps with and without a focus of cancer. The results showed that there was no significant difference in driver mutation burden when also including monosomy and CN-LOH of *APC* and *TP53* as well as 18q allelic imbalance ($U = 22$, $p = 0.846$) (**Fig. 2.30**).

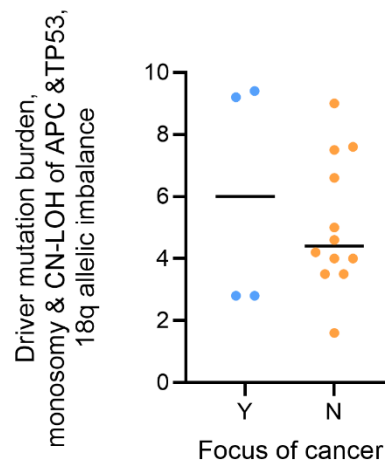


Figure 2.30 A comparison of average Intogen cancer driver mutation burden (SNAs/indels), monosomy and CN-LOH of *APC* and *TP53*, and 18q allelic imbalance in polyps with and without a focus of cancer. Horizontal line = median. Y = Yes, polyp has a focus of cancer. N = No, polyp does not have a focus of cancer.

2.4.6 Polyp size

Polyp size refers to maximum diameter (mm). This dataset is of particular interest as it is focused on the study of large adenomatous polyps ($> 1\text{cm}$). It is currently unknown why many of these large polyps did not progress to a cancer. Again, this section exclusively analyses MSS polyps due to the hypermutated phenotype of MSI⁺ samples.

Interestingly, although cancer-containing polyps had a slightly higher median size than polyps without a focus of cancer, the difference did not reach significance. ($T = 2.031$, $p = 0.062$) (**Fig. 2.31**).

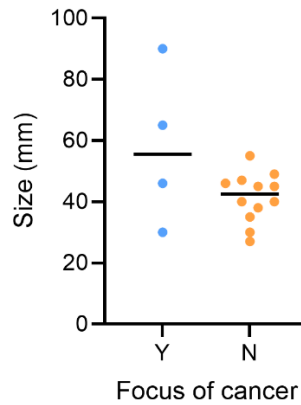


Figure 2.31 A comparison of size between polyps with and without a focus of cancer. Horizontal line = median. Y = Y = Yes, polyp has a focus of cancer. N = No, polyp does not have a focus of cancer.

Notably, a significant, inverse relationship was found between average mutation burden and polyp size, the smaller the polyp, the higher the mutation burden ($F = 5.016$, $p = 0.042$). No correlation was found between average number of mutations in Intogen colorectal cancer driver genes and size ($F = 0.837$, $p = 0.376$) (**Fig. 2.32**).

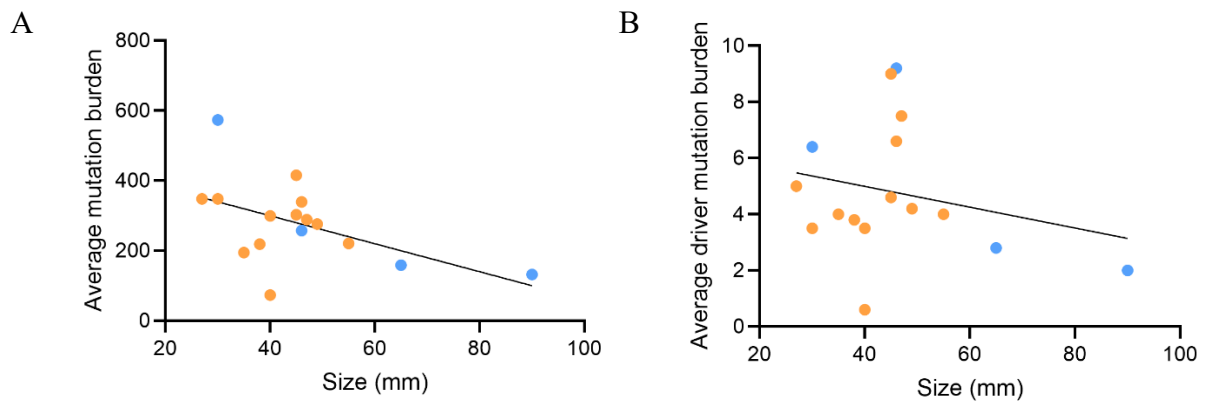


Figure 2.32 The relationship between polyp size (mm) and average mutation burden. Blue = focus of cancer. Orange = no focus of cancer. A) Shows the relationship between polyp size (mm) and average exonic mutation burden. B) Shows the relationship between polyp size (mm) and average Intogen colorectal cancer driver mutation burden.

An analysis was then carried out to determine the relationship between average Intogen colorectal cancer driver mutation burden and age. The results indicated that driver mutation burden increases significantly with age ($F = 6.071$, $p = 0.027$). The relationship between polyp size and age was also explored. In this case, a negative relationship was shown between polyp size and age, which was just over the significance threshold ($F = 4.581$, $p = 0.0504$) (**Fig. 2.33**). A big influence on this test was a 36-year-old FAP patient with a tumour diameter of 90mm. When removed from the analysis, the relationship between age and tumour size is definitively, non-significant ($F = 0.01$, $p = 0.920$).

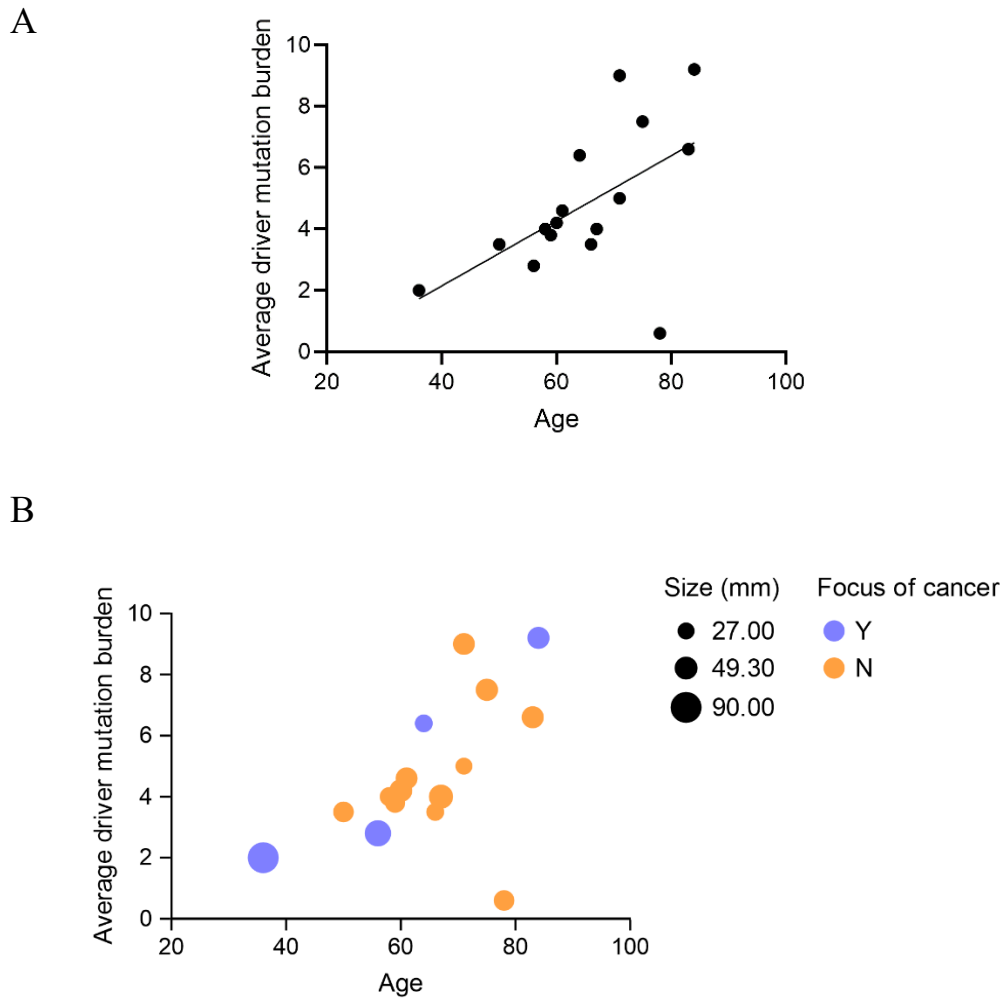


Figure 2.33 The relationship between average Intogen colorectal cancer driver mutation burden, size (mm) and age. A) Shows the relationship between average Intogen colorectal cancer driver mutation burden and age. B) Shows the relationship between average Intogen colorectal cancer driver mutation burden (y-axis), age in years (x-axis), size of polyp (mm) and focus of cancer (blue = focus of cancer present, orange = no focus of cancer).

2.4.7 Copy number

The numbers of CNAs for both MSS polyps with and without a focus of cancer were analysed. Again, to mitigate for correlations between samples, an average CNA value was calculated for each polyp, this included the number of regions with copy loss, CN-LOH and amplification. (**Fig 2.34**).

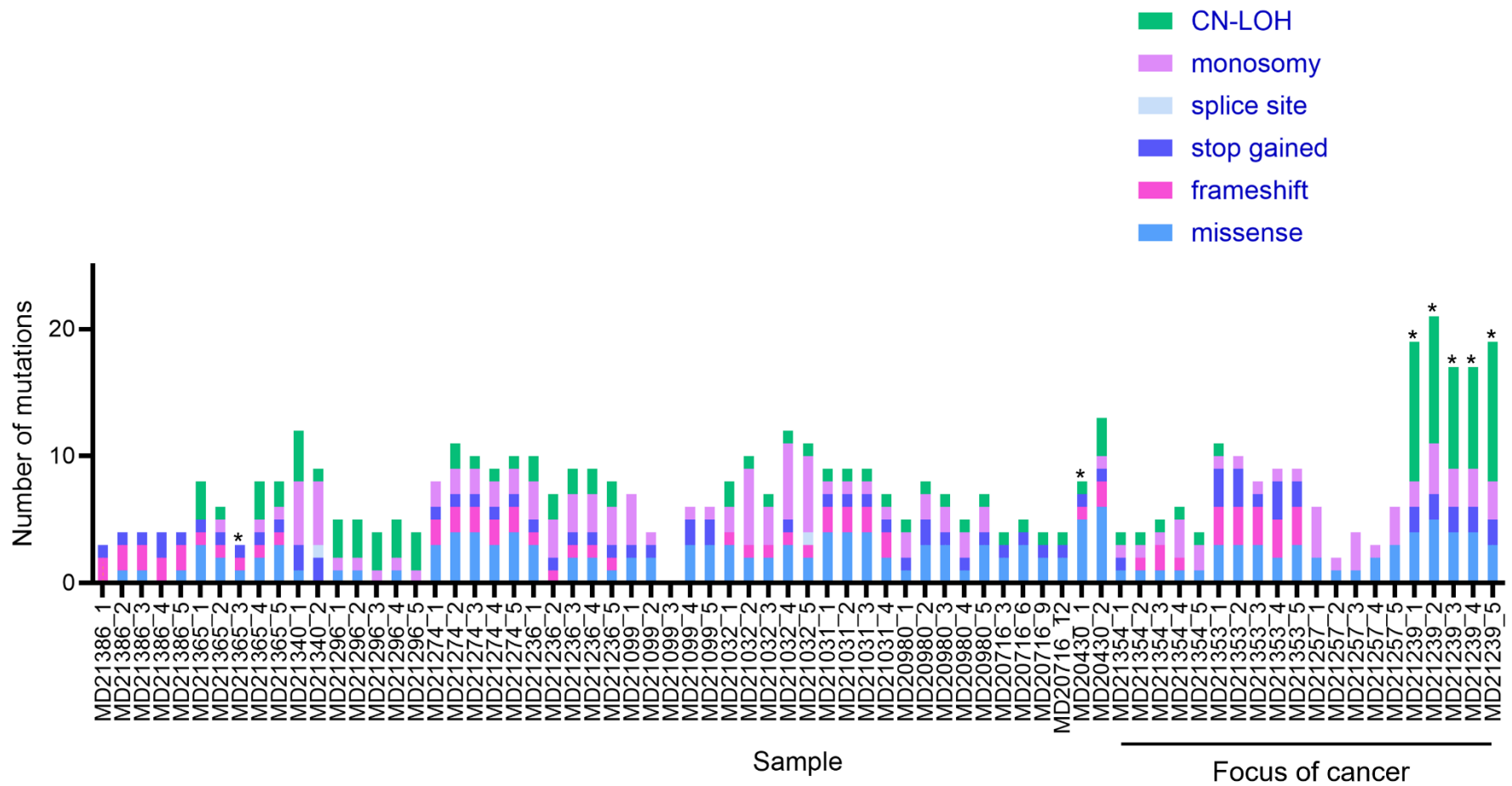


Figure 2.36 An overview of all SNV types, monosomy, CN-LOH and WGD events in all MSS polyp samples. * = WGD.

2.4.8 ITH

A measure of ITH was assessed for both samples taken from an MSS polyp with a focus of cancer and without a focus of cancer. Polyps with only two samples were removed from this analysis. Firstly, the proportion of unique variants (variants identified in one sample alone) to total variants was calculated for each polyp. The median proportion of unique variants for polyps with a focus of cancer was 16.32%. The median proportion of unique variants for polyps without a focus of cancer was 21.84%. Although slightly higher in polyps without a focus of cancer, a Mann-Whitney U test showed no significant difference in the proportion of unique variants in polyps with and without a focus of cancer ($U = 12$, $p = 0.304$) (**Fig. 2.37**).

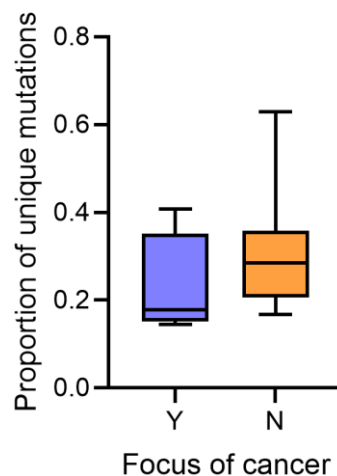


Figure 2.37 The distribution of mutations present in one (unique) or more than one sample in MSS polyps with and without a focus of cancer. Y = Yes, polyp has a focus of cancer. N = No, polyp does not have a focus of cancer.

In addition, the number of sub-clonal driver mutations (any driver mutation that was not clonal) were calculated as a proportion of total driver mutations. A Mann-Whitney U test was carried out to determine whether there was a significant difference in the number of sub-clonal driver mutations in polyps with and without a focus of cancer. The results showed that there was no significant difference in the proportion of sub-clonal driver mutations in polyps with and without a focus of cancer ($U = 15$, $p = 0.52$) (**Fig. 2.38**).

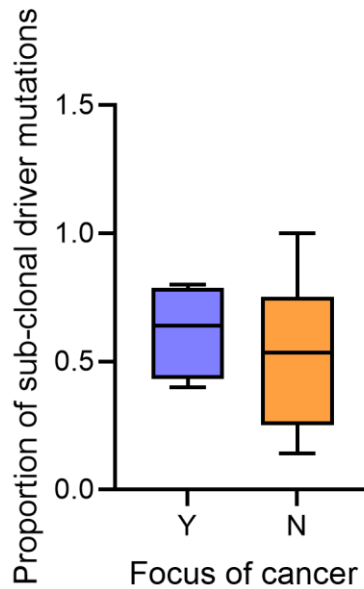


Figure 2.38 The proportion of sub-clonal driver mutations compared to total driver mutations in polyps with and without a focus of cancer. Y = Yes, polyp has a focus of cancer. N = No, polyp does not have a focus of cancer.

As shown above, parsimony phylogenetic trees were created for each polyp. The topology of phylogenetic trees provides insight into the clonality of mutations e.g. shorter trunks indicate early divergence and high ITH whereas longer trunks indicate more shared mutations and lower ITH. Trunk length was calculated as a proportion of total tree length. No significant difference in proportional trunk length was identified between MSS polyps with and without a focus of cancer ($U = 23$, $p = 0.953$). Average branch and leaf length was also calculated, then average branch and leaf length was converted into proportion of total tree length. A Mann-Whitney U test was then utilised to determine if branch length was significantly different in polyps with and without a focus of cancer. This test determined that branches were significantly longer in polyps with a focus of cancer ($U = 0$, $p = 0.004$). Finally, no significant difference in leaf length was found between polyps with and without a focus of cancer ($U = 10$, $p = 0.188$) (**Fig. 2.39**). MSI⁺ samples were excluded from this analysis as two polyps, out of a total of three MSI⁺ polyps, were from the same patient. The influence of this would not be accounted for in this analysis.

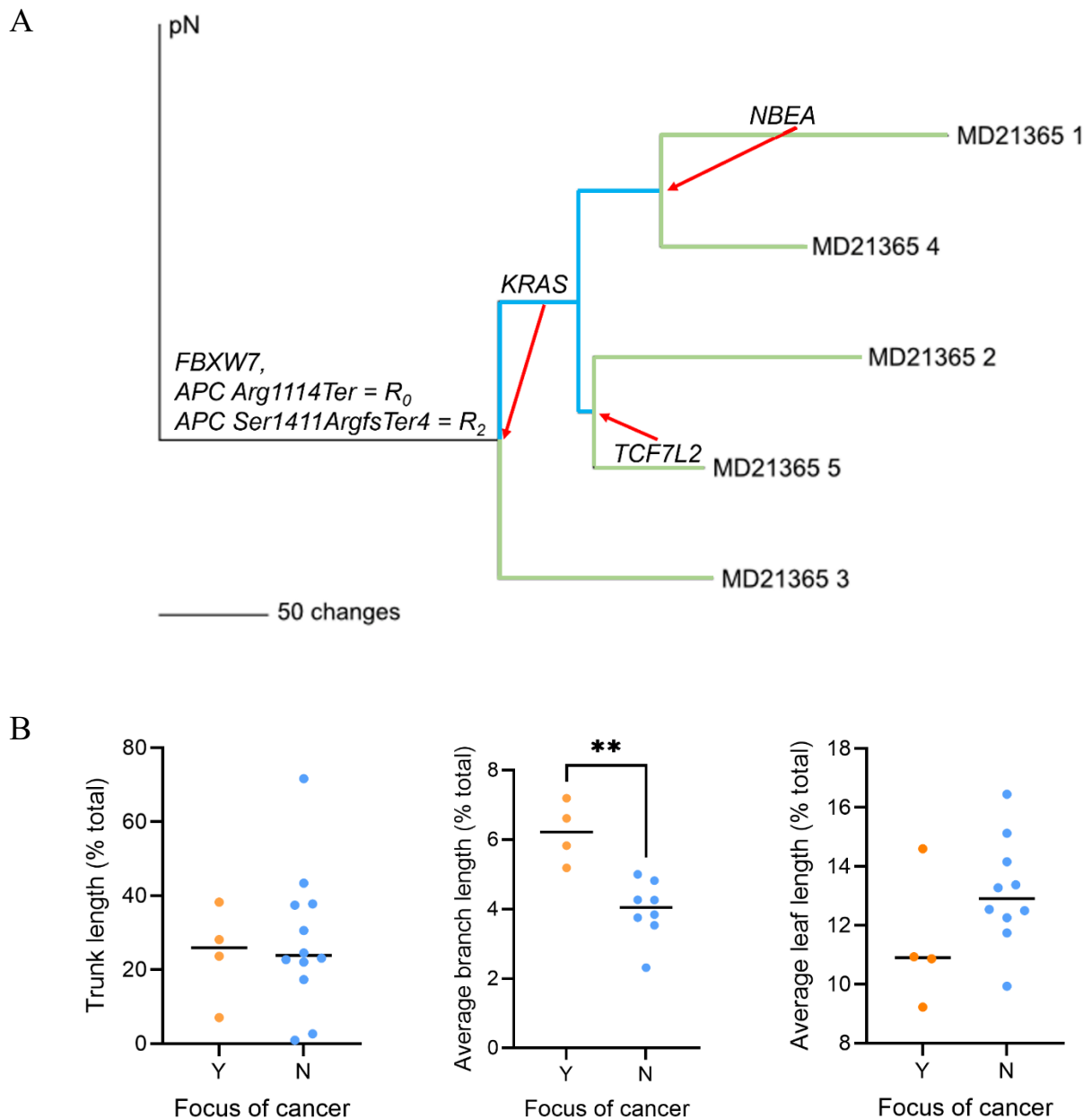


Figure 2.39 A comparison of relative trunk, branch and leaf lengths in MSS polyps with and without a focus of cancer. A) Shows where the trunk, branches and leaves are located. Black = trunk, blue = branches, green = leaves. B) Comparisons of phylogenetic tree segment lengths in polyps with and without a focus of cancer. Horizontal line = median. Y = Yes, polyp has a focus of cancer. N = No, polyp does not have a focus of cancer.

2.5 Discussion

The findings of this chapter challenge and expand upon current models of CRC evolution. According to the linear/Vogelstein model of CRC evolution, tumour progression occurs through the stepwise accumulation of driver mutations and occurrence of selective sweeps, continuously resulting in the production of a dominant clone and a monoclonal tumour (Fearon and Vogelstein, 1990b). The results of this chapter negate this theory in multiple ways. Firstly, through the examination of driver mutations in all 19 polyps, each individual polyp was shown to exhibit ITH, suggesting a lack of monoclonality through the occurrence of selective sweeps. This result is in direct conflict with the linear model. Clones with differing driver mutations appear to be able to co-exist without becoming a dominant clone in the tumour. Driver ITH and a lack of selective sweeps was also evidenced in the study by Cross *et al.*, (2018) suggesting that polyps of differing sizes exhibit ITH.

The presence of driver ITH is also a problem for the big bang model. The big bang model posits that all mutations capable of driving tumourigenesis are present from the very early stages of development. According to this model, each cell is “born to be bad” (Sottoriva *et al.*, 2015). However, in this chapter, as well as in the study by Cross *et al.*, (2018), the examination of multiple polyps revealed the sub-clonal occurrence of pathogenic driver mutations e.g. *KRAS*^{Gly13Asp} and *APC*^{Thr1556AsnfsTer3}. This indicates that these driver mutations arise later in tumour development than the big bang model suggests. Driver ITH is better explained by branched evolution, which suggests that different clones with different driver mutations are able to coexist and evolve in parallel, each clone representing a different route of tumourigenesis (Gerlinger *et al.*, 2014).

Although there is a lack of monoclonality and selective sweeps, this study showed that most polyps harboured mutations in driver genes predicted by the Vogelstein model such as *APC*, *KRAS*, *TP53* and 18q loss. Nine polyps harboured truncal mutations in *KRAS*, whilst nine additional polyps had sub-clonal *KRAS* mutations. A polyp (with a focus of cancer) that seemed to follow this evolutionary pattern was MD21239 that had clonal *APC* CN-LOH, clonal *TP53* CN-LOH, a clonal *KRAS* mutation and sub-clonal 18q allelic imbalance. Cancerous transformation was evidenced without the entire mutation complement, for example MD21354 and MD21354 possessed a focus of cancer but not the entire complement of driver mutations as suggested by the Vogelstein model. Further research into the timing of these mutations, e.g. via MutationTimer would provide further insight into whether the mutations occur in the order predicted by the Vogelstein model (Gerstung *et al.*, 2020), although Cross *et al.*, (2018) did not find evidence for this.

Interestingly, the frequency of *FBXW7* mutations is notable, with clonal mutations observed in five polyps and sub-clonal mutations present in an additional four polyps. Despite *FBXW7* being included as a Wnt-pathway mini-driver in previous literature, its frequency might suggest a greater role for its mutation in CRC development (Brunet *et al.*, *). In addition, MD21296 displaying a sub-clonal

KRAS mutation, it is questionable whether this polyp should be viewed and analysed in the same way as the other polyps as this polyp showed evidence of inflammatory origin. Evidence suggests that polyps of inflammatory origin show distinct genetic differences to other polyp types. For example, a study by Chatila *et al.*, (2023) showed a lack of Wnt pathway alteration in colitis-associated cancers (Chatila *et al.*, 2023).

Notably, no difference was found in the proportion of unique mutations in MSS polyps with and without a focus of cancer. This implies that the transition from non-cancerous polyp to a cancerous state does not necessarily involve a broad increase in the number of new, unique mutations across the tumour population. This analysis provides more information regarding differing levels of ITH throughout cancer development, as Saito *et al.*, (2018) found significantly higher levels of ITH in MSS precursor lesions (polyps with a focus of cancer) to advanced CRCs. This is also supported by Cross *et al.*, (2018) who found significantly higher ITH in MSS CRCs when compared with MSS CRAs. No difference in ITH in pre-cancerous polyps and polyps with a focus of cancer is supportive of the big bang model which predicts that a high level of ITH will be present from an early stage of tumour development (Sottoriva *et al.*, 2015).

As stated above, the analysis of ITH by comparing the proportion of unique variants indicates no significant difference in ITH whether a polyp develops a focus of cancer or not. However, this approach is the equivalent to comparing only the lengths of leaves on phylogenetic trees of polyps with and without a focus of cancer, as it focuses exclusively on mutations present in individual samples, not on sub-clonal mutations present in two or more samples. Phylogenetic analysis corroborates this finding, revealing no significant difference in the lengths of leaves between phylogenetic trees of MSS polyps with and without a focus of cancer. This consistency suggests that the rate and extent of mutation acquisition remain similar at this stage of tumourigenesis, regardless of cancer development.

When comparing the proportion of sub-clonal driver mutations; any driver mutation that was not clonal, between polyps with and without a focus of cancer there was no significant difference. Nonetheless, a significant difference was observed in the branch lengths of polyps with and without a focus of cancer. Polyps that develop a focus of cancer exhibit significantly longer branches compared to those without, indicating an accelerated rate of mutation accumulation and genetic diversity early in tumour development following the most recent common ancestor of all clones. This disparity suggests that critical genetic diversifications, potentially involving key driver mutations, occur during the initial stages of tumour development in polyps that progress to cancer. It may therefore be of interest to carry out another statistical test to determine if there is a difference in the number of sub-clonal driver mutations using an altered definition of driver mutation. Cornish *et al.*, (2024) (of which I am a contributing author) identified over 250 putative CRC driver genes. It may be of interest to investigate

whether the addition of these drivers provides more insight into driver ITH than using only Intogen CRC driver genes.

Further investigation into the nature of these early mutations and their biological impacts could provide valuable insights into tumour development. For example, it would be possible to compare the ratio of non-synonymous mutation to synonymous mutations situated on the trunks, leaves and branches of phylogenetic trees of polyps with a focus of cancer. If the ratio of non-synonymous mutations to synonymous mutations increases on the branches of the trees, followed by a decrease on the leaves, this would indicate that positive selection happens at the branch stage of tumour development, followed by ITH as a result of neutral evolution on the leaves (Cross et al., 2018). This would provide additional support for the big bang model, which suggests that following a big initial expansion, further evolution is neutral and further ITH is driven by mistakes in DNA repair mechanisms. It would also be of great insight to analyse any differences in mutation signatures between polyps with and without a focus of cancer to establish whether these mutations are driven by selection rather than underlying mutational processes. Cross et al., (2018) analysed mutational processes in CRAs and CRCs and did not find any evidence for an association with enhanced evolution.

Another modification to the study design would be to increase the number of samples taken from each polyp. This would allow for a more detailed examination of ITH. By increasing the number of samples taken from each polyp, this would increase the power of the analysis sufficiently to compare variations in leaf length between multiregional samples from the same polyp, potentially allowing for the identification of differing rates of genetic diversity and mutation accumulation in different regions of the polyp. Ideally, a power calculation would be utilised to determine how many samples are required to ensure that meaningful comparisons can be made between regions. By predicting and inputting effect size and desired power (~80%), these calculations ensure the study has enough statistical power to detect significant differences between samples such as mutation burden or significant differences in mutation type.

Samples could also be taken from a greater depth within the polyp itself. The samples utilised in this analysis were all taken from the surface of the polyp, potentially missing out genetically different populations located deeper within the polyp. Stage and level of invasion of the polyps with a focus of cancer would also be of great interest to incorporate into future studies. If incorporating tumour stage into the analysis, dissecting samples from different depths of the polyp is of high importance as this has the potential to capture cells able to invade into the submucosa and muscularis propria ($\geq T_1$), which may harbour different mutations to those on the surface of the polyp. This can be accomplished by carefully excising tissue samples from distinct layers, ensuring that both superficial and deeper areas of the polyp are included. Additional comparisons could also be made between polyps with high and low-grade dysplasia as this is indicative of progression along the pathway towards cancer (Chang *et al.*, 2020).

Notably, one of the main findings of the Cross *et al.*, (2018) paper was the occurrence of punctuated evolution through the increased number of CNAs found in CRCs when compared to CRAs. This finding was not replicated in the current study, perhaps due to a different stage of sampling. This could indicate that punctuation events occur later in tumour development.

Multiple polyps showed supportive evidence for the just-right model in Wnt-pathway related genes. Out of a total of 19 polyps, nine polyps had two detectable, clonal hits to *APC*. Each of these nine polyps had between one and four 20AARs remaining in *APC* following two alterations. As stated in Section 2.1, rather than total inactivation of *APC*, the just-right model predicts that an optimal number of 20AARs will be selected to adjust the level of Wnt activity. A β -catenin inhibitory domain (CID) between 20AAR1 (R1) and 20AAR2 (R2) downregulates levels of β -catenin. Evidence suggests that removal of the CID is selected, however, neoplasms still rely on the binding of β -catenin to R1 for tumour development, providing supporting evidence that intermediate Wnt-activity is selected for. Multiple polyps showed this exact relationship between *APC* hits. For example, MD21031 had two clonal *APC* mutations, one in region 0 (R₀) and another in region 1 (R₁). MD21320_1, MD21340 and MD21365 all had two 20AARs remaining following two clonal *APC* alterations. Interestingly, MD20716, MD21274, MD21353 and MD21386 all had three 20AARs remaining following two clonal *APC* alterations. MD21236 had four 20AARs remaining. The most common number of 20AARs in this dataset was between two and three. This supports the just-right model as the model predicts a tendency for a minimum of one 20AAR to be maintained for the purpose of adjusting levels of Wnt signalling.

It is of importance to assess this model in the context of the wider Wnt pathway. In many of these polyps, there was the presence of an additional, clonal mutation in a Wnt pathway mini-driver gene, followed by a gain of additional sub-clonal Wnt mini-driver mutations. Hypothetically, a mutation in a Wnt pathway mini-driver gene could adjust for a sub-optimal selection of 20AARs in *APC* resulting in a correlation between *APC* mutations and mutations in Wnt pathway mini-driver genes. This is generally supportive of the just-right and mini-driver models which predict a correlation between mutations in major-driver and mini-driver genes, and a pattern of mutual exclusivity between mutations in major-driver genes. Although generally there is evidence of co-occurrence between mutations in mini-driver and major-driver genes, supporting the just-right model, there are a range of different combinations of mini-driver mutations with different numbers of 20AARs. Although, the most frequent combination was presence of a clonal *FBXW7* mutation in three polyps with three 20AARs. This data is limited by sample size. Patterns of co-occurrence between mutations in mini-driver and major-driver genes would benefit from exploration using a much larger dataset. This is carried out in chapter 3. Notably, only MD21320_1 harboured mutations in two different Wnt pathway major-driver genes, *APC* and *RNF43*, that previous studies have reported to be mutually exclusive. The usually mutually exclusive relationship between mutations in these two major-driver genes is again supportive of the just-right model as they theoretically provide too much aberration to Wnt signalling to be beneficial to the

cell (Segditsas and Tomlinson, 2006). This polyp is however MSI⁺ and therefore has a hypermutator phenotype. In addition, there is controversy regarding the pathogenicity of the mutation that was present in *RNF43* (*RNF43^{Gly659ValfsTer41}*) (Tu *et al.*, 2019).

The just-right model proposes that only certain branches and sub-clones harbouring the right combination of mutations, resulting in a just-right level of pathway activation, are likely to drive tumour progression. This phenomenon may occur via convergent evolution, meaning that although different mutations may occur within the same gene, the key factor is achieving the precise level of pathway activation that is “just-right” for tumour development, this can be achieved through varying mutations within the same genes (Segditsas and Tomlinson, 2006). An example of convergent evolution can be found in MD20980 in which a *GNAS^{Arg844Cys}* mutation is observed in sample 5 and a *GNAS^{Arg844His}* mutation is observed in sample 3.

Although none of the polyps were classified as serrated in their pathological reports, several polyps lacked two detectable clonal alterations in *APC*. There are a few reasons why this might have occurred. Firstly, this analysis was limited to somatic mutations and CNAs. It’s important to note that *APC* function can also be lost epigenetically through hypermethylation of its promoter (Zhu *et al.*, 2021). Additionally, to normalise the differences in read depth between polyps sequenced using WGS and WES, down-sampling was implemented. This method could lead to under-detection of certain mutations potentially causing some driver mutations to appear sub-clonal. It could also affect the detection of low frequency variants only present in a small fraction of cells which would require greater read depth for accurate identification. This issue might have influenced the observed levels of ITH and the ability to accurately assess the just-right model. To address this, bam files were examined using the Integrative Genomics Viewer (IGV) to identify mutations that appeared in the majority of samples but were not identified as clonal through standard variant calling.

Notably, two polyps exhibited only one detectable, clonal alteration in *APC*, alongside another sub-clonal alteration. It remains uncertain whether this scenario is due to detection limitation, suggesting that the second mutation might actually represent a third *APC* alteration, or if it reflects a section of the polyp that developed through an alternative mechanism of *APC* inactivation. To resolve these ambiguities, conducting WGS at a higher depth across all samples in future studies would be crucial for more accurate mutation detection and assessment of the just-right model. Additionally, investigating epigenetic modifications of *APC* would also be invaluable.

According to the linear model, transition from CRA to CRC occurs due to a stepwise accumulation of driver mutations. If true, this would result in CRAs and CRCs being distinguishable from one another via mutation burden. Through a comparison of polyps with and without a focus of cancer, changes in mutation burden can be assessed through the course of these early evolutionary stages. No significant difference in either mutation burden or driver mutation burden was evidenced,

contrary to the linear model. This result is also reflected in the study by Cross *et al.*, (2018) who showed no significant difference in either mutation burden or driver mutation burden between CRAs and CRCs. One limitation of the approach utilised in the Cross *et al.*, (2018) study, which is replicated in this chapter, is the exclusive focus on *APC* and *TP53* when analysing CN-LOH and monosomy, thereby overlooking other TSGs that contribute to CRC through similar mechanisms. For example, PTEN is a TSG in which biallelic inactivation occurs frequently in CRC (Molinari and Frattini, 2014). By restricting the analysis of CN-LOH and monosomy to *APC* and *TP53*, this methodology does not fully investigate whether known TSGs have undergone biallelic inactivation.

It is of importance to note that a key difference between the Cross *et al.*, (2018) study and the sampling methodology in this chapter is that this analysis is based on the multiregional sequencing of large polyps that do or do not harbour a focus of cancer, in contrast to the direct comparison of CRA versus CRC conducted by Cross *et al.*, (2018). Due to this, we compare CRAs to polyps with a focus of cancer to provide an overview of early-stage development, as opposed to Saito *et al.*, (2018) who compared polyps with a focus of cancer to CRCs. By carrying out sampling this way, it is unclear whether the multiregional samples from polyps with a focus of cancer, are cancerous or if they are an admixture of cancer and adenoma, or a sample of adenoma adjacent to a focus of cancer. One way to improve upon the current sampling technique is by accessing the block in which the polyp is mounted, carrying out H&E staining and extracting DNA from benign and cancerous regions via histopathological identification (Mamlouk *et al.*, 2020). This method was unfortunately out of the scope of this project. However, Cross *et al.*, (2018) argue that malignancy is difficult to decipher cytologically as malignancy is better distinguished via invasion and location of tumour cells. They argue that different neoplastic components may both be cancerous.

It has been shown in previous studies that larger polyps of over 10mm are more likely to contain early cancer (Rembacken *et al.*, 2000). Only four of the large 19 polyps progressed to a cancer. This suggests that although there is an association, size alone is not the only predictor of progression. In addition, within this population of large polyps, polyps ranging in size developed a focus of cancer. This further underscores that when looking at polyps over 20mm, it is not size alone that dictates the development of a cancer.

The relationship between polyp size and mutation burden was investigated. This identified an inverse relationship between polyp size and mutation burden; the smallest polyps harboured the greatest mutation burden. An increased number of mutations could confer a selective growth disadvantage due to the occurrence of too much genetic instability, or even through increased interactions with the tumour microenvironment, resulting in the production of a small polyp with a high mutation burden. One such growth disadvantage is the production of neoantigens as a result of a high mutation burden. This results in the cells being targeted by activated immune cells. This may result in a decrease in growth rate due to interactions with the surrounding microenvironment (Chalmers *et al.*, 2017). Fox and Loeb (2010)

also suggest the existence of a mutational threshold, above which lethal mutagenesis occurs (Fox and Loeb, 2010).

In contrast, larger polyps may grow through mechanisms involving fewer mutations that effectively enhance growth and survival without causing too much genetic instability. This may promote a more sustained growth rate. This is perhaps another example of the just-right model in action. Despite mutation burden having a negative association with polyp size, further analysis in this chapter and in studies by both Cross *et al.*, (2018) and Saito *et al.*, (2018) showed no association between mutation burden and progression to cancer suggesting that investigation into the action of specific mutations is more important than the number of mutations, contrary to the main premise of the linear model.

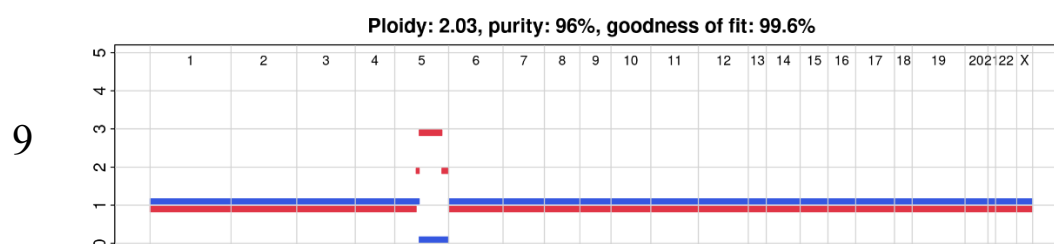
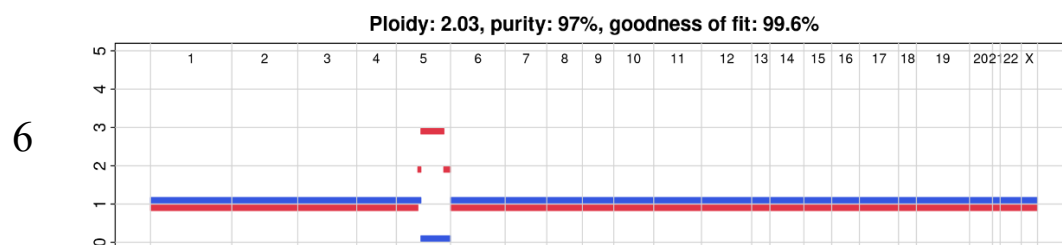
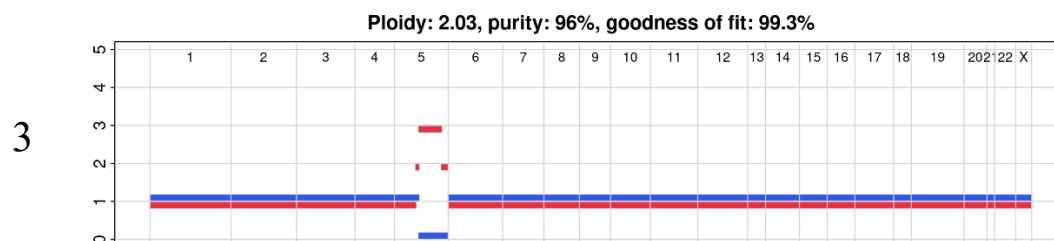
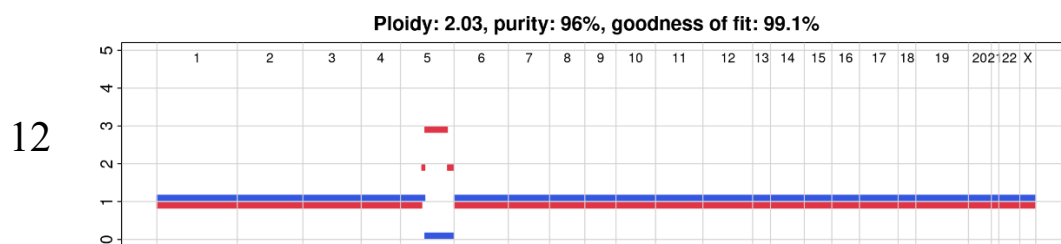
These results suggest that even though larger polyps have an enhanced risk of developing a cancer, as suggested by previous studies, not all large polyps progress and therefore there are additional complexities that contribute to its occurrence. Although mutation burden may differ between polyps of varying sizes, evidence suggests that mutation burden does not necessarily influence cancer progression and so further research into potential reasons for transition to a cancer in large polyps is required.

Altogether, these results show a lack of sufficient evidence for the linear model of evolution through a lack of monoclonality and selective sweeps. Although, mutations do accumulate in the driver genes as suggested by the linear model, cancers can occur without the full complement and drivers exist in sub-clonal regions. Evidence of driver ITH is also problematic for the big bang model of CRC evolution which suggests that all cells in a tumour should harbour mutations capable of driving tumourigenesis from a very early stage of development. There is no evidence for a punctuation event in these tumours via CNAs or WGD. When incorporating this result with the findings of Cross *et al.*, (2018), this suggests that punctuated events may occur later in tumour development than is visible in these samples at an earlier stage. These results do show an increased level of ITH at branch level, following the most recent common ancestor in polyps with a focus of cancer. These results are better described by branched and parallel evolution in which clones with varied driver mutations can exist and evolve within the same tumour. Evidence for the just-right model was also seen in multiple polyps through the selection of a just-right number of 20AARs in *APC*; a pattern of co-occurrence between *APC* and putative Wnt pathway mini-drivers and a mutually exclusive relationship between pathogenic mutations in major-driver genes *APC*, *RNF43* and *CTNNB1*. Although this dataset is of limited size, evidence for an association between mutations in mini-driver mutations and a particular number of 20AARs in *APC* would benefit from exploration using a much larger dataset. This is carried out in the following chapter.

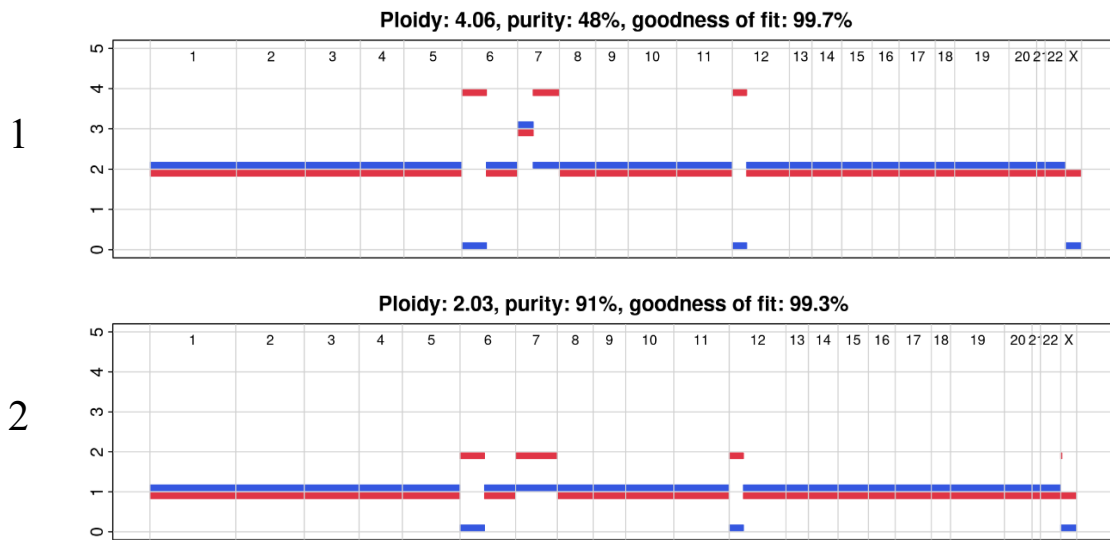
Appendix 2

76 Intogen colorectal cancer driver genes: *APC*, *TP53*, *KRAS*, *PIK3CA*, *SMAD4*, *FBXW7*, *TCF7L2*, *FAT4*, *AMER1*, *LRP1B*, *BRAF*, *SOX9*, *ATM*, *NRAS*, *ARID1A*, *KMT2C*, *PCBP1*, *EPHA3*, *BCL9L*, *BCL9*, *SMAD2*, *ERBB2*, *SMAD3*, *PTEN*, *RNF43*, *HSPG1*, *CTNNB1*, *BRCA2*, *MTOR*, *GNAS*, *PTPRC*, *SMARCA4*, *ERBB3*, *TGIF1*, *MAP2K4*, *DUSP16*, *MAP2K1*, *SIRPA*, *ACVR2A*, *FANCD2*, *RBM10*, *ARID1B*, *MYH9*, *TBX3*, *NF1*, *BMPR1A*, *BCOR*, *NIN*, *USP6*, *EP300*, *EGFR*, *FLCN*, *PTPN11*, *PARP4*, *EIF3E*, *MAP3K1*, *BIRC6*, *NBEA*, *FAT3*, *CCDC6*, *RNF6*, *NCOR2*, *HRAS*, *GRIN2A*, *WNK2*, *PPP2R1A*, *TGFBR2*, *MTCPI1*, *CARD11*, *DICER1*, *BCORL1*, *LATS1*

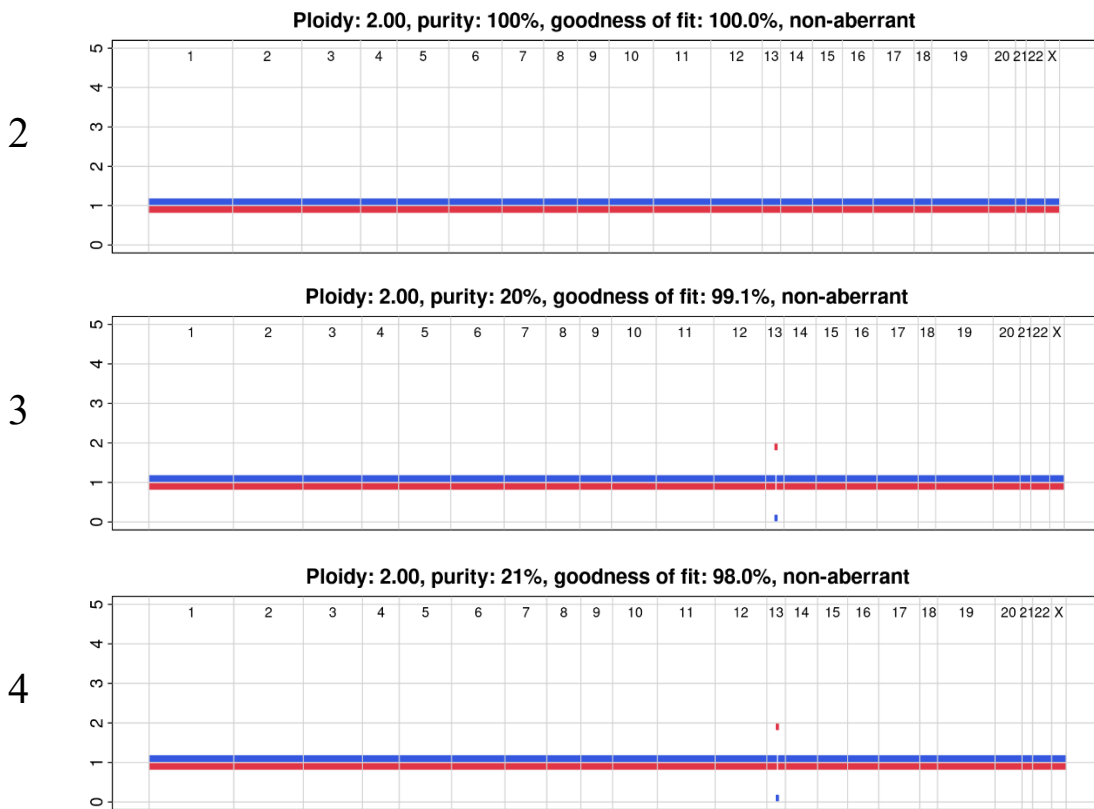
MD20716



MD20430



MD20730

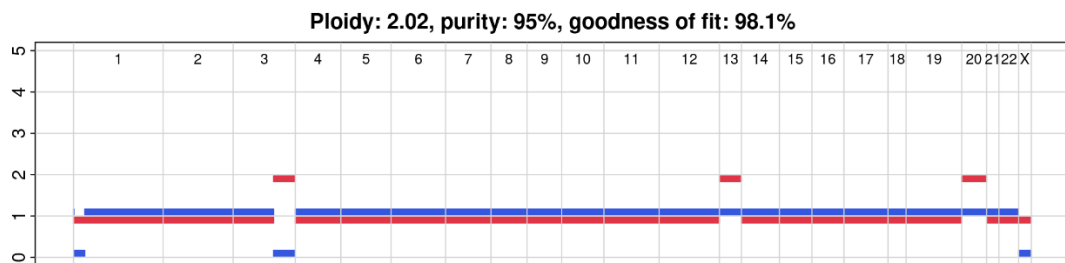


5

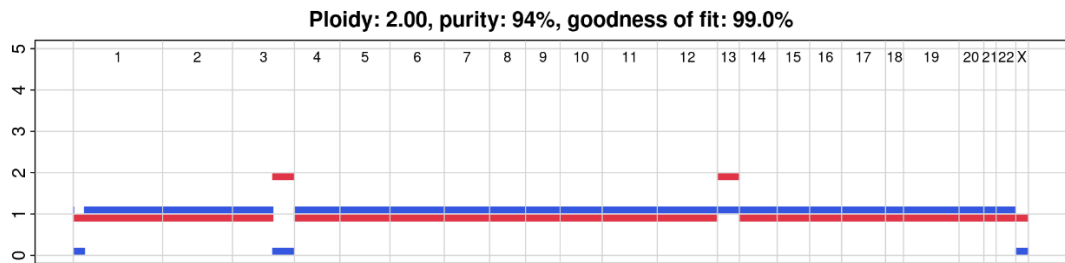


MD20980

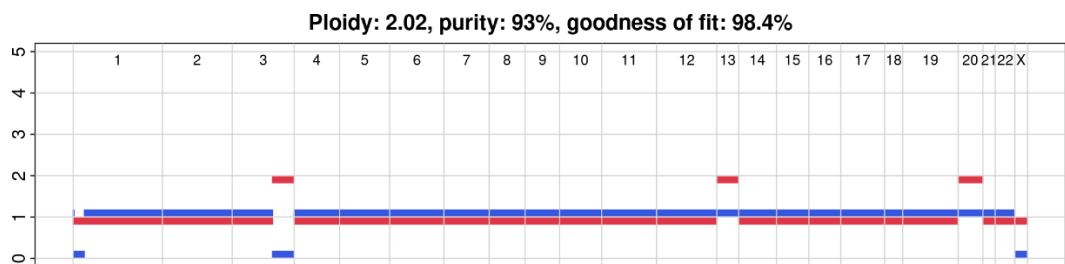
1



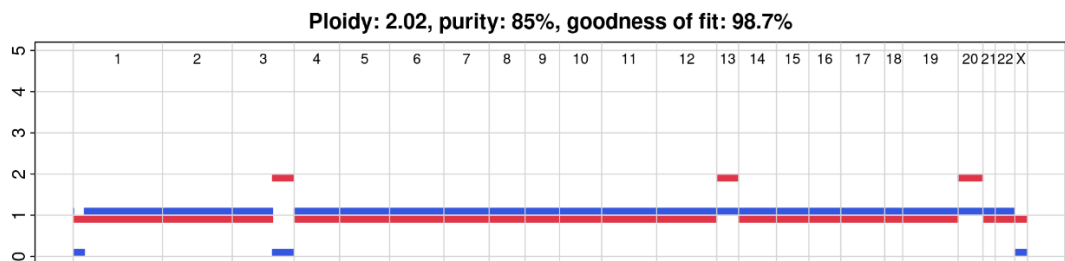
2



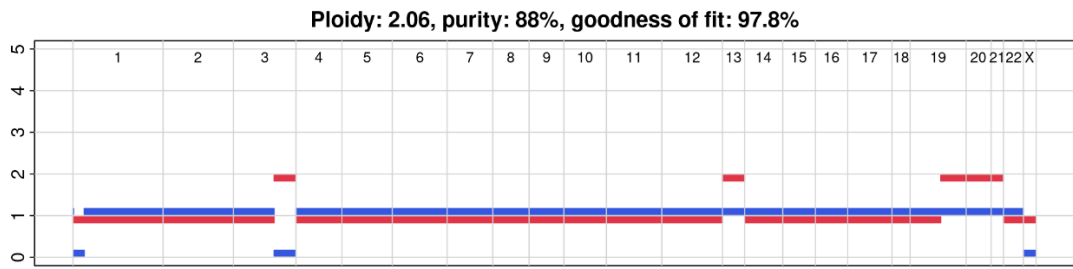
3



4



5



MD21031

1



2



3

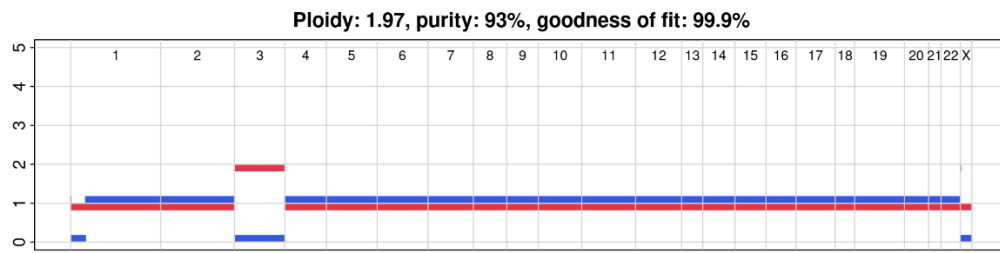


4

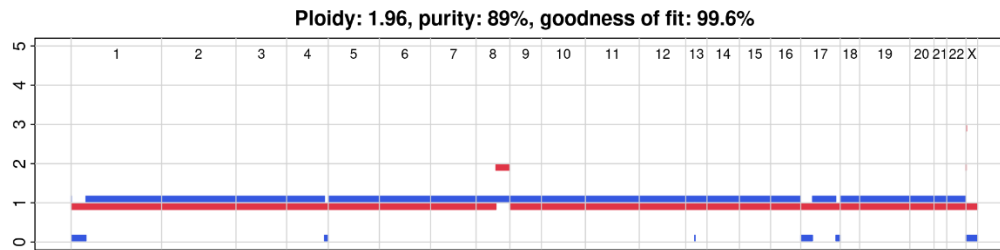


MD21032

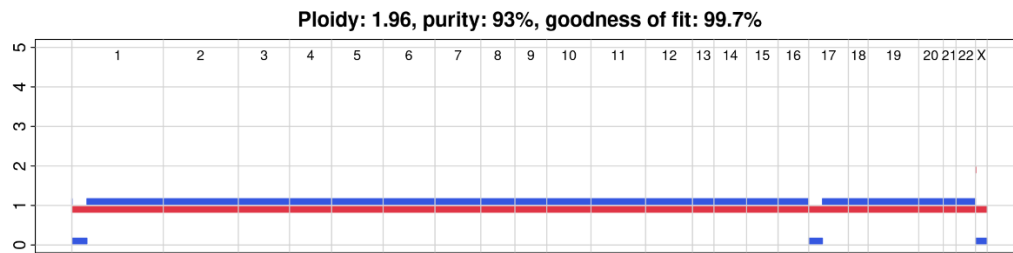
1



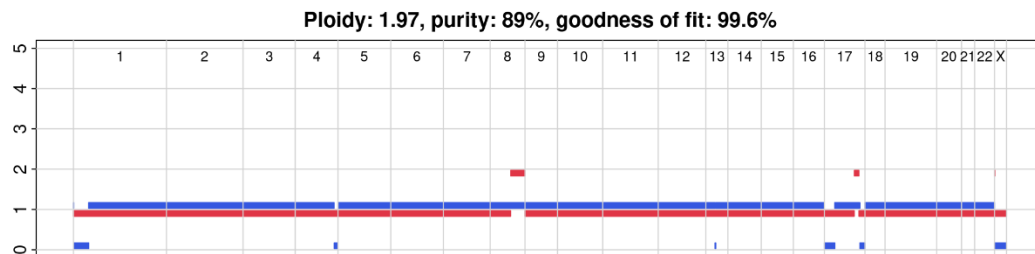
2



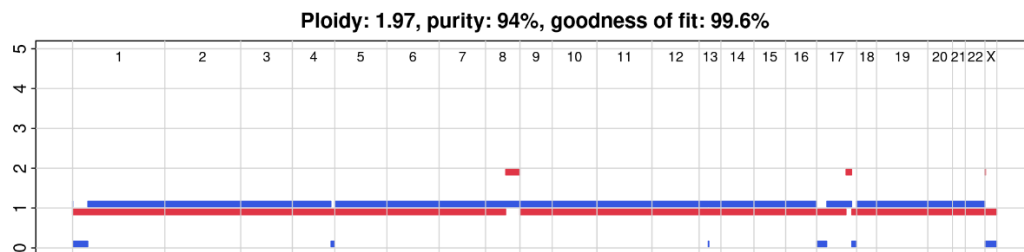
3



4

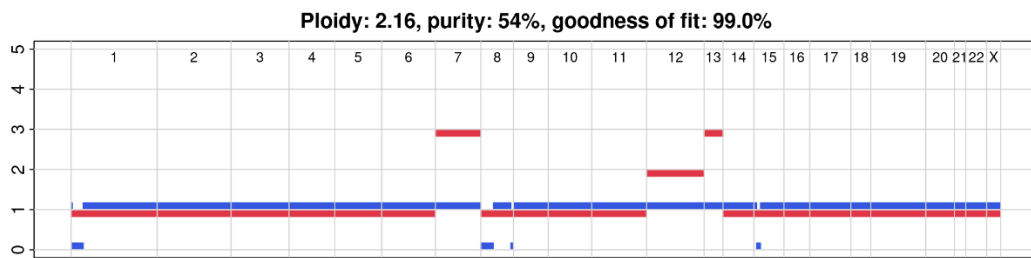


5

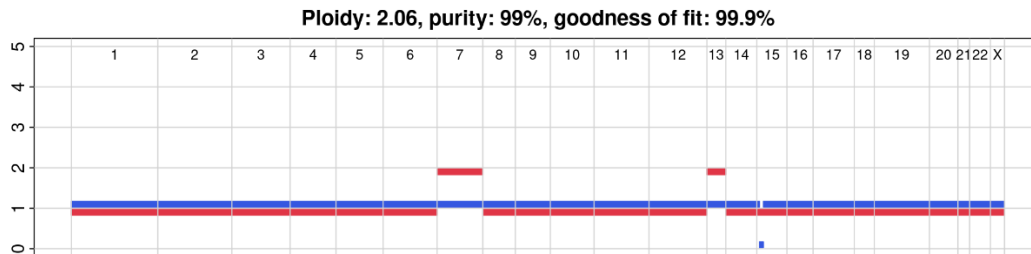


MD21099

1



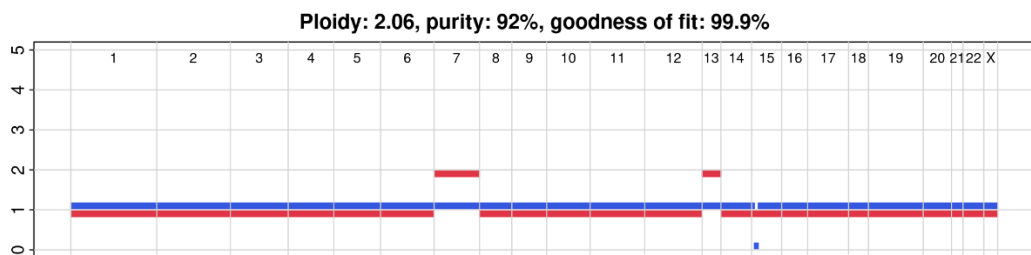
2



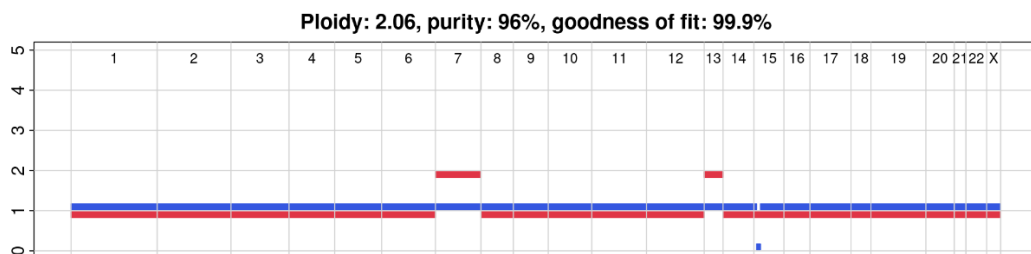
3



4

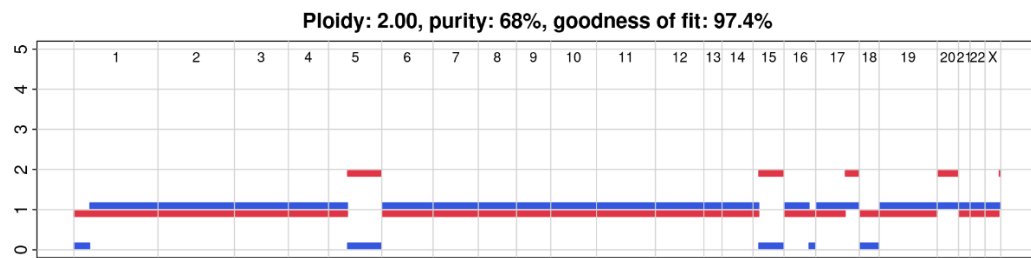


5

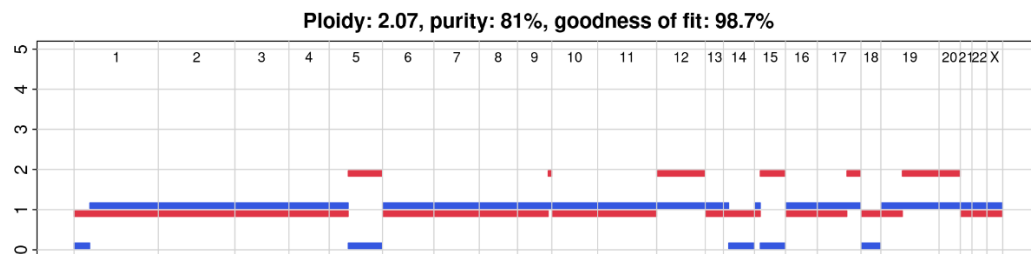


MD21236

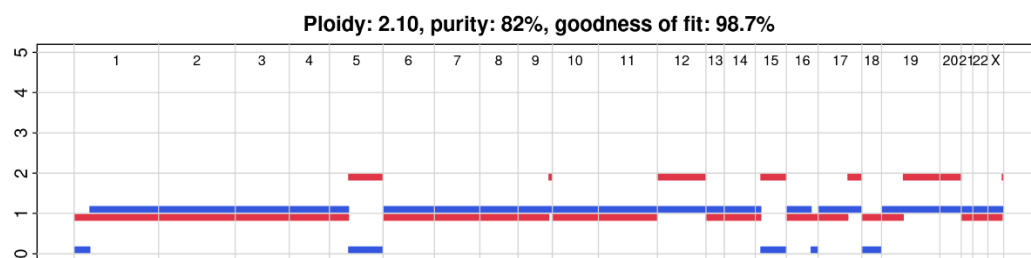
1



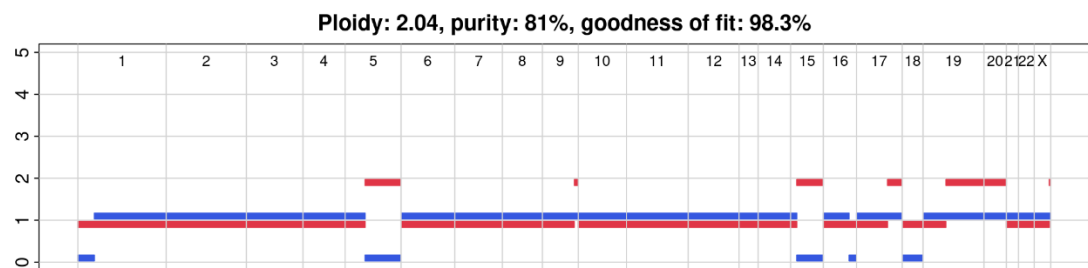
2



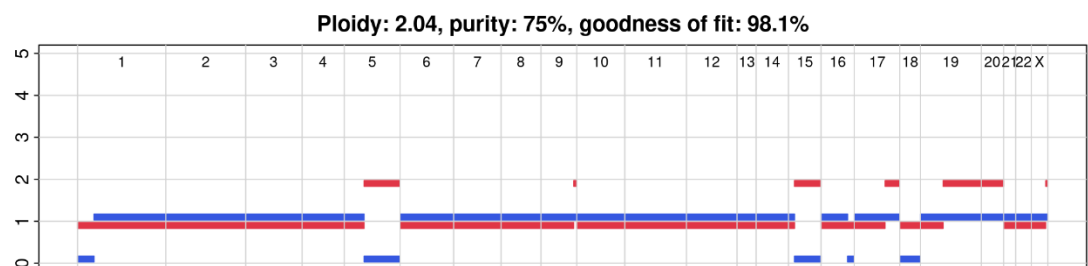
3

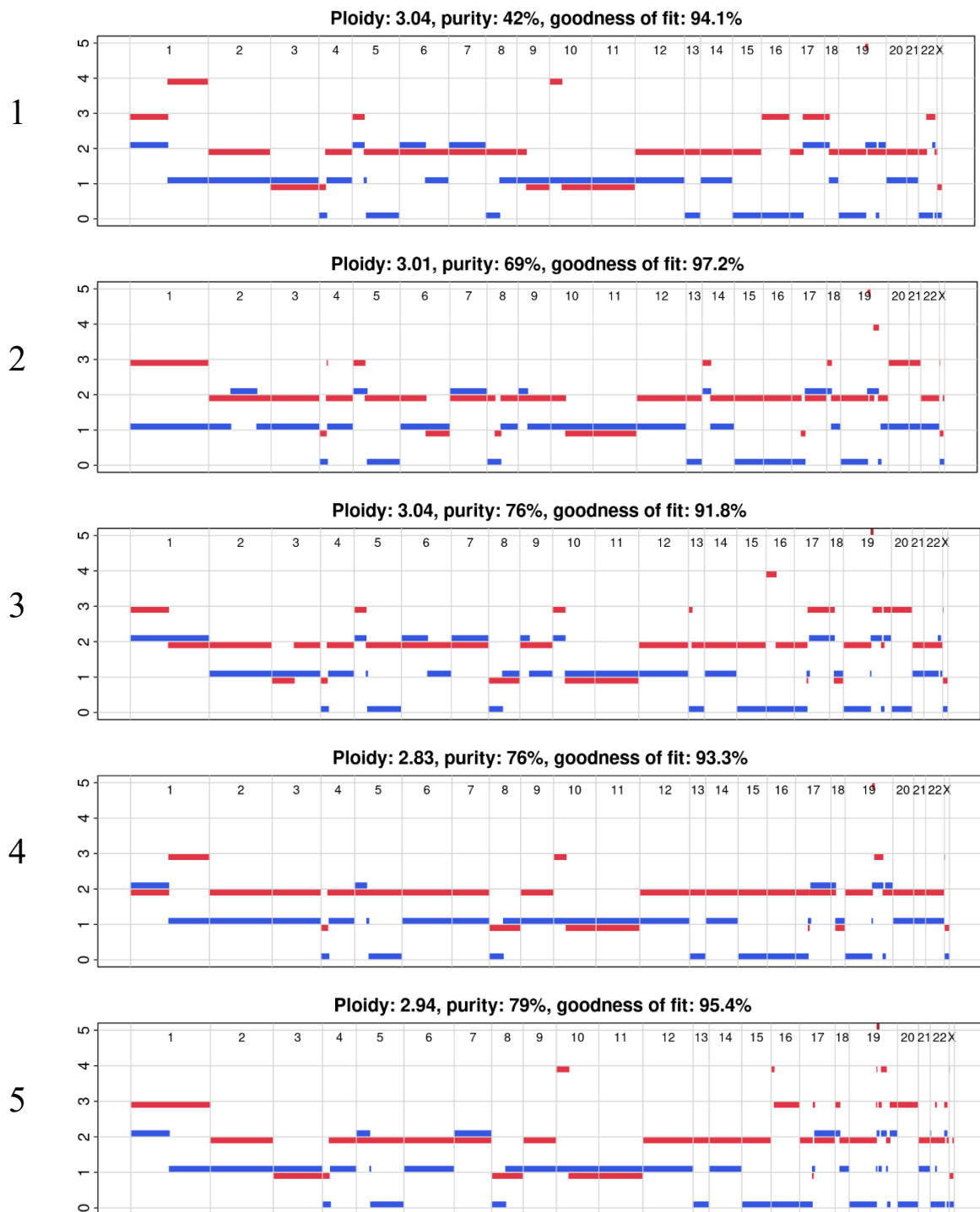


4



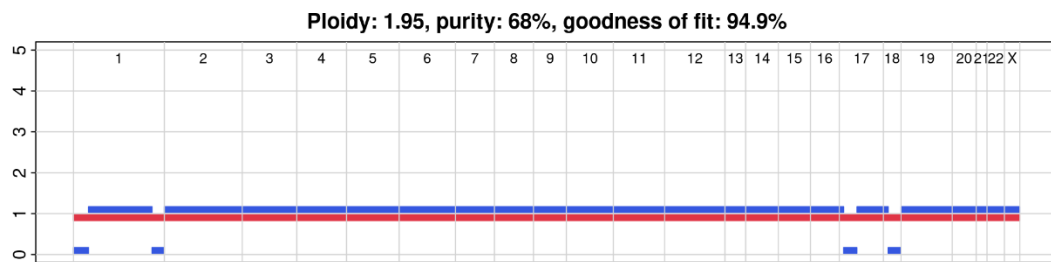
5



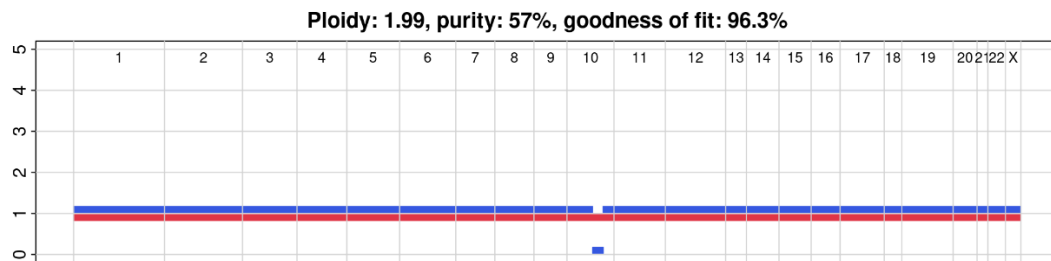


MD21257

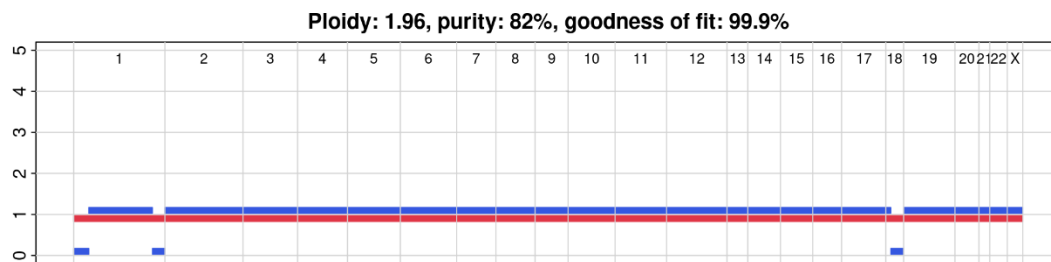
1



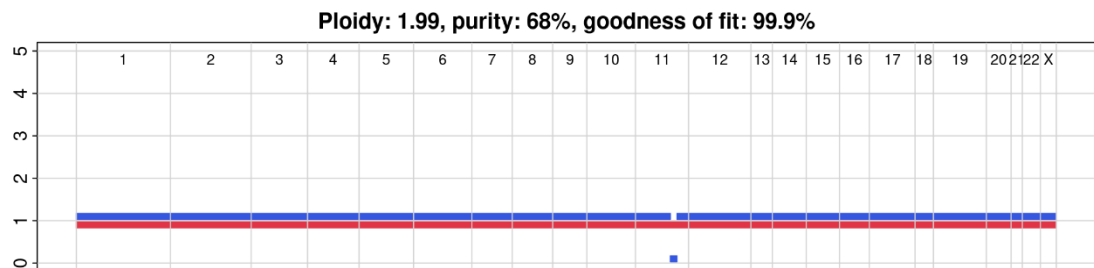
2



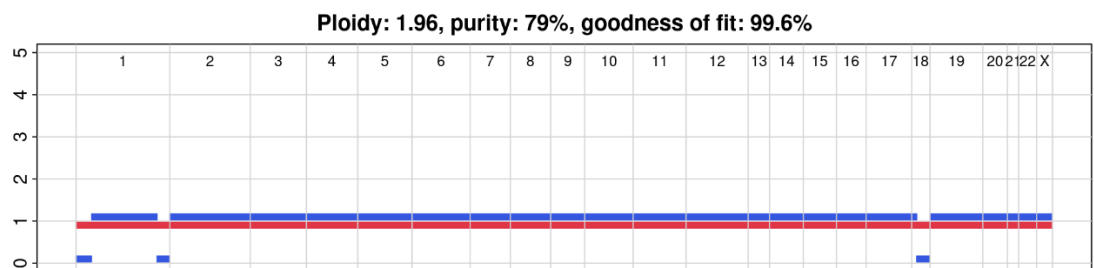
3



4

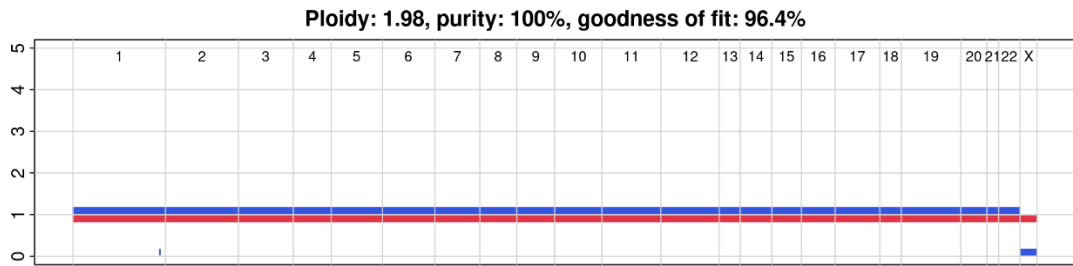


5

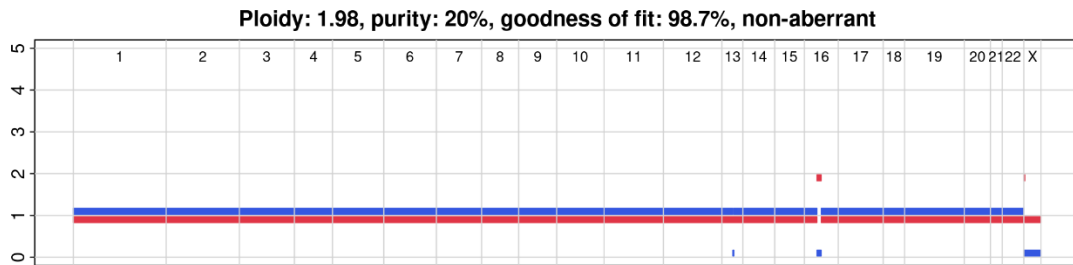


MD21274

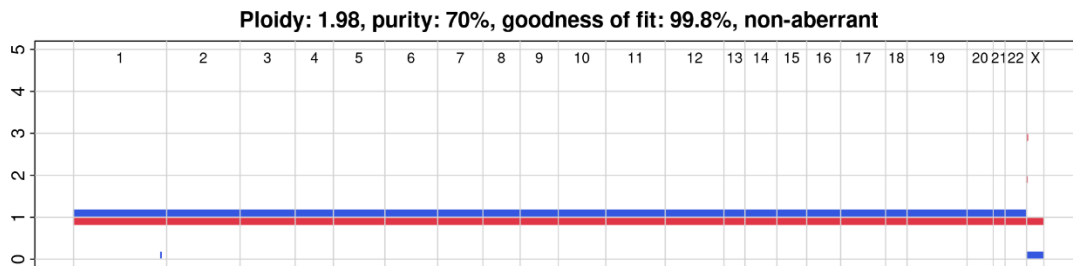
1



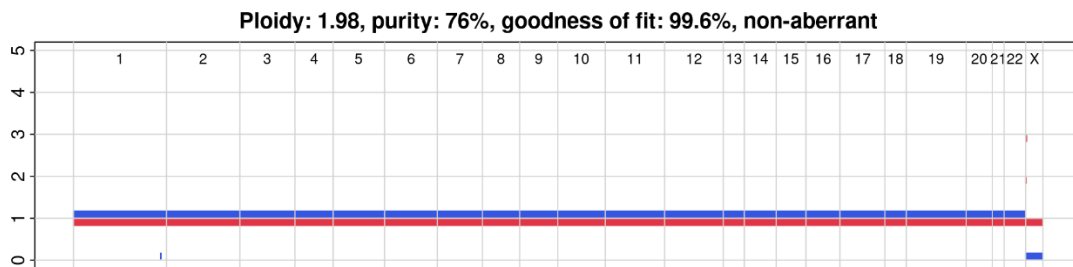
2



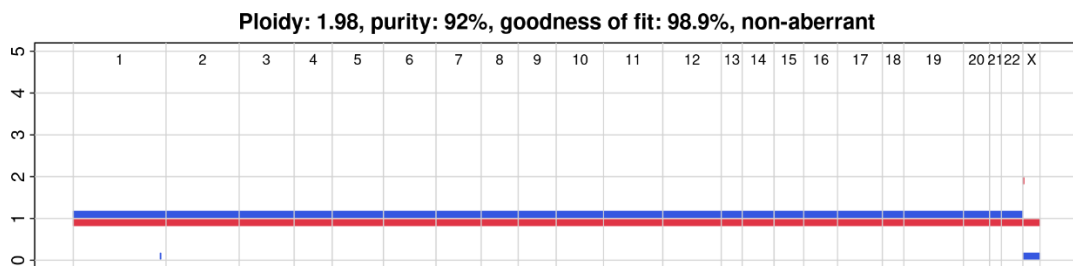
3



4

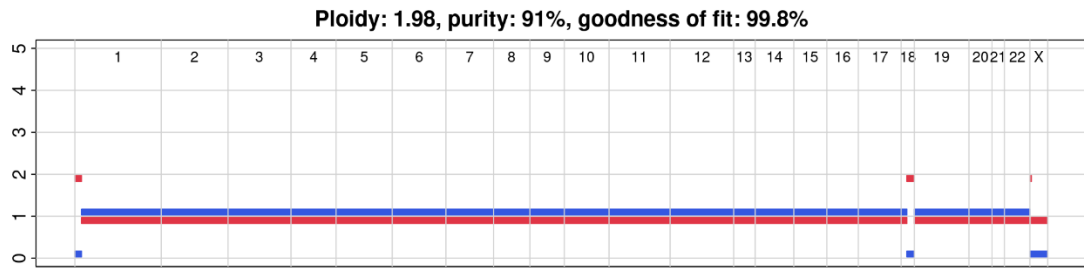


5

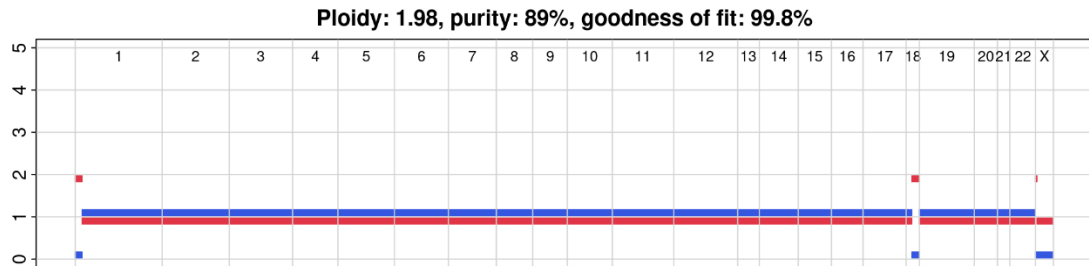


MD21296

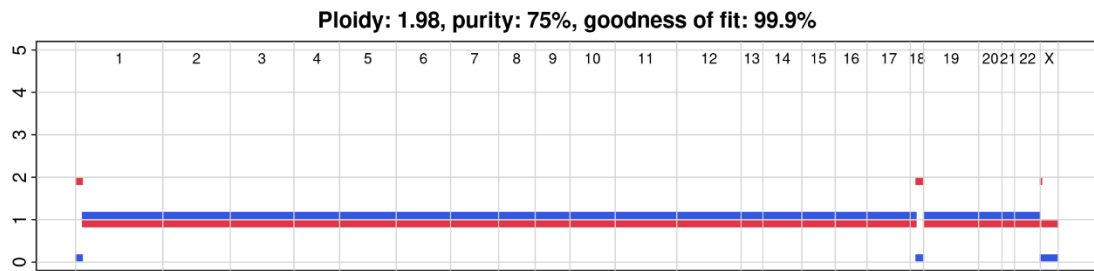
1



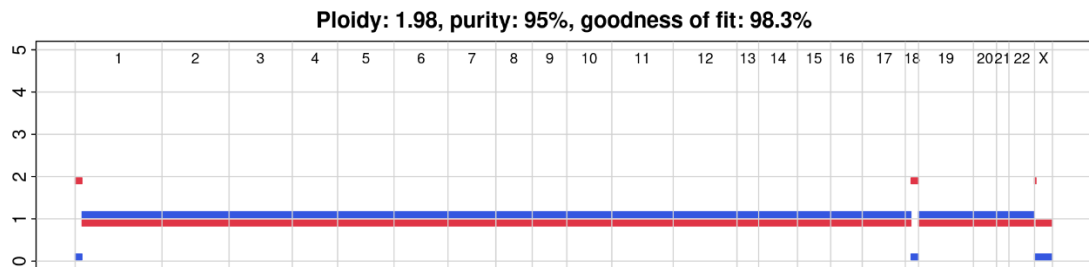
2



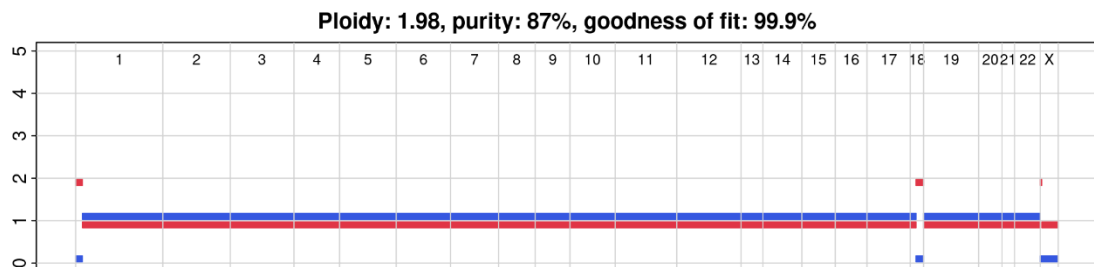
3



4

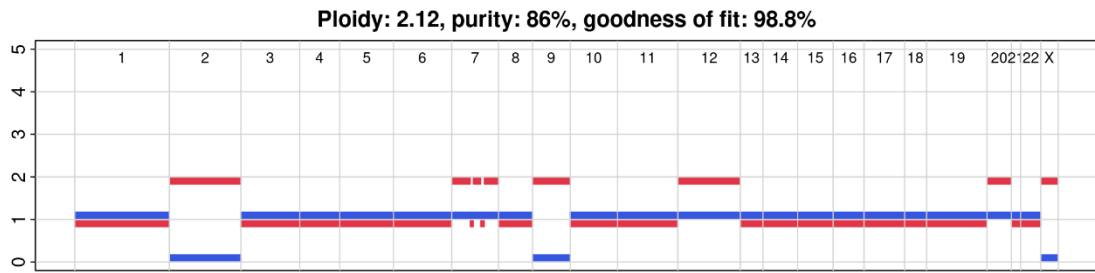


5

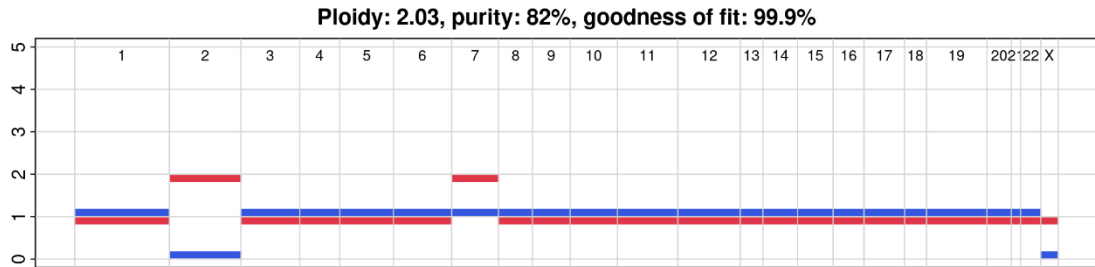


MD21365

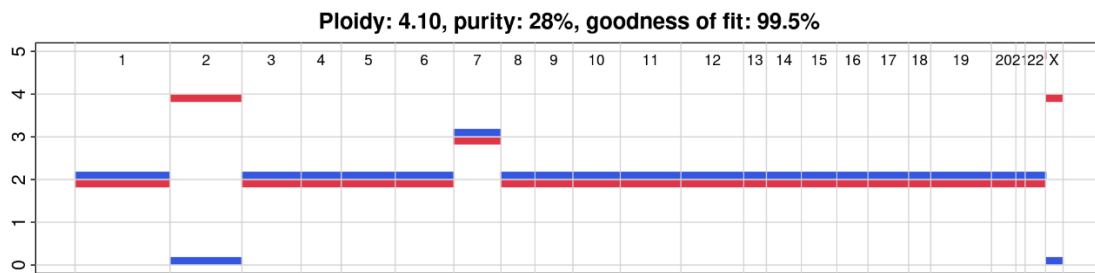
1



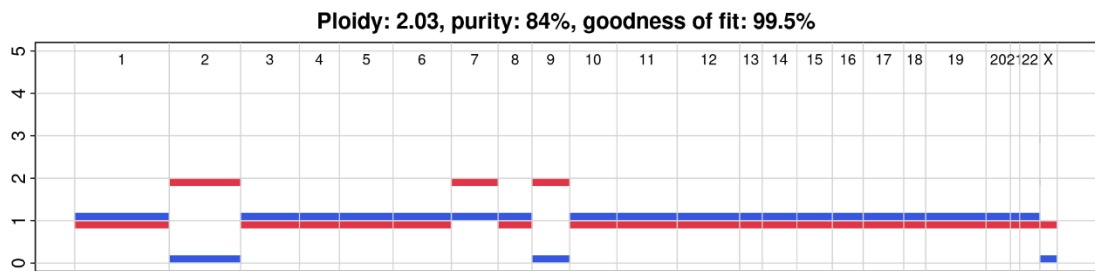
2



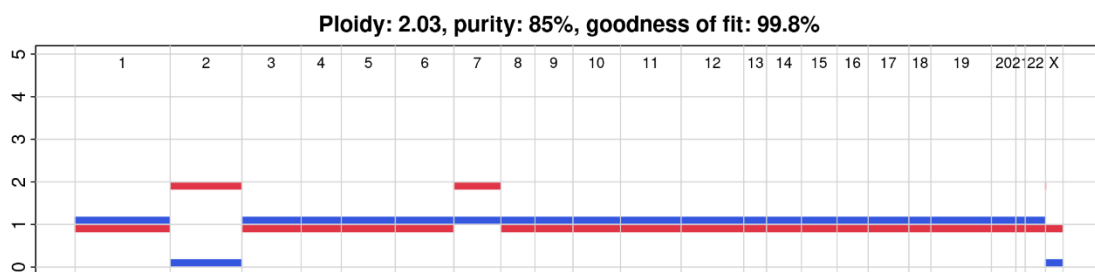
3



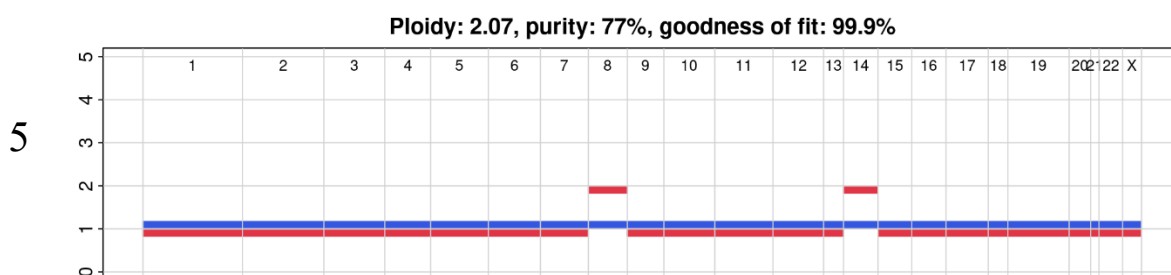
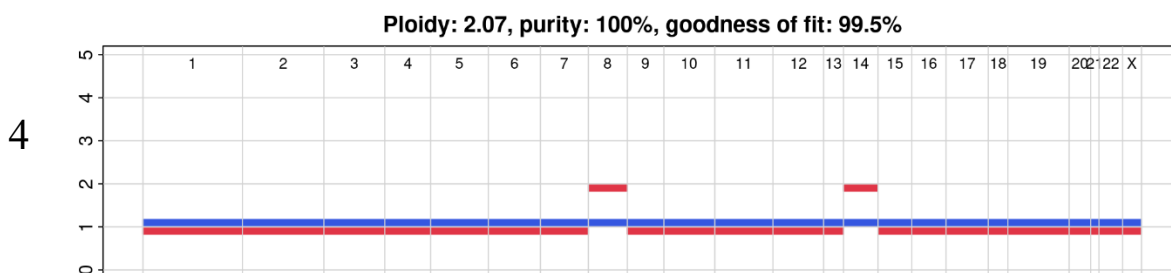
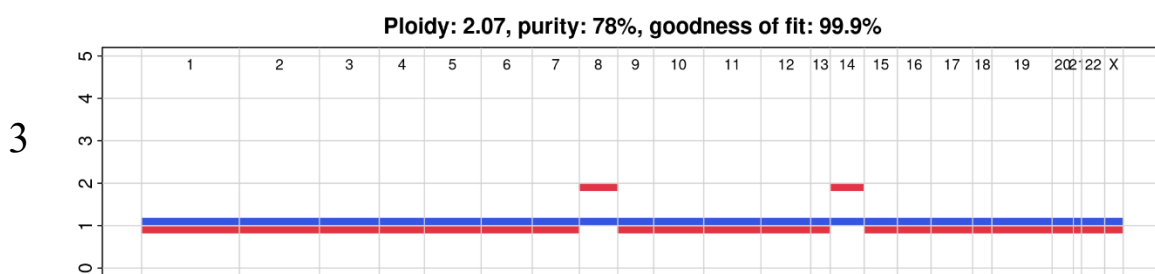
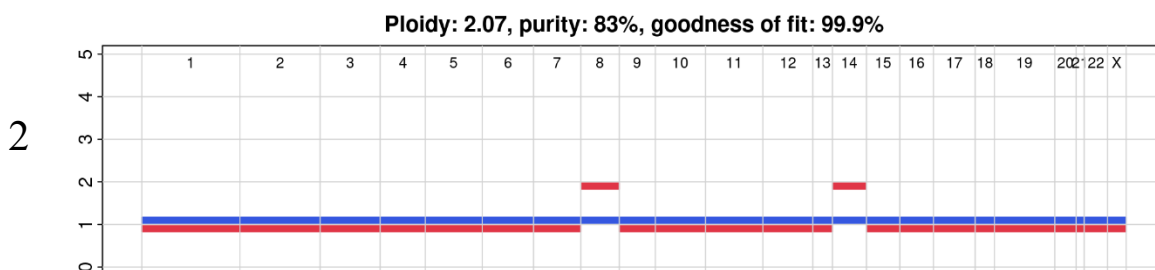
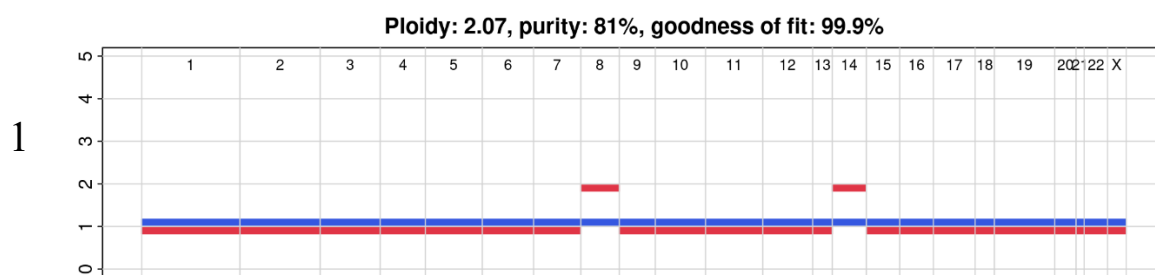
4



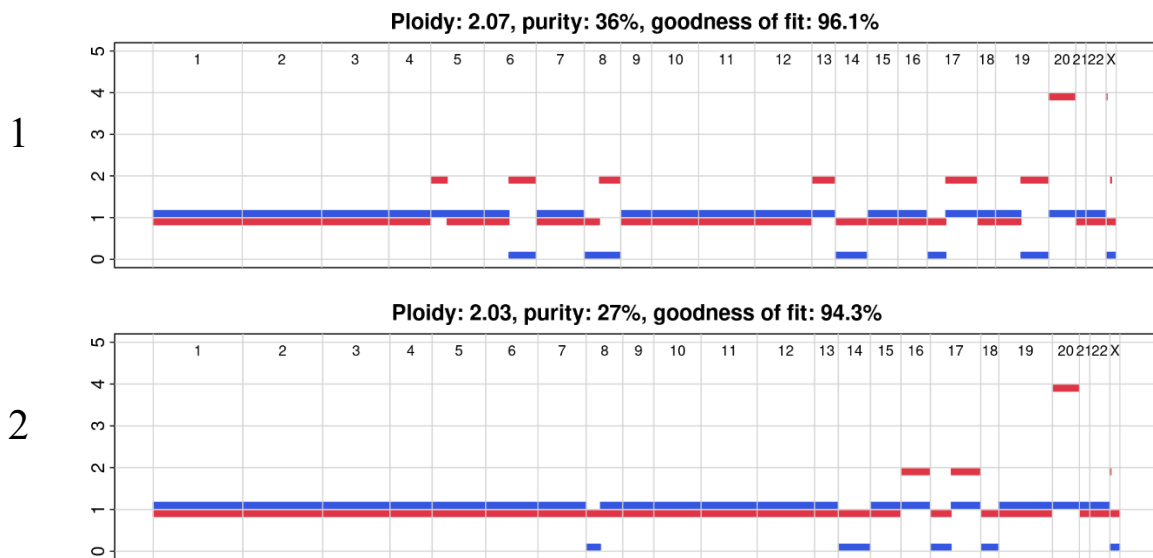
5



MD21386



MD21340



MD21353



4

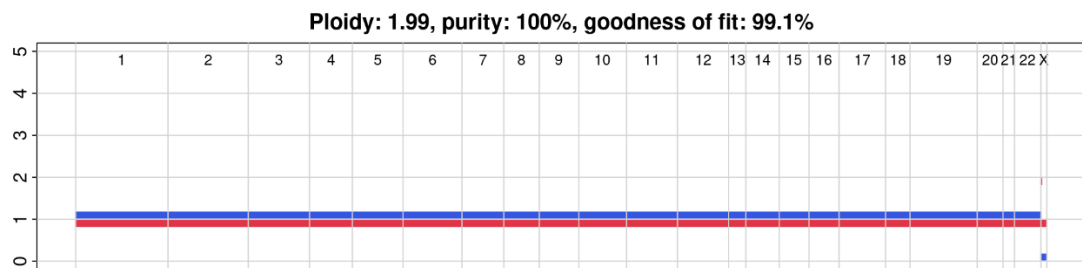


5



MD21354

1



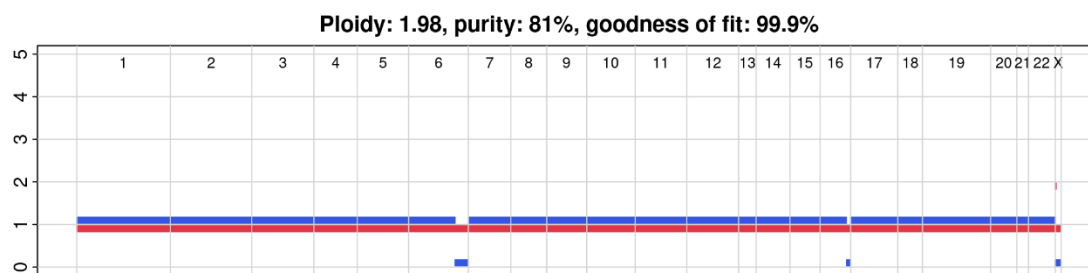
2



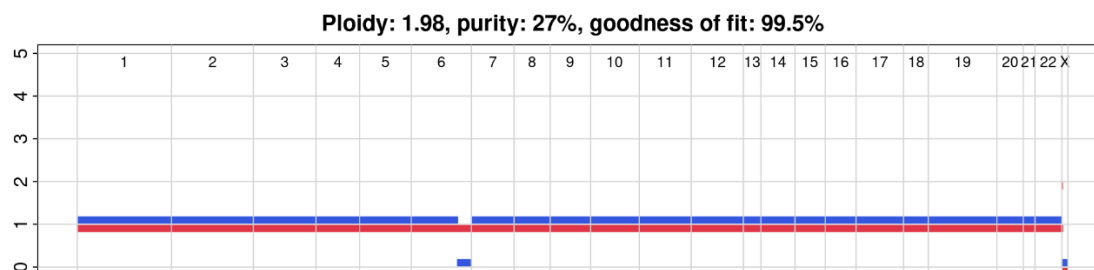
3



4

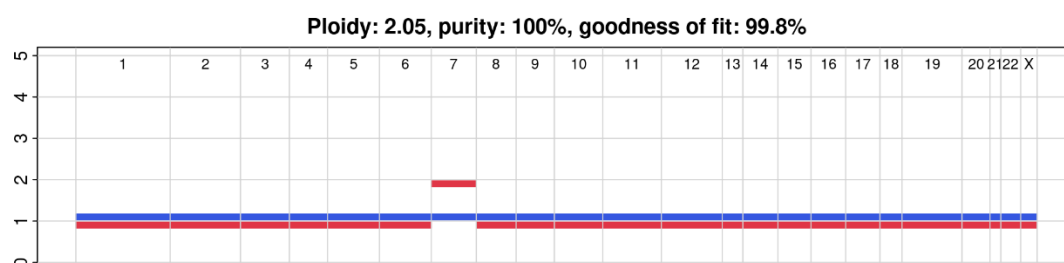


5

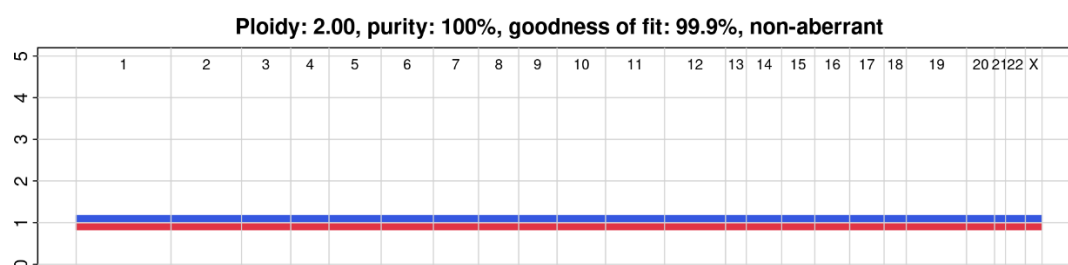


MD21320_1

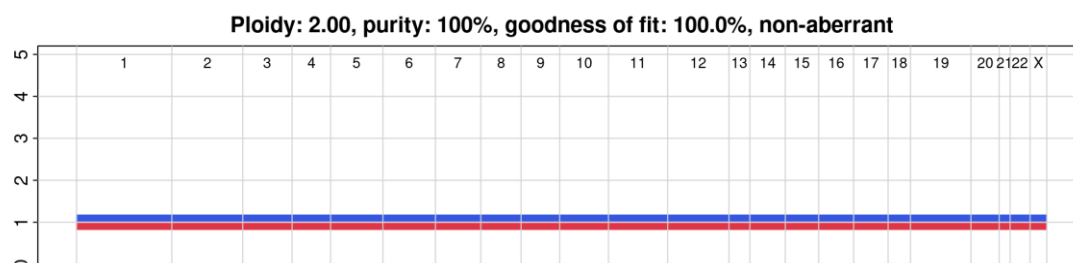
1



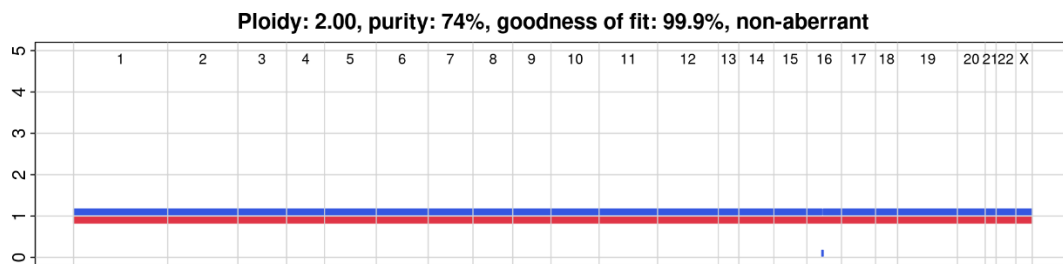
2



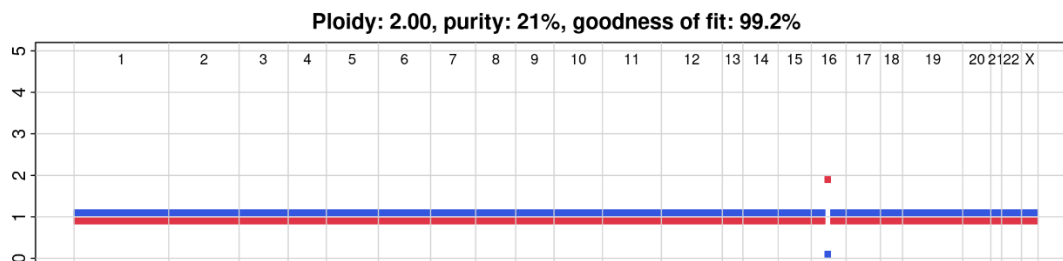
3



4

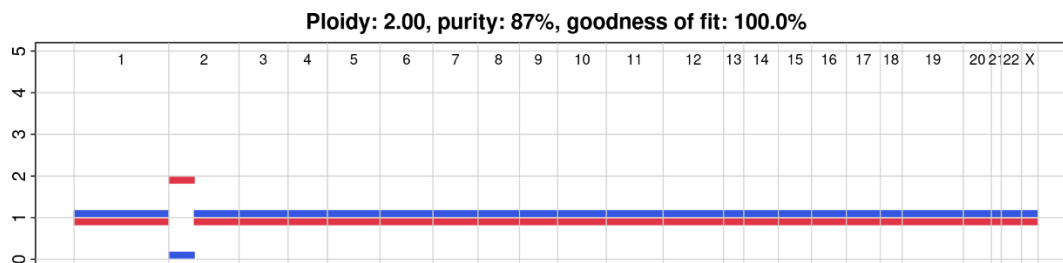


5

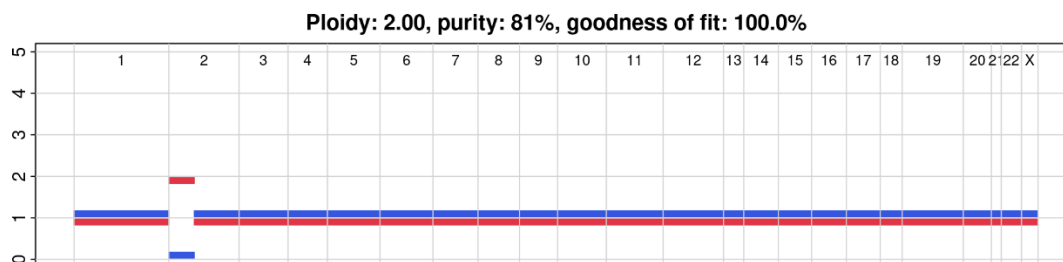


MD21320_2

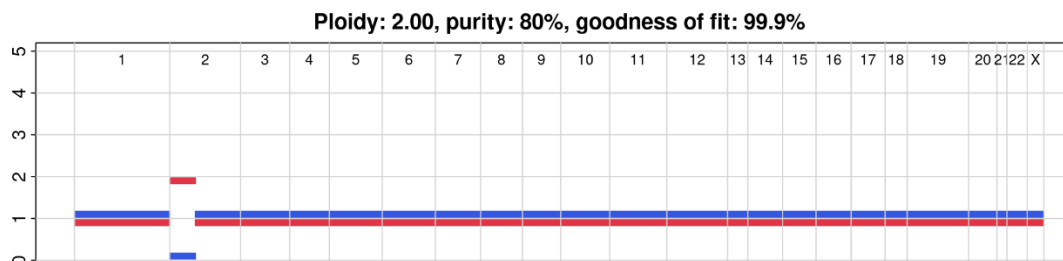
1



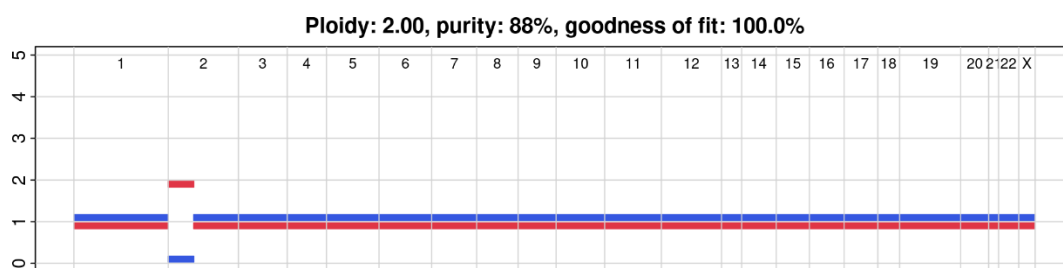
2



3



4



Chapter 3

A data-driven investigation of the just-right/mini-driver
model of colorectal cancer evolution in the Wnt signalling
pathway

3.1 Background

It is well established that CRC is strongly associated with aberrant Wnt signalling. Studies have shown that tumour cells with invasive and migratory properties, have stained positive for nuclear β -catenin. It is however, still in question whether there is a specific amount of nuclear β -catenin that allows for transition from CRA to CRC (Zhao *et al.*, 2022).

As stated in the previous chapter, the investigation of FAP/AFAP-related *APC* mutations resulted in the formulation of the just-right model of CRC evolution. This model suggests that there is a tendency for the most advantageous combination of germline and somatic *APC* mutations to be selected, resulting in the maintenance of at least one 20AAR across both alleles.

The just-right model was later expanded upon by Segditsas & Tomlinson (2006) who suggested that the model may extend across the Wnt pathway. For example, evidence suggests that both activating mutations in *CTNNB1* and *RNF43* are mutually exclusive with mutations in *APC*. This suggests that the simultaneous presence of these two major-driver mutations might lead to excessively high levels of Wnt signalling, at a level that is disadvantageous to the cell (Segditsas and Tomlinson, 2006; Giannakis *et al.*, 2014a). A similar relationship of mutual exclusivity has been observed between RSPO fusions and *APC* mutations. This exclusivity, combined with the ability of RSPO fusions to significantly alter Wnt signalling levels, supports the notion that RSPO fusions act as major-drivers of CRC, and a just-right level of Wnt signalling is surpassed if they occur alongside other Wnt-related major-driver mutations (Seshagiri *et al.*, 2012).

To further the just-right model, Castro-Giner, Ratcliffe and Tomlinson (2015) proposed the existence of mutations of small selective advantage, present at low frequencies in CRCs. They coined these mutations “mini-driver” mutations. Theoretically, a mini-driver mutation may have the ability to adjust the level of Wnt signalling to a just-right level, that may not be provided by the presence of a major Wnt pathway driver alone. In which case, a pattern of co-occurrence would be visible between known, major Wnt pathway driver mutations in *APC*, *CTNNB1* and *RNF43* and mini-driver mutations. Similarly, an accumulation of mini-driver mutations in the Wnt pathway may have a similar overall effect on the tumour as a major-driver mutation.

CRCs often harbour mutations in Wnt genes other than the key major-driver genes. For example, *AMER1* mutations are observed in approximately 10% of CRCs, but their contribution to CRC formation remains unclear. As described in section 1.3, mutations in *AMER1* can result in Wnt pathway dysregulation (Tanneberger, Pfister, Brauburger, *et al.*, 2011; Tanneberger, Pfister, Kriz, *et al.*, 2011). Although classed as a driver gene of CRC, *AMER1* may be a candidate mini-driver, capable of fine-tuning and optimising dysregulated Wnt signalling. Evidence for *AMER1* as a mini-driver of the Wnt pathway potentially would present itself as a pattern of co-occurrence with other major-driver mutations

in CRC. The occurrence of a mini-driver mutation alongside mutations in *APC*, may hypothetically provide a just-right amount of Wnt signalling to compensate for a sub-optimal number of 20AARs in *APC*.

Mutations in genes of low frequency and recognised as lacking in phenotypic consequence, may lead to reduced interest in further study. However, it is these genes that may be candidate mini-drivers and their presence in combination with major-driver mutations hypothetically assist in achieving a just-right level of pathway signalling. For example, it has been evidenced that genetic deletion of *BCL9* and *BCL9L* in the intestinal epithelium of mice does not cause any phenotypic consequences. However, this was not tested on the background of a major-driver mutation. *BCL9* and *BCL9L* could act as Wnt-related mini-drivers, and it is of interest to establish if mutations such as these, co-occur with other, Wnt pathway major or mini-driver mutations (Tanton *et al.*, 2022).

With the use of a large CRC dataset (100kGP), this chapter investigates the relationship between mutations in key Wnt pathway major-driver mutations and potential Wnt mini-driver mutations, analysing patterns of mutual exclusivity and co-occurrence. It also aims to further assess evidence for the just-right model by investigating the influence of mini-driver mutations on the number of 20AARs and therefore functionality of *APC*. This chapter investigates mutations in Wnt pathway-related candidate driver genes identified in a study by Cornish *et al.*, (2024), to which I am a contributing author. Cornish *et al.*, (2024) analysed 2023 CRC samples from the 100kGP database, identifying over 250 candidate driver genes, refining CRC subtypes, identifying new mutational processes, immune-escape mechanisms and structural variant hotspots. The discovery of novel Wnt pathway driver genes provides insight into potential mini-driver genes, enabling the exploration of relationships between mutations in candidate mini-driver genes and known major-drivers.

Notably, it has been established that MSS and MSI⁺ CRCs develop through the alteration of different pathways. Salahshor *et al.* (1999) based this conclusion on their study of mutations in *APC*, *KRAS*, *TP53* and *TGFβR2* in MSS and MSI⁺ CRCs. Through their analysis of 191, sporadic CRCs, they found that 86% of MSI⁺ tumours harboured a mutation in *TGFβR2*. This is in stark contrast to the frequency of *TGFβR2* mutations that they evidenced in MSS CRCs, at 0.6%. In addition, a strong negative correlation was found between MSI⁺ CRCs and mutations in *APC* and *TP53* (Salahshor *et al.*, 1999). This suggests such strong differences in mutational patterns, the just-right model has the potential to present differently in these two contexts. Mutations in MSS and MSI⁺ CRCs are therefore analysed separately in this chapter, with an aim to investigate evidence for the just-right model in the context of both MSS and MSI.

In a similar regard, previous evidence also suggests that differing numbers of 20AARs in *APC* are associated with different locations in the colon (Christie *et al.*, 2013). This chapter therefore aims to

further investigate whether there is evidence for the just-right model and whether it does indeed differ depending on tumour location.

By elucidating whether the just-right model/mini-driver model has a role in CRC evolution, therapeutic opportunities can be realised. For example, in 2023, Chang *et al.*, highlighted synthetic lethality triggered by overexpression of β -catenin, particularly in cell lines harbouring a truncation mutation in the MCR region of *APC*. In addition, they also found that the perturbation of other proteins in the destruction complex such as CK1 α and AXIN2, led to similar effects (Chang *et al.*, 2023). These results highlight the clinical importance of thoroughly understanding the role of the just-right model in different tissues and tumour subtypes (MSS/MSI).

3.2 Chapter aims

- Statistically assess whether there is evidence for the just-right model in the Wnt pathway utilising a large CRC dataset (100KGP).
- Assess whether previously evidenced mutually exclusive relationships between mutations in major-driver genes in the Wnt pathway can be replicated.
- Investigate whether mutations in potential mini-driver genes show patterns of mutual exclusivity or co-occurrence with major-driver genes.
- Investigate whether mutations in mini-driver genes co-occur with one another.
- Investigate whether any mutational relationships are restricted to either MSS or MSI-associated tumours meaning that the just-right model is achieved in different ways in these two cancer types.
- Investigate whether the optimal number of 20AARs differ between the proximal and distal colorectum.
- Investigate whether the presence of mini-driver mutations is associated with a change in the optimal number of *APC* 20AARs.
- Investigate whether the addition of a mini-driver mutation alters the optimal number of 20AARs in the proximal and distal colon/rectum
- Further explore any other correlations found within the data whether that be between genes or covariates.

3.3 Methods

3.3.1 Statistical assessment of mutational relationships between Wnt pathway genes

Version 8, hg38-aligned WGS data from the 100KGP was utilised which comprises 2023 CRC samples. These samples were also utilised in a study carried out by (Cornish *et al.*, 2024a), of which I am a contributing author. Cornish *et al.* (2024) performed SNV and indel calling using Strelka (v2.4.7) (Kim *et al.*, 2018). At the time of carrying out this analysis, Cornish *et al.* (2024) had identified 194 putative driver genes via utilisation of the IntOGen pipeline. An updated version of this manuscript identifies 250 putative driver genes. Utilising this list of 194 putative drivers, Cornish *et al.* (2024) produced a binary matrix in which CRC samples were annotated as “1” or “0” depending on whether each sample harboured a pathogenic mutation in one of the listed 194 candidate driver genes.

Non-synonymous mutations were considered pathogenic based on various criteria. Each mutation was annotated using variant effect prediction tool OncoKB (Chakravarty *et al.*, 2017). Each mutation was listed as pathogenic if annotated by OncoKB as “Oncogenic”, “Likely Oncogenic” or “Predicted Oncogenic”. Alternatively, for missense mutations in candidate driver genes classed as oncogenes and TSGs and those genes classed as ambiguous, if the mutation was seen in three or more tumours, the mutation was classed as recurrent and annotated pathogenic. For TSGs and those genes whose function is ambiguous, truncation mutations (splice acceptor, splice donor, frameshift, stop lost, stop gained or start lost) were all classed as pathogenic.

Out of these candidate driver genes, I filtered the results down to include only Wnt-related genes of interest: *AMER1*, *APC*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2*, *ZNRF3*. The binary matrix was utilised to identify which samples contained pathogenic mutations in each of the aforementioned genes. In addition, I utilised the structural variant calling carried out by Cornish *et al.* (2024) to identify structural variants in *RSPO* genes *RSPO1*, *RSPO2*, *RSPO3* and *RSPO4*. Structural variants were called by Delly (Rausch *et al.*, 2012), Lumpy (Layer *et al.*, 2014) and Manta (Chen, Schulz-Trieglaff, Shaw, Barnes, Schlesinger, Källberg, Anthony J Cox, *et al.*, 2016). I investigated all structural variants in these genes utilising the IGV browser to identify *RSPO* fusions. *RSPO* fusions were added to the binary matrix, noting “1” if the sample contained a fusion and “0” if not.

I assembled additional information regarding these samples/patients, such as MSI status, age at sampling, sex, mutation burden (somatic coding variants per Mb) and tumour location. I also investigated which samples harboured a pathogenic mutation in *POLE*. These mutations were limited to the exonuclease domain of *POLE* to ensure pathogenicity. 16 patients harboured a mutation in the exonuclease domain of *POLE*. These mutations included: *POLE*^{S459F}, *POLE*^{V411L} and *POLE*^{P286R}.

Samples were listed as “0” if the patient harbouring the tumour was female and “1” if male. Primary tumour location was also categorised into the following 3 categories: proximal colon; distal colon or rectum; colon or unknown tumour location. Proximal was listed as “0” in the matrix, distal colon and rectum “1” and colon or unknown tumour location “2”. There were 617 proximal samples, 607 distal samples and 310 samples in the colon/unknown location category.

I then filtered the samples to include only primary tumours, removing cases of both metastases and recurrence. I also removed any patients who had received systemic or radiotherapy prior to sampling. Following this, there were 1690 samples remaining. I then removed any samples with missing mutation burden information, resulting in a final sample number of 1534. 297 of these samples were MSI⁺ and 1237 MSS. Spearman’s rank correlations were then carried out via Python package SciPy (Virtanen *et al.*, 2020). The same sample set, covariates and filtering process were once again utilised to perform logistic regressions via Python package Statsmodels (GLM) (Seabold and Perktold, 2010). Logistic regressions were carried out via two distinct methods. Method A employed a pairwise logistic regression strategy, focussing on the direct relationship between mutations in two specific genes. Method B offers a more holistic approach by selecting one gene as the dependent variable (Y) and incorporating the rest of the Wnt pathway genes, along with key covariates as independent variables:

Prior to performing logistic regressions, a logarithmic transformation was performed on mutation burden (somatic coding variants per Mb) values. Post-Method A and Method B, further statistical analysis was then performed on the data in the form of feature selection with cross validation which utilised python package Scikit-learn (Pedregosa, F. Varoquaux, G. Gramfort *et al.*, 2011). In this case, Method B was utilised to perform the logistic regression before feature selection and cross validation were applied to each model.

3.3.2 Further just-right analysis in the context of *APC* 20AARs

A recent study by Brunet *et al.*, (currently unpublished*), to which I am a contributing author, assessed the number of 20AARs remaining in *APC* in those samples harbouring two pathogenic *APC* mutations. This study utilised the same dataset as described in 2.2.1, 2023 CRC samples from the 100KGP database. Samples were again limited to treatment naïve, primary cancers however filtering criteria expanded only those samples with no pathogenic mutations (exonuclease domain) in *POLE* (n=1477). *APC* mutations were more stringently filtered than in the Cornish *et al.* (2024) study. Only stop gained and frameshift truncating mutations were considered pathogenic. Truncation mutations occurring downstream of the SAMP repeat (AAR position 1569) were excluded. Any samples with only one mutation and no copy number change were also excluded (n=140).

The position of each stop gained and frameshift mutation was placed into one of four categories depending on the position of the mutation in regards to *APC*’s amino acids. The number of 20AARs

retained due to presence of a truncation mutation at this amino acid position, was then noted (**Table 3.1**). The influence of CNAs on 20AAR number was also taken into account. WGD was presumed to occur after two initial *APC* alterations. If copy loss acted as a second hit to *APC*, the number of 20AARs was calculated as the remaining number of 20AARs following the initial truncation mutation on one allele. If CN-LOH occurred, or an amplification of the major allele occurred with loss of the second allele, the number of 20AARs as a result of the initial truncation mutation was multiplied by the number of major alleles (Brunet *et al.*, *).

Table 3.1 The distribution of pathogenic mutations in *APC* and the corresponding number of 20AAR repeats retained. Adapted from Brunet *et al.* (*).

Region	Amino acid position	20AARs retained
R ₀	0-1256	0
R ₁	1256-1370	1
R ₂	1370-1486	2
R ₃	1486-1569	3

I applied all 20AAR information to the binary matrix described in 2.2.1 which contains additional sample information including the status of mutations within other Wnt pathway genes and covariates such as age, tumour site etc. From this matrix I removed samples with missing 20AAR information listed as having a pathogenic *APC* mutation by Cornish *et al.* (2024) (n=27). These samples are likely those that harboured recurrent missense or splice-site mutations considered pathogenic by Cross *et al.* (2022), but not by Brunet *et al.* (*). I also removed any samples included in the analysis by Brunet *et al.* (*) that did not have mutation burden data available. The final number of samples was 1109.

Following data preparation, I carried out a Shapiro-Wilk test for normality to determine the distribution of the 20AAR data and mutational data in the binary matrix, using Python package Scipy (Virtanen *et al.*, 2020). I then conducted a Mann-Whitney U test to statistically investigate the relationship between mutations in Wnt pathway genes and the number of 20AARs in *APC*. This was also carried out using Scipy (Virtanen *et al.*, 2020). This methodology was repeated in both MSS and MSI sample subsets.

3.3.3 Proximal, distal/rectal analysis

A comparison was made between the number of 20AARs in the proximal and distal colon/rectum. Any samples without exact tumour site data (unknown/colon) were excluded from this analysis. A total of 361 proximal samples and 507 distal samples were utilised in the analysis. The

Python package Scipy was utilised to conduct both Shapiro-Wilk and Mann-Whitney U tests (Virtanen *et al.*, 2020).

3.3.4 Lollipop plots

As stated in Section 2.1.1 mutations in *AMERI* and *TCF7L2* were annotated by Cornish *et al* (2024). Pathogenic mutations and their protein position were plotted in the form of a lollipop plot using Python package Seaborn (Waskom *et al.*, 2017). Lollipop plots were also created for both MSS and MSI samples. In addition, lollipop plots were created for those mutations present in the context of an *APC* mutation.

3.3.5 FAP

Two FAP patients undergoing surgery at the Western General Hospital were consented and enrolled into the SOCCS3 (Scottish Colorectal Cancer Genetic Susceptibility Study 3) study led by Professor Malcolm Dunlop, by research nurse Donna Markie. Additional information for these patients such as age at sampling, sex, follow up information, location of polyp, foci of invasion and dysplasia was gathered by Adam Gerrard. Malcolm Dunlop took biopsies from multiple polyps and tumours from the two patients, ranging from 30 mm to >2 cm. DNA and RNA extraction from these samples were carried out by Marion Walker (Dunlop group). Tissue samples were homogenised in Trizol using the Qiagen Tissue Lyser and RNA extracted using the Invitrogen RiboPure kit (AM1944). DNA was extracted using the Qiagen DNeasy Blood and Tissue kit (69504). Sequencing of normal blood samples was carried out via WGS. Sequencing of tumour tissue was carried out via WES. Sequencing data was then gifted to us by the Dunlop group.

Aligning and variant calling of the resulting WGS/WES data was carried out by Steve Thorn (Tomlinson group) using the bcbio-nextgen 1.2.4 pipeline. I then subset the BAM files utilising the hg38 Twist Illumina Exome 2.0 Plus Panel BED file to account for the differences in regions sequenced by WGS and WES. Following this, I ran the FACETS copy number caller with the assistance of scripts by Alison Meynert (Shen and Seshan, 2016).

Following this, I investigated the number of 20AARs in *APC* by assessment of the position of the germline and somatic *APC* mutation, combined with copy number status of *APC*. This was carried out via the same methodology described in section 2.2.2. I then investigated the presence of mutations in the Wnt pathway including *AMERI*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2*, *ZNRF3*. Mutations in these genes were only included if they were stop gained, frameshift, start lost, stop lost, splice site or missense mutations. Missense mutations were analysed with variant effect predictor tools SIFT (Ng and Henikoff, 2003) and Polyphen (Adzhubei, Jordan and Sunyaev, 2013) and only included in the analysis if listed as deleterious or probably damaging. No

mutations were found in *RNF43* or *CTNNB1*. Mutations in these genes are filtered in a more specific way in chapter 2.

3.3.6 Comparison of MSI/MSS and age in the context of *APC* and *RNF43* mutations

An investigation was carried out to assess the age of patients in MSS and MSI subsets. This was again assessed in the presence or absence of mutations in *APC* and *RNF43*. Statistical tests were performed using SciPy and distributions were plotted using Seaborn.

3.4 Results

To assess the applicability of the just-right model in CRC evolution across the entirety of the Wnt pathway, I conducted a statistical evaluation of relationships between pathogenic mutations in Wnt pathway genes. This examination aimed to illustrate patterns of co-occurrence and mutual exclusivity amongst pathogenic Wnt pathway mutations to further investigate the just-right model of CRC evolution. This study leverages the 100KGP dataset comprising 1534 treatment-naïve, primary tumour samples. Of these samples, 1237 were MSS and 237 were MSI⁺. Information on the presence of pathogenic mutations in candidate driver genes was collected by Cornish *et al.* (2024). Information regarding the pathogenic mutations in Wnt-related, candidate driver genes was collected from the aforementioned study. These genes included: *AMER1*, *APC*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2* and *ZNRF3*. In addition, samples were examined for fusions in *RSPO1/2/3/4*. Nine presumed pathogenic *RSPO* fusions were present in these 1534 samples. Less than 5 samples of each fusion were identified. The type of fusions present are shown in **Table 2.2**.

Table 3.2 RSPO2/3 fusions in treatment naïve, primary CRC tumour samples.

RSPO Fusion
RSPO3-PTPRK
RSPO2-EIF3E
RSPO3-ADGRB3
RSPO2-DECR1
RSPO2-TGS1

Results of all statistical tests analysing patterns of co-occurrence and mutual exclusivity between pathogenic mutations in Wnt pathway genes are firstly presented as heatmaps in Section 3.4.1.1 before being further described and analysed below.

3.4.1.1 Wnt pathway mutations in primary, treatment naïve CRCs.

Firstly, a Spearman's rank correlation was utilised to investigate basic associations between pathogenic mutations in primary, treatment naïve CRC samples. Presence of a pathogenic *POLE* mutation, sex, age at sampling, tumour site and mutation burden were also tested for any patterns of co-occurrence or mutual exclusivity with the presence of pathogenic Wnt pathway mutations (**Fig. 3.1**).

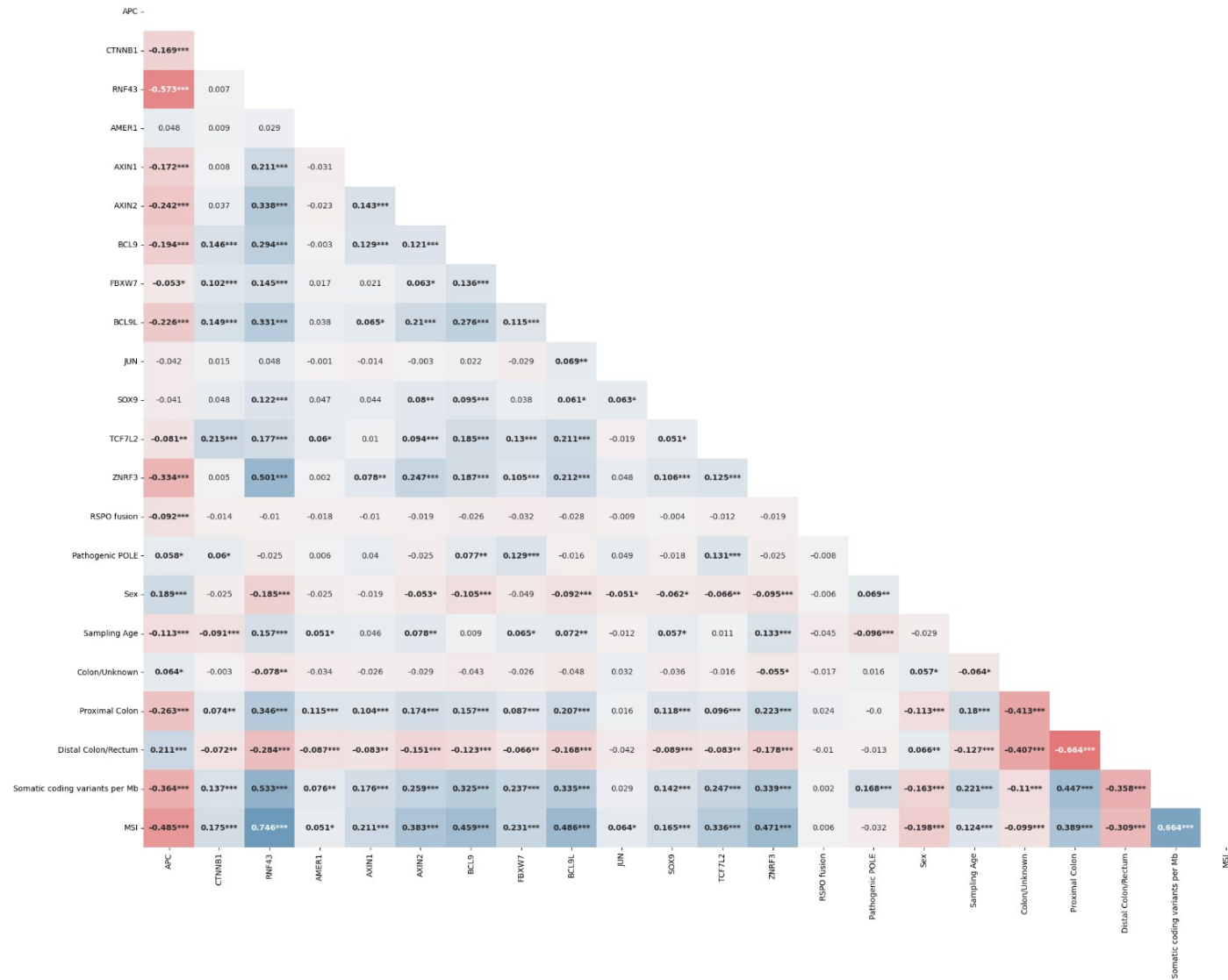


Figure 3.1 A pairwise comparison of pathogenic mutations in Wnt pathway genes, with additional covariates, in treatment naïve, primary CRCs. A Spearman's rank correlation was performed to explore patterns of co-occurrence and mutual exclusivity between pathogenic mutations in Wnt pathway genes *AMER1*, *APC*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2* and *ZNRF3*. This was performed in treatment naïve, primary, CRC samples. Relationships between pathogenic mutations and various covariates are also explored. These include the presence of a pathogenic *POLE* mutations (exonuclease domain), sex (male), age at sampling, tumour site and mutation burden (somatic coding variants per Mb). Red indicates mutual exclusivity, blue represents co-occurrence. For each relationship, a Spearman's rank statistic is displayed as well as asterisks for significance.

As seen in Figure 3.1, the presence of pathogenic mutations in most genes is associated with MSI. The exception to this rule are mutations in *APC* which show a negative correlation with MSI ($r = -0.485$, $p = < 0.001$). This negative relationship between mutations in *APC* and MSI is widely reported in previous literature (Fang *et al.*, 2002). Reassuringly, MSI is also showing a strongly positive association with mutation burden ($r = 0.664$, $p = < 0.001$). Interestingly, MSI⁺ samples show a negative correlation with the covariate listed as sex which corresponds to males ($r = -0.198$, $p = < 0.001$). Studies suggest that MSI is related to right-sided or proximal tumours which are more frequent in females (Choi *et al.*, 2023). This result therefore supports existing evidence for the relationship between sex and MSI. This study also supports the positive association between MSI and proximal tumours ($r = 0.389$, $p = < 0.001$).

Spearman's rank tests were also performed on MSS, treatment naïve, primary CRC samples, utilising the same covariates as previously described (**Fig 3.2**). Heatmaps for all tests are presented first, followed by a description of the results.

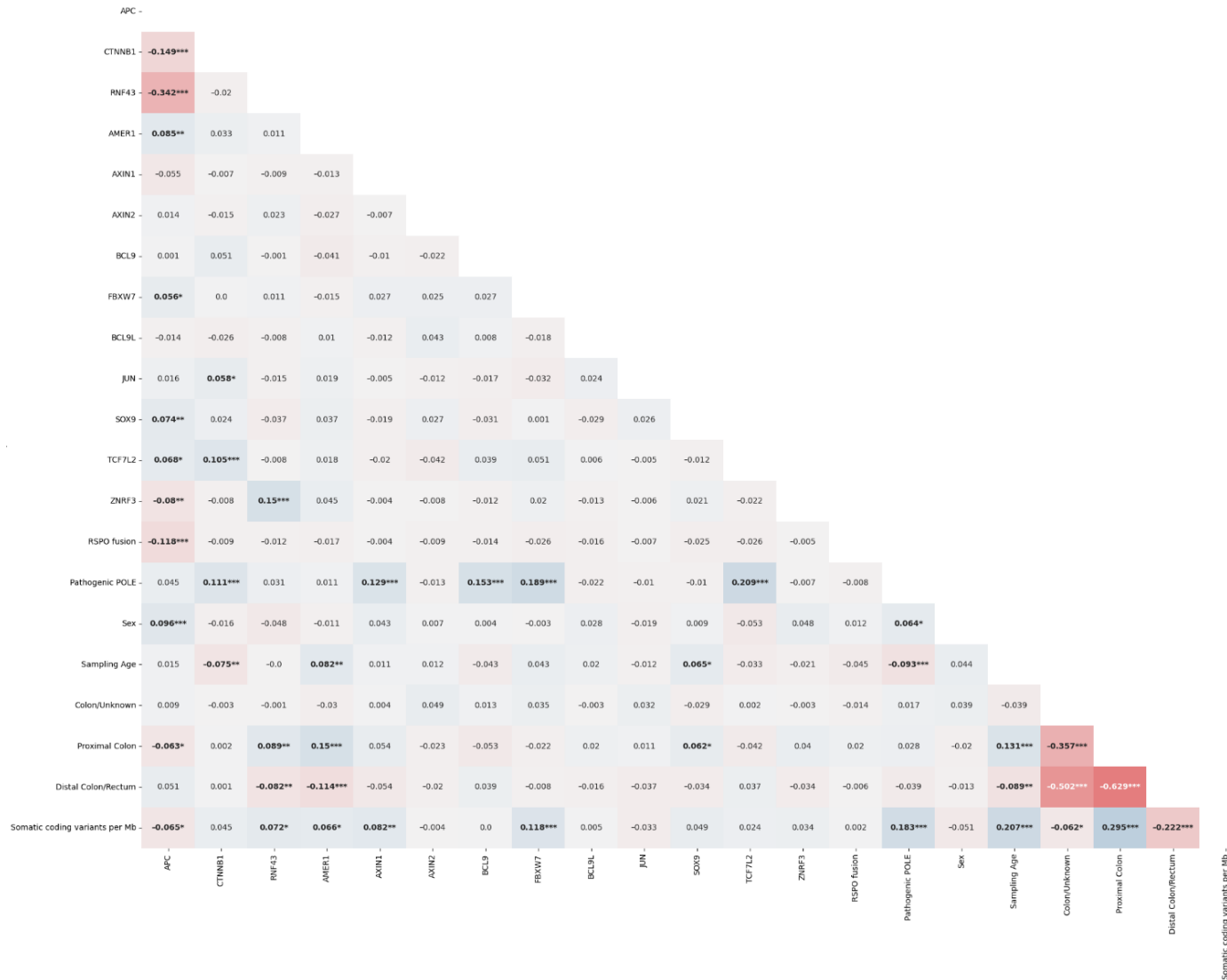


Figure 3.2 A pairwise comparison of pathogenic mutations in Wnt pathway genes with additional covariates in treatment naïve, MSS, primary CRCs. A Spearman's rank correlation was performed to explore patterns of co-occurrence and mutual exclusivity between pathogenic mutations in Wnt pathway genes *AMER1*, *APC*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2* and *ZNRF3*. This was performed in MSS, treatment naïve, primary, CRC samples. Relationships between pathogenic mutations and various covariates are also explored. These include the presence of a pathogenic *POLE* mutations (exonuclease domain), sex (male), age at sampling, tumour site and mutation burden (somatic coding variants per Mb). Red indicates a negative correlation/mutual exclusivity, blue represents a positive correlation/co-occurrence. For each relationship, a Spearman's rank statistic is displayed as well as asterisks for significance.

The same test was then repeated utilising MSI⁺, treatment naïve, primary CRC samples (**Fig 3.3**).

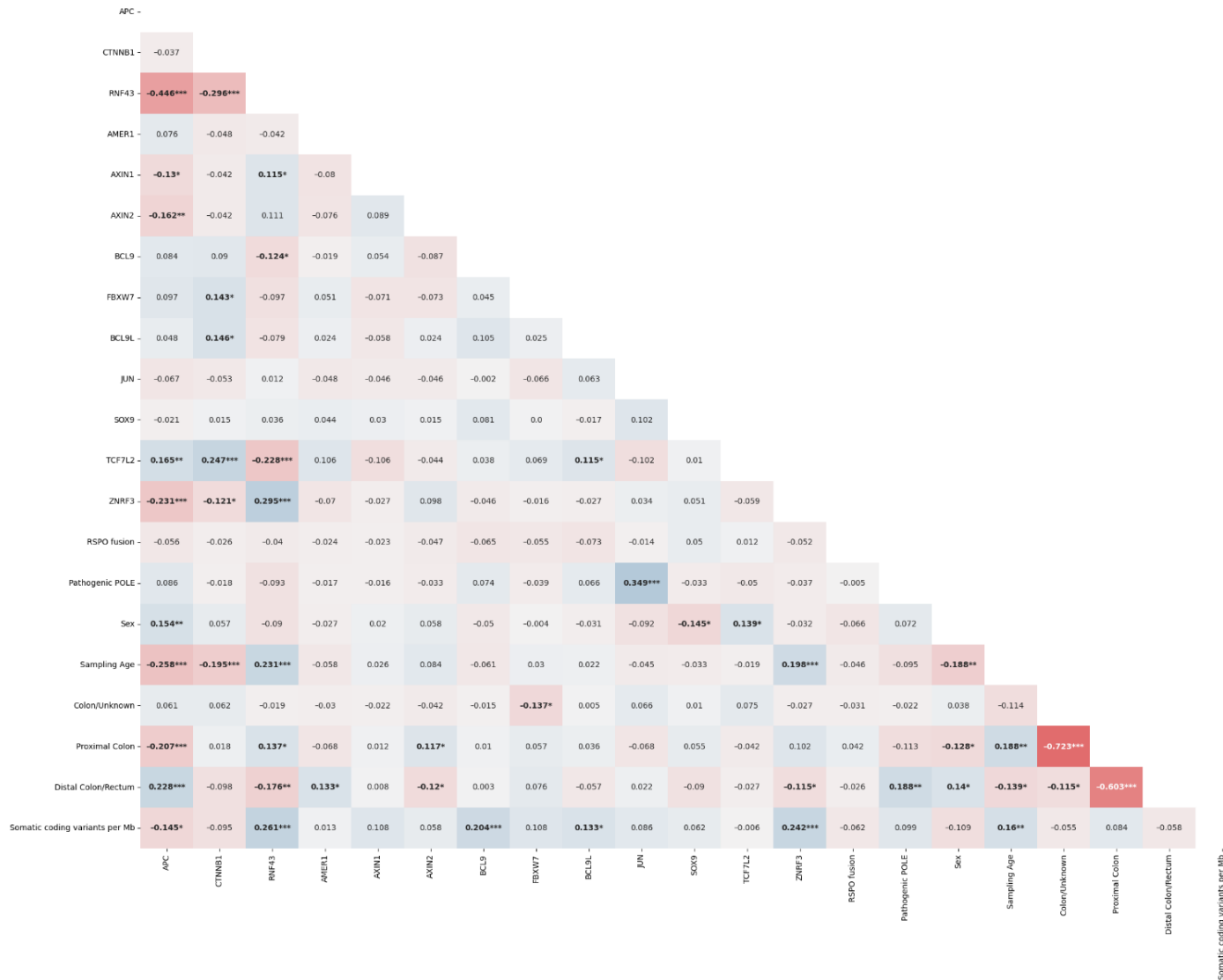


Figure 3.3 A pairwise comparison of pathogenic mutations in Wnt pathway genes with additional covariates in treatment naïve, MSI⁺, primary CRCs. A Spearman's rank correlation was performed to explore patterns of co-occurrence and mutual exclusivity between pathogenic mutations in Wnt pathway genes *AMER1*, *APC*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2* and *ZNRF3*. This was performed in MSI⁺, treatment naïve, primary, CRC samples. Relationships between pathogenic mutations and various covariates are also explored. These include the presence of a pathogenic *POLE* mutations (exonuclease domain), sex (male), age at sampling, tumour site and mutation burden (somatic coding variants per Mb). Red indicates a negative correlation/mutual exclusivity, blue represents a positive correlation/co-occurrence. For each relationship, a Spearman's rank statistic is displayed as well as asterisks for significance.

Following the Spearman's rank analyses, pairwise logistic regressions were then utilised on the dataset to provide a comprehensive framework for the analysis of multiple variables, adjusting for the effects of confounding variables. Mutations in a singular gene, plus multiple covariates were utilised as independent variables (X). Covariates included: presence of a pathogenic *POLE* mutation (exonuclease domain), sex, age at sampling, tumour site and mutation burden (somatic coding variants per Mb). This method is referred to as method A in the section 3.2.1. From this point onwards, in this section, MSS and MSI⁺ CRCs are analysed separately.

Firstly, pairwise logistic regressions were carried out utilising MSS, treatment naïve, primary CRCs (**Fig 3.4**), followed by MSI⁺, treatment naïve, primary CRCs (**Fig 3.5**).

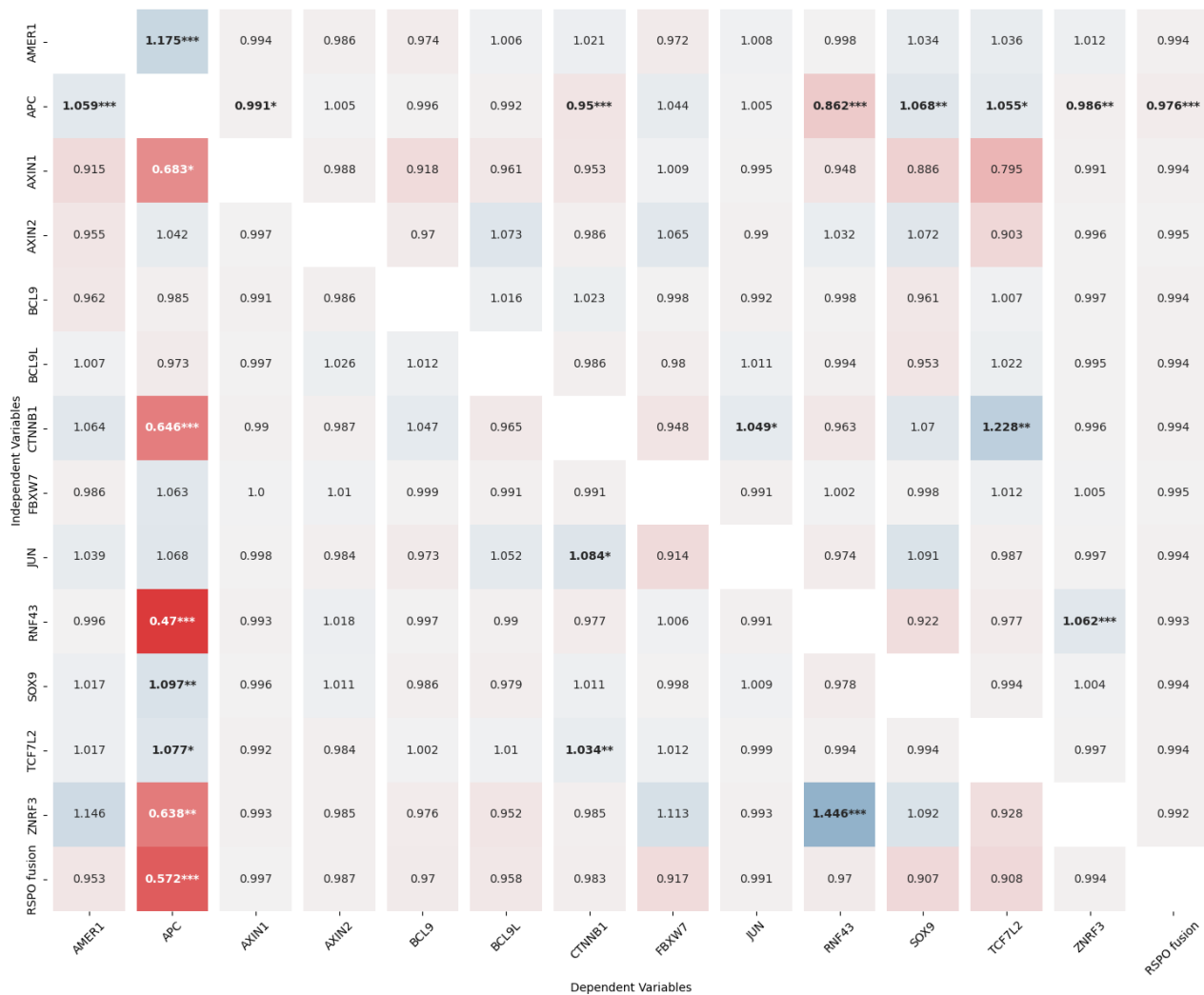


Figure 3.4 An assessment of associations between mutations in Wnt pathway genes and covariates utilising multiple, pairwise logistic regression models (Method A) in MSS, primary, treatment naïve, CRC samples. Multiple, pairwise logistic regression models were utilised to explore patterns of co-occurrence and mutual exclusivity among Wnt pathway genes *AMER1*, *APC*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2* and *ZNRF3*. This was performed in MSS, treatment naïve, primary CRC samples. Each regression was performed in a pairwise fashion with the presence of pathogenic mutations in a single gene acting as the independent variable and another as a dependent variable. Multiple covariates were included as predictor variables to account for their effect on these mutational relationships. Covariates include: the presence of a pathogenic *POLE* mutation (exonuclease domain), sex (male), age at sampling, tumour site and mutation burden (somatic coding variants per Mb). Red represents mutual exclusivity, blue represents co-occurrence. Odds ratios are shown in each square of the heatmap along with asterisks for significance.

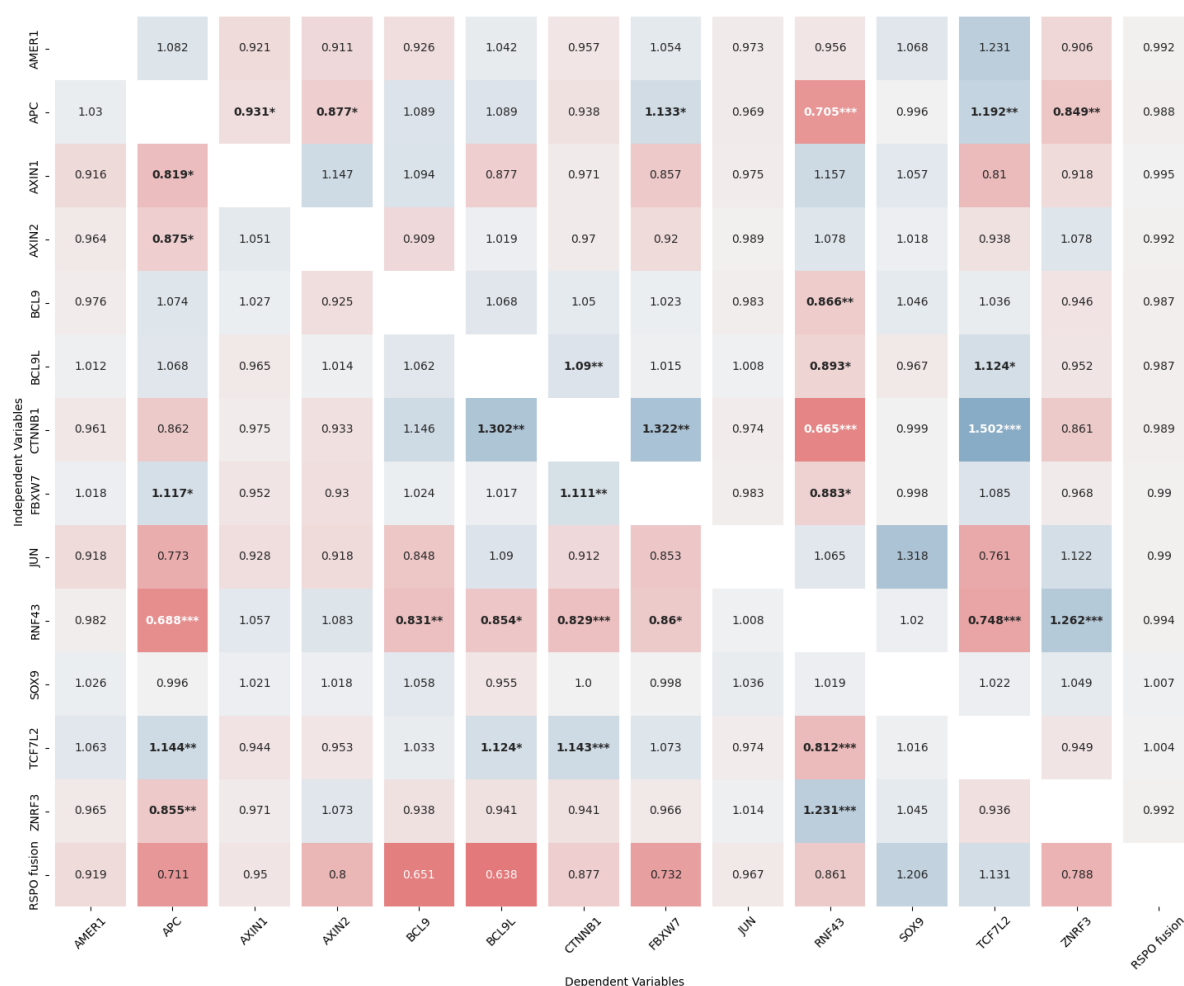


Figure 3.5 An assessment of associations between mutations in Wnt pathway genes and covariates utilising multiple, pairwise logistic regression models (Method A) in MSI⁺, primary, treatment naïve, CRC samples. Multiple, pairwise logistic regression models were utilised to explore patterns of co-occurrence and mutual exclusivity among Wnt pathway genes *AMER1*, *APC*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2* and *ZNRF3*. This was performed in MSI⁺, treatment naïve, primary CRC samples. Each regression was performed in a pairwise fashion with the presence of pathogenic mutations in a single gene acting as the independent variable and another as a dependent variable. Multiple covariates were included as predictor variables to account for their effect on these mutational relationships. Covariates include: the presence of a pathogenic *POLE* mutation (exonuclease domain), sex (male), age at sampling, tumour site and mutation burden (somatic coding variants per Mb). Red represents mutual exclusivity, blue represents co-occurrence. Odds ratios are shown in each square of the heatmap along with asterisks for significance.

Logistic regression analyses were then repeated via Method B to explore a broader landscape of gene interactions within the Wnt signalling pathway. By selecting pathogenic mutations in one gene as the dependent variable (Y) and incorporating all other pathogenic mutations in other Wnt pathway genes, along with the key covariates as independent variables (X), this method allows for a more expansive view of the interactions between mutations and covariates in the Wnt pathway. This allows for the capture of a web of interactions and a more informative view of the impact of multiple genes and covariates on the mutation status of a focal gene. Initially, this test was performed on MSS,

treatment naïve, primary CRCs (**Fig. 3.6**), before being repeated on MSI⁺, treatment naïve, primary CRCs (**Fig. 3.7**).

AMER1 -		1.178***	0.996	0.984	0.973	1.011	1.03	0.964	1.005	1.021	1.019	1.018	1.014	0.999
APC -	1.072***		0.988*	1.009	1.0	0.983	0.933***	1.05	1.007	0.856***	1.064*	1.069*	0.992	0.97***
AXIN1 -	0.947	0.663*		0.988	0.916	0.954	0.938	1.027	1.0	0.892	0.904	0.825	0.992	0.978
AXIN2 -	0.952	1.065	0.997		0.968	1.079	0.998	1.062	0.989	1.04	1.075	0.9	0.994	0.997
BCL9 -	0.96	0.999	0.991	0.985		1.017	1.023	0.999	0.991	0.998	0.962	0.998	0.998	0.994
BCL9L -	1.012	0.956	0.996	1.028	1.013		0.982	0.98	1.013	0.987	0.954	1.029	0.996	0.993
CTNNB1 -	1.095	0.602***	0.986	0.998	1.05	0.951		0.974	1.052*	0.898**	1.095	1.266***	0.994	0.98
FBXW7 -	0.982	1.059	1.001	1.01	1.0	0.991	0.996		0.992	1.008	0.994	1.011	1.005	0.996
JUN -	1.024	1.082	1.0	0.983	0.97	1.058	1.085*	0.917		0.995	1.078	0.963	0.999	0.997
RNF43 -	1.044	0.471***	0.983	1.026	0.998	0.976	0.931**	1.033	0.998		0.959	1.036	1.055***	0.971*
SOX9 -	1.009	1.075*	0.996	1.012	0.987	0.979	1.015	0.994	1.007	0.99		0.987	1.006	0.996
TCF7L2 -	1.009	1.08*	0.993	0.984	0.999	1.013	1.038***	1.011	0.996	1.008	0.987		0.998	0.997
ZNRF3 -	1.166	0.806	0.993	0.979	0.981	0.956	0.977	1.13	0.997	1.341***	1.135	0.943		0.989
RSPO fusion -	0.993	0.568***	0.987	0.992	0.967	0.948	0.95	0.942	0.996	0.891*	0.938	0.946	0.992	
Pathogenic POLE -	0.987	1.361*	1.055*	0.985	1.256**	0.954	1.045	1.336*	1.048	0.99	0.885	1.9***	0.982	1.003
Distal colon/Rectum -	0.873**	2.354***	1.002	1.001	1.079*	1.022	1.092***	0.896	1.017	1.162***	0.9	1.114	1.008	1.047**
Proximal Colon -	0.935	2.291***	1.008	0.999	1.058	1.032	1.088**	0.873*	1.024	1.188***	0.927	1.081	1.011	1.051**
Colon/Unknown -	0.887**	2.333***	1.005	1.015	1.074	1.023	1.089***	0.911	1.027	1.173***	0.894	1.102	1.009	1.046**
Somatic coding variants per Mb -	1.011	0.969	1.006	1.002	1.006	1.003	1.018	1.07**	0.986	1.022	1.03	0.987	1.002	1.001
Sex -	0.99	1.063**	1.005	1.0	0.999	1.014	1.002	0.987	0.995	0.995	1.002	0.956*	1.008*	1.004
Sampling age -	1.136*	1.007	1.0	1.0	0.941	1.028	0.942*	1.138	1.004	0.962	1.163*	0.987	0.988	0.978
	AMER1	APC	AXIN1	AXIN2	BCL9	BCL9L	CTNNB1	FBXW7	JUN	RNF43	SOX9	TCF7L2	ZNRF3	RSPO fusion

Dependent Variables

Figure 3.6 An assessment of associations between mutations in Wnt pathway genes and covariates utilising multiple, logistic regression models (Method B) in MSS, primary, treatment naïve, CRC samples. Multiple, logistic regression models were utilised to explore patterns of co-occurrence and mutual exclusivity among Wnt pathway genes *AMER1*, *APC*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2* and *ZNRF3*. This was performed in MSS, treatment naïve, primary CRC samples. Each regression was performed in a pairwise fashion with the presence of pathogenic mutations in a single gene acting as the independent variable and another as a dependent variable. Multiple covariates were included as predictor variables to account for their effect on these mutational relationships. Covariates include: the presence of a pathogenic *POLE* mutation (exonuclease domain), sex (male), age at sampling, tumour site and mutation burden (somatic coding variants per Mb). Red represents mutual exclusivity, blue represents co-occurrence. Odds ratios are shown in each square of the heatmap along with asterisks for significance.

AMER1 -		0.981	0.932	0.917	0.906	1.034	0.925	1.035	0.969	0.986	1.096	1.205	0.901	0.986
APC -	0.991		0.944	0.884	1.012	1.061	0.836***	1.108	0.961	0.721***	1.03	1.125	0.89	0.978
AXIN1 -	0.927	0.877		1.109	1.117	0.905	0.994	0.892	0.964	1.052	1.074	0.878	0.86	0.991
AXIN2 -	0.967	0.904	1.038		0.907	1.053	0.972	0.946	0.979	0.986	1.031	0.983	1.046	0.988
BCL9 -	0.969	1.008	1.033	0.923		1.047	1.014	0.99	0.981	0.905*	1.064	0.988	0.982	0.985
BCL9L -	1.01	1.037	0.973	1.04	1.043		1.062	0.973	1.015	0.965	0.963	1.063	0.979	0.989
CTNNB1 -	0.919	0.669***	0.994	0.925	1.046	1.246		1.308**	0.967	0.688***	1.033	1.412***	0.899	0.98
FBXW7 -	1.012	1.076	0.964	0.952	0.989	0.969	1.09**		0.987	0.963	1.006	0.999	1.008	0.991
JUN -	0.906	0.77	0.9	0.845	0.829	1.173	0.906	0.887		0.883	1.376	0.809	1.036	0.978
RNF43 -	0.993	0.689***	1.026	0.98	0.845*	0.937	0.827***	0.942	0.978		1.035	0.848*	1.184*	0.982
SOX9 -	1.036	1.025	1.026	1.031	1.078	0.951	1.012	1.007	1.041	1.025		1.033	1.051	1.01
TCF7L2 -	1.062	1.082	0.962	0.986	0.988	1.068	1.109***	0.999	0.978	0.907*	1.027		1.016	1.006
ZNRF3 -	0.961	0.912	0.948	1.044	0.979	0.973	0.963	1.008	1.004	1.124*	1.049	1.018		0.989
RSPO fusion -	0.866	0.617	0.912	0.732	0.621	0.676	0.818	0.752	0.929	0.704	1.292	1.225	0.742	
Pathogenic POLE -	0.857	1.201	0.96	1.035	1.302	1.29	0.959	0.624	2.583***	0.658	0.618	0.661	0.918	1.011
Distal colon/Rectum -	1.115	3.43***	0.922	1.062	0.682	0.577	1.564**	0.67	1.0	0.756	1.177	0.753	0.651	1.056
Proximal Colon -	0.989	3.025***	0.877	1.228	0.706	0.633	1.762***	0.595	0.999	0.85	1.315	0.819	0.687	1.066
Colon/Unknown -	0.975	3.242***	0.877	1.153	0.694	0.63	1.818***	0.499*	1.04	0.897	1.303	0.903	0.682	1.054
somatic coding variants per Mb -	1.052	1.003	1.049	1.036	1.333***	1.227**	0.987	1.181*	1.028	1.267***	1.033	1.11	1.11	1.007
Sex -	0.965	1.055	1.021	1.077	0.919	0.946	1.011	0.981	0.971	1.007	0.886*	1.136*	1.022	0.988
Sampling age -	0.869	0.476***	1.033	0.949	0.645	1.056	0.664***	1.254	0.946	1.13	0.752	1.226	1.211	0.946
	AMER1	APC	AXIN1	AXIN2	BCL9	BCL9L	CTNNB1	FBXW7	JUN	RNF43	SOX9	TCF7L2	ZNRF3	RSPO fusion

Dependent Variables

Figure 3.7 An assessment of associations between mutations in Wnt pathway genes and covariates utilising multiple, logistic regression models (Method B) in MSI⁺, primary, treatment naïve, CRC samples. Multiple, logistic regression models were utilised to explore patterns of co-occurrence and mutual exclusivity among Wnt pathway genes *AMER1*, *APC*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2* and *ZNRF3*. This was performed in MSI⁺, treatment naïve, primary CRC samples. Each regression was performed in a pairwise fashion with the presence of pathogenic mutations in a single gene acting as the independent variable and another as a dependent variable. Multiple covariates were included as predictor variables to account for their effect on these mutational relationships. Covariates include: the presence of a pathogenic *POLE* mutation (exonuclease domain), sex (male), age at sampling, tumour site and mutation burden (somatic coding variants per Mb). Red represents mutual exclusivity, blue represents co-occurrence. Odds ratios are shown in each square of the heatmap along with asterisks for significance.

Finally, feature selection was performed on the data within each logistic regression model to mitigate overfitting, incorporating only crucial variables relevant to the model. Cross-validation was additionally conducted to evaluate the stability and resilience of the selected features obtained via feature selection. Theoretically, this approach emphasises the most significant variables and outcomes in each logistic regression model. Firstly this approach was carried out utilising MSS, treatment naïve, primary CRCs (Fig 3.8), followed by MSI⁺, treatment naïve, primary CRCs (Fig 3.9).

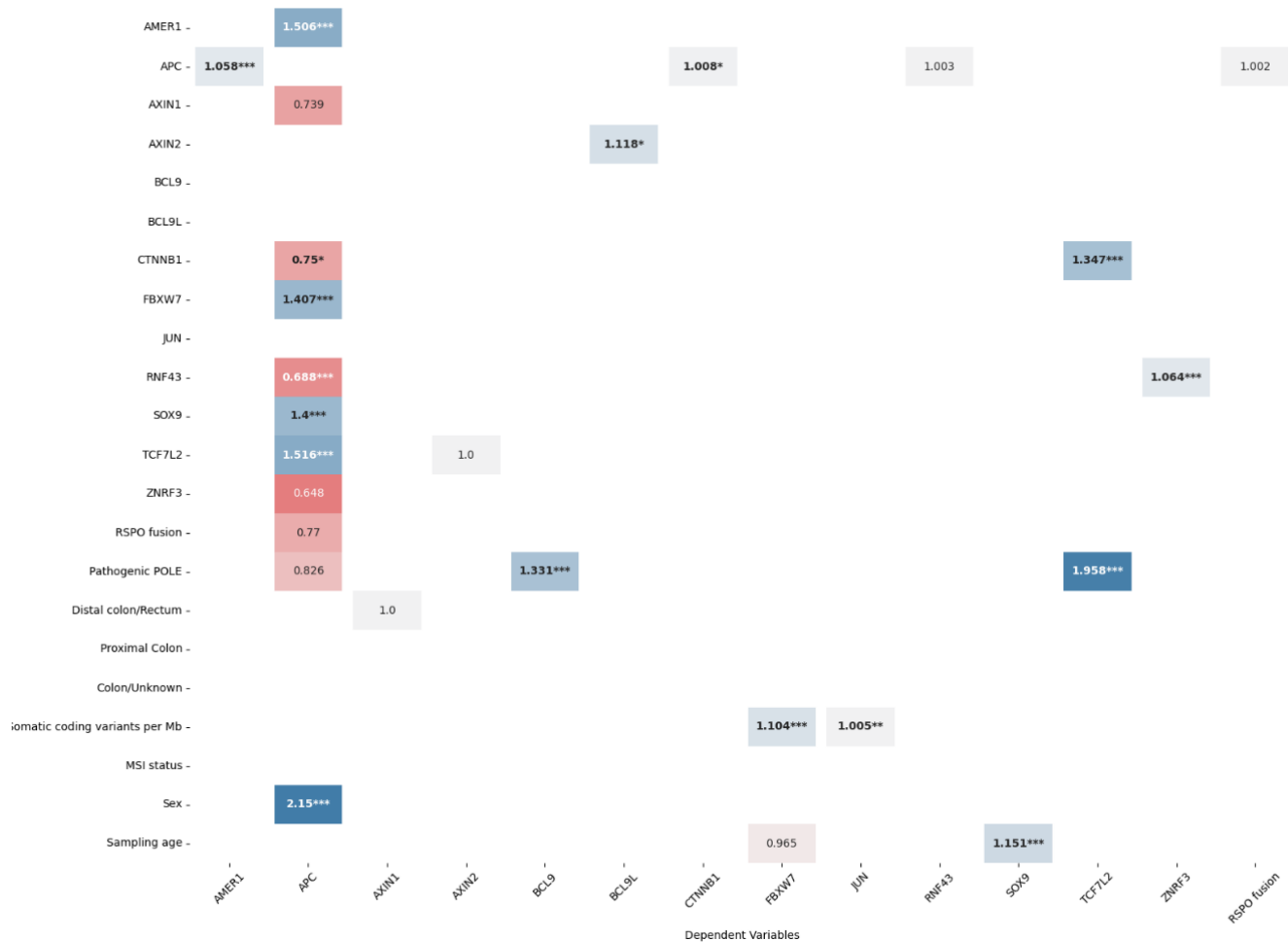


Figure 3.8 Multiple logistic regression models with feature selection and cross validation for the purpose of exploring relationships between pathogenic mutations in Wnt pathway genes and covariates, in MSS, treatment naïve, primary CRCs. Multiple logistic regression models were employed in conjunction with feature selection and cross validation to highlight significant features and relationships between pathogenic Wnt pathway mutations and various covariates in MSS, treatment naïve, primary CRC samples. Red indicates mutual exclusivity, blue indicates co-occurrence. Odds ratios for each examined relationship are displayed along with asterisks representing significance.

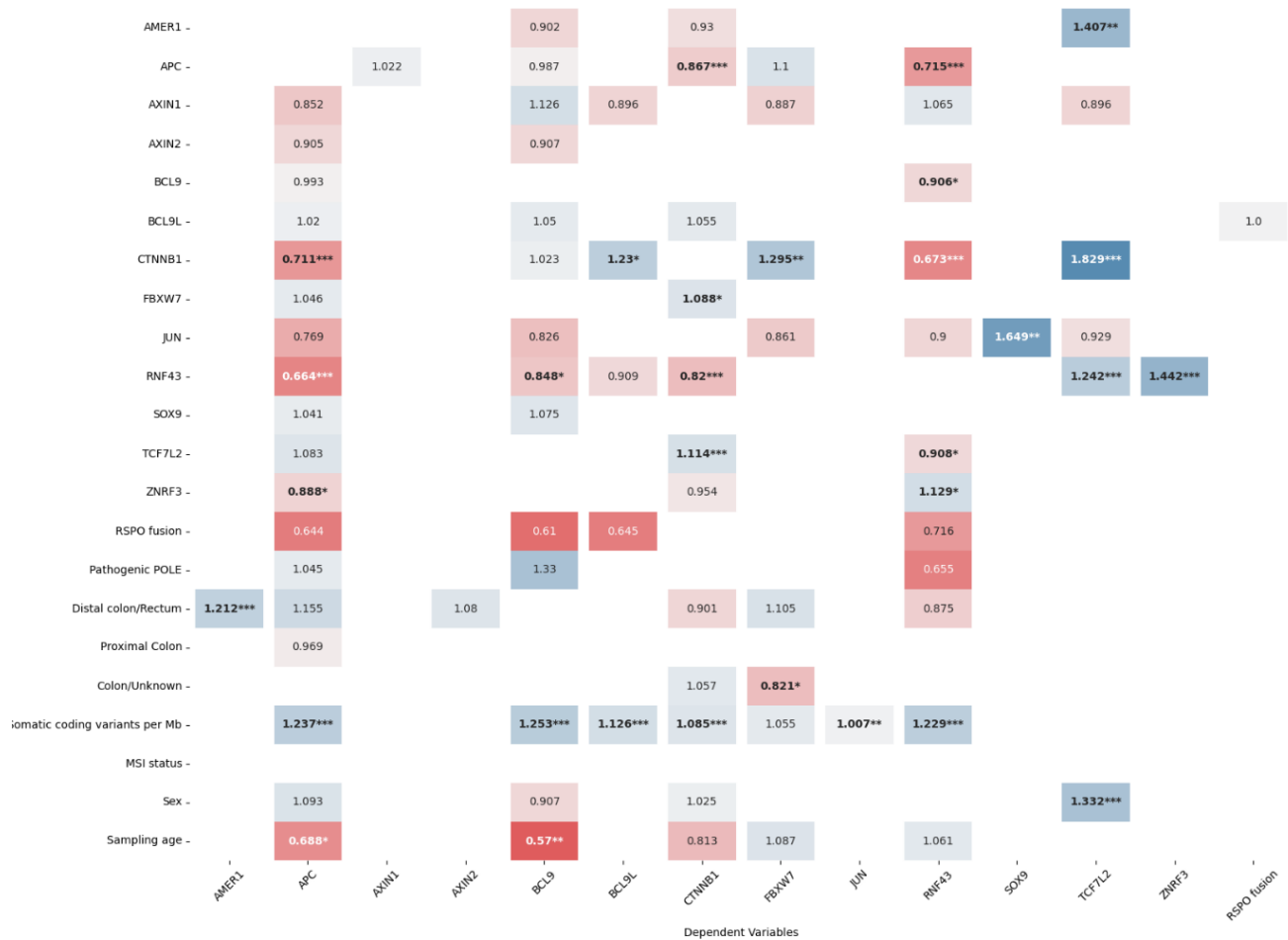


Figure 3.9 Multiple logistic regression models with feature selection and cross validation for the purpose of exploring relationships between pathogenic mutations in Wnt pathway genes and covariates, in MSI⁺, treatment naïve, primary CRCs. Multiple logistic regression models were employed in conjunction with feature selection and cross validation to highlight significant features and relationships between pathogenic Wnt pathway mutations and various covariates in MSI⁺, treatment naïve, primary CRC samples. Red indicates mutual exclusivity, blue indicates co-occurrence. Odds ratios for each examined relationship are displayed along with asterisks representing significance.

3.3.1.2 Wnt-pathway major-driver mutations and *APC*, in MSS CRCs

Via multiple statistical methodologies, this data confirms a mutually exclusive relationship between mutations in Wnt pathway, major-driver genes *RNF43*, *CTNNB1*, *RSPO* fusions and mutations in *APC*, in MSS CRCs.

Firstly, as seen in Figure 3.3, *CTNNB1* and *APC* are negatively correlated via Spearman's rank ($r = -0.149$, $p < 0.001$). This relationship is replicated via pairwise logistic regression (Method A) where *CTNNB1* mutations are used as predictor variables alongside additional covariates such as mutation burden, age, sex and tumour site, and *APC* mutations are utilised as the outcome variable (**Fig 3.4**) (OR = 0.646, [95% CI: 0.549, 0.759], $p < 0.001$). A negative correlation between mutations in *CTNNB1* and *APC* are also shown via logistic regression Method B which utilises mutations in all other Wnt genes as predictor variables, in addition to *CTNNB1* mutations, for the purpose of investigating the

relationship between *CTNNB1* and *APC* mutations as part of a wider context (**Fig. 3.6**) (OR = 0.602, [95% CI: 0.518, 0.699], $p < 0.001$). Logistic regression with feature selection and cross-validation also supports that a mutation in *CTNNB1* is an important negative predictor of the occurrence of *APC* mutations (OR = 0.75, [95% CI: 0.583, 0.966], $p = 0.026$) (**Fig. 3.8**).

A strong negative association between mutation in *RNF43* and *APC* was observed via Spearman's rank (**Fig. 3.2**) ($r = -0.342$, $p < 0.001$). This negative relationship was also observed via pairwise logistic regression, in which presence of an *RNF43* mutation was associated with a decreased likelihood of the occurrence of an *APC* mutation (OR = 0.47, [95% CI: 0.417, 0.529], $p < 0.001$) (**Fig. 3.4**). When carrying out a logistic regression via Method B, this negative relationship was once again evidenced (OR = 0.471, [95% CI: 0.419, 0.529], $p < 0.001$) (**Fig. 3.6**). Feature selection with cross validation once again highlighted the significance of *RNF43* mutations in predicting the presence of an *APC* mutation (OR = 0.688, [95% CI: 0.567, 0.836], $p < 0.001$) (**Fig. 3.8**). These findings support a robust negative association between mutations in *RNF43* and *APC*.

Again, a negative relationship was shown between RSPO fusions and mutations in *APC* via Spearman's rank ($r = -0.118$, $p < 0.001$) (**Fig. 3.2**). This finding was replicated through a pairwise logistic regression in which the presence of an RSPO fusion was associated with a decreased likelihood of a mutation in *APC* (OR = 0.572, [95% CI: 0.44, 0.745], $p < 0.001$) (**Fig. 3.4**). Again, this finding was replicated via logistic regression Method B (OR = 0.568, [95% CI: 0.446, 0.723], $p < 0.001$) (**Fig. 3.6**). However, the logistic regression with feature selection and cross-validation did not identify RSPO fusions as statistically significant determinant features in the occurrence of an *APC* mutation (**Fig. 3.8**).

3.3.1.3 Wnt-pathway major-driver mutations and *APC*, in MSI⁺ CRCs

Notably, in the context of MSI, a reliably negative relationship was once again evidenced between mutations in *APC* and *RNF43* via Spearman's rank ($r = -0.446$, $p < 0.001$) (**Fig. 3.3**). This negative relationship was replicated via pairwise logistic regression (Method A), in which *RNF43* mutations were associated with a decreased likelihood of the occurrence of an *APC* mutation (OR = 0.688, [95% CI: 0.616, 0.769], $p < 0.001$) (**Fig. 3.5**). A very similar result was found in the context of additional Wnt pathway mutations, via logistic regression Method B (OR = 0.689, [95% CI: 0.613, 0.775], $p < 0.001$) (**Fig. 3.7**). *RNF43* was also found to be a negative predictor of *APC* mutations via logistic regression with feature selection and cross-validation (OR = 0.664, [95% CI: 0.588, 0.749], $p < 0.001$) (**Fig. 3.9**).

Notably, *CTNNB1* mutations were not found to be significantly associated with *APC* mutations via Spearman's rank or via pairwise logistic regression. However, once involving a wider gene subset, *CTNNB1* became a negative predictor of *APC* mutations (OR = 0.669, [95% CI: 0.564, 0.793], $p < 0.001$) (**Fig. 3.7**). Logistic regression with feature selection and cross-validation highlighted *CTNNB1*

mutations as negative predictors for the presence of an *APC* mutation (OR = 0.711, [95% CI: 0.589, 0.845], $p < 0.001$) (**Fig. 3.9**).

RSPO fusions were not significantly associated with *APC* mutations in MSI⁺ CRCs, via any statistical test.

3.3.1.4 Mutations in mini-driver genes in the context of *APC*-mutant, MSS CRCs

Via various analyses, mutations in multiple genes have shown a positive association with mutations in *APC*. These include mutations in *AMER1*, *FBXW7*, *SOX9* and *TCF7L2*.

The Spearman's rank test evidences a positive association between mutations in *AMER1* and *APC* ($r = 0.085$, $p = 0.003$) (**Fig. 3.2**). In addition the pairwise logistic regression (Method A) also evidenced a positive association between mutations in *AMER1* and *APC*. *AMER1* mutations were associated with an increased likelihood of the occurrence of *APC* mutations (OR = 1.175, [95% CI: 1.07, 2.29], $p < 0.001$) (**Fig. 3.4**). This relationship remained strongly positive following logistic regression Method B (OR = 1.178, [95% CI: 1.081, 1.283], $p < 0.001$) (**Fig. 3.6**). *AMER1* was also indicated an important positive predictor of *APC* mutations via logistic regression with feature selection and cross-validation (OR = 1.506, [95% CI: 1.306, 1.737], $p < 0.001$) (**Fig. 3.8**). These results indicate *AMER1* to be a putative candidate for a Wnt pathway mini-driver gene through its strong association with major-driver *APC*. Hypothetically, the presence of an *AMER1* mutation may allow an adjustment of Wnt signalling to account for a sub-optimal number of 20AARs in *APC*. This is explored further in Section 3.3.2.

The Spearman's rank test also indicated a subtle, positive association between mutations in *FBXW7* and *APC* ($r = 0.056$, $p = 0.049$) (**Fig. 3.2**). This was not found by either logistic regression Method A or B. However, *FBXW7* was highlighted as a positive predictor of *APC* mutations via logistic regression with feature selection and cross-validation (OR = 1.407, [95% CI: 1.273, 1.556], $p < 0.001$) (**Fig. 3.8**). These results also implicate *FBXW7* as a possible Wnt pathway mini-driver gene through its association with *APC*. Further insight is provided as to whether *FBXW7* is associated with a change in 20AARs in *APC* in section 3.3.2.

A positive association was also found between mutations in *SOX9* and *APC*, initially via Spearman's rank ($r = 0.074$, $p = 0.009$) (**Fig. 3.2**). This positive association was also evidenced via pairwise logistic regression (Method A) in which *SOX9* mutations were associated with an increased likelihood of the occurrence of a mutation in *APC* (OR = 1.097, [95% CI: 1.026, 1.172], $p = 0.006$) (**Fig. 3.4**). This result was replicated via logistic regression Method B (OR = 1.075, [95% CI: 1.012, 1.142], $p = 0.019$) (**Fig. 3.6**). Logistic regression with feature selection and cross-validation also indicated *SOX9* as a statistically significant, positive predictor of mutations in *APC* (OR = 1.4, [95%

CI: 1.266, 1.55], $p < 0.001$) (**Fig. 3.8**). Collectively, these results implicate *SOX9* as a potential Wnt pathway, mini-driver gene.

Finally, a positive correlation between mutations in *TCF7L2* and *APC* is observed via Spearman's rank ($r = 0.068$, $p = 0.017$) (**Fig. 3.2**). A positive association between mutations in *TCF7L2* and *APC* was also evidenced via logistic regression (Method A) in which *TCF7L2* mutations were associated with an increased likelihood of the occurrence of *APC* mutations (OR = 1.077, [95% CI: 1.008, 1.149], $p = 0.027$) (**Fig. 3.4**). A similar, positive association was found via logistic regression Method B (OR = 1.08, [95% CI: 1.017, 1.147], $p = 0.012$) (**Fig. 3.6**). Finally, *TCF7L2* was found to be a statistically significant, positive predictor of mutations in *APC* via logistic regression with feature selection and cross-validation (OR = 1.516, [95% CI: 1.373, 1.674]) (**Fig. 3.8**). These results indicate that *TCF7L2* as a candidate Wnt pathway mini-driver gene through its positive association with *APC*.

3.3.1.5 Mutations in mini-driver genes in the context of *APC*-mutant, MSI⁺ CRCs

Notably, in MSI⁺ CRCs, relationships between mutations in *APC* and other Wnt pathway genes showed key differences to the patterns observed in MSS CRCs, perhaps owing to a low frequency of *APC* mutations in MSI⁺ CRCs.

TCF7L2 mutations were the only mutations that showed a positive association with *APC* mutations via Spearman's rank ($r = 0.165$, $p = 0.004$) (**Fig. 3.3**). This finding was replicated via pairwise logistic regression (Method A) in which *TCF7L2* mutations were associated with an increased likelihood of the occurrence of *APC* mutations (OR = 1.144, [95% CI: 1.036, 1.264], $p = 0.008$). A slightly higher odds ratio was found when utilising *APC* as the predictor variable and *TCF7L2* as the outcome variable, meaning that *APC* mutations were associated with an increased likelihood of *TCF7L2* mutation (OR = 1.192, [95% CI: 1.235, 1.826], $p = 0.008$) (**Fig. 3.5**). *TCF7L2* was not a statistically significant predictor of mutations in *APC* via logistic regression Method B or by logistic regression with feature selection and cross-validation.

Interestingly, *FBXW7* mutations were shown to be associated with an increased likelihood of *APC* mutations, only via pairwise logistic regression (OR = 1.117, [95% CI: 1.003, 1.244], $p = 0.044$) (**Fig 3.5**). This positive association was not evidenced by any other test.

3.3.1.6 Other mini/major-driver relationships

Notably, the majority of statistical tests, both in MSS and MSI⁺ CRCs, *TCF7L2* mutations were positively associated with mutations in *CTNNB1*.

Starting with MSS, the Spearman's rank evidenced a strong, positive association between mutations in *TCF7L2* and *CTNNB1* ($r = 0.105$, $p < 0.001$) (**Fig. 3.2**). This association became even

stronger in the MSI⁺ CRCs ($r = 0.247$, $p < 0.001$) (**Fig. 3.3**). For MSS samples, logistic regression Method A indicated that *TCF7L2* mutations are associated with an increase in the likelihood of *CTNNB1* mutations (OR = 1.034, [95% CI: 1.011, 1.058], $p = 0.004$) (**Fig. 3.4**). This positive association was also shown via logistic regression Method A using MSI⁺ CRCs (OR = 1.143, [95% CI: 1.071, 1.219], $p < 0.001$) (**Fig. 2.6**). Again, through logistic regression Method B, *TCF7L2* was associated with an increase in the likelihood of a *CTNNB1* mutation in the context of MSS CRCs (OR = 1.038, [95% CI: 1.015, 1.061], $p < 0.001$) (**Fig. 3.6**). This was replicated utilising MSI⁺ CRCs (OR = 1.109, [95% CI: 1.041, 1.182], $p < 0.001$) (**Fig. 3.7**). Notably in MSS CRCs, *TCF7L2* mutations were not implicated as statistically significant predictors of *CTNNB1* mutations. However, *CTNNB1* mutations were statistically significant predictors of *TCF7L2* mutations, meaning that *CTNNB1* mutations were associated with an increased likelihood of *TCF7L2* mutation (OR = 1.347, [95% CI: 1.165, 1.556], $p < 0.001$) (**Fig. 3.8**). In MSI⁺ CRCs, *CTNNB1* mutations were positively associated with an increased likelihood of *TCF7L2* mutations (OR = 1.114, [95% CI: 1.046, 1.188], $p < 0.001$) and vice versa (OR = 1.829, [95% CI: 1.501, 2.23], $p < 0.001$) (**Fig. 3.9**). These results implicate *TCF7L2* as a potential Wnt pathway mini-driver in the context of *CTNNB1* mutations.

ZNRF3 mutations also consistently co-occurred with *RNF43* mutations in both MSS and MSI⁺ CRCs. Firstly, a positive association was found between the two genes via Spearman's rank in MSS ($r = 0.15$, $p < 0.001$) (**Fig. 3.2**) and MSI⁺ CRCs ($r = 0.295$, $p < 0.001$) (**Fig. 3.3**). Following this, a pairwise logistic regression showed that *ZNRF3* mutations were associated with an increased occurrence of *RNF43* mutations in MSS (OR = 1.446, [95% CI: 1.26, 1.659], $p < 0.001$) (**Fig. 3.4**) and MSI⁺ CRCs (OR = 1.231, [95% CI: 1.107, 1.369], $p < 0.001$) (**Fig. 3.5**). This result was replicated by logistic regression Method B in MSS (OR = 1.341, [95% CI: 1.177, 1.528], $p < 0.001$) (**Fig. 3.6**) and MSI⁺ CRCs (OR = 1.124, [95% CI: 1.02, 1.239], $p < 0.001$) (**Fig. 3.7**). In MSS CRCs, *ZNRF3* mutations were not implicated as statistically significant predictors of *RNF43* mutations, however, *ZNRF3* mutations were associated with a slight increase in the likelihood of an *RNF43* mutation (OR = 1.064, [95% CI: 1.04, 1.088], $p < 0.001$) (**Fig. 3.8**). In MSI⁺ CRCs, *ZNRF3* mutations were associated with a slight increased likelihood of *RNF43* mutation (OR = 1.129 [95% CI: 1.026, 1.242], $p = 0.013$). However, the presence of an *RNF43* mutation had a strong, positive association with *ZNRF3* mutations (OR = 1.442, [95% CI: 1.361, 1.528], $p < 0.001$) (**Fig. 3.9**).

In the context of MSI⁺ CRCs, *RNF43* and *TCF7L2* mutations were consistently negatively associated. Initially, the Spearman's rank showed a negative association ($r = -0.228$, $p < 0.001$) (**Fig. 3.3**). The pairwise logistic regression showed *TCF7L2* to be strongly associated with a decrease in the likelihood of an *RNF43* mutation (OR = 0.812, [95% CI: 0.739, 0.892], $p < 0.001$) (**Fig. 3.5**). Logistic regression Method B also showed *TCF7L2* mutations to be a negative predictor of *RNF43* mutations (OR = 0.907, [95% CI: 0.829, 0.992], $p = 0.033$) (**Fig. 3.7**). This was again replicated through logistic

regression with feature selection and cross validation (OR = 0.908, [95% CI = 0.832, 0.991], $p = 0.03$) (Fig. 3.9).

Again, in MSI⁺ CRCs, *RNF43* and *BCL9* mutations were consistently negatively associated. Initially, the Spearman's rank showed a negative association between the two mutations ($r = -0.124$, $p = 0.032$) (Fig. 3.3). The pairwise logistic regression (Method A) then showed *BCL9* mutations to be associated with a reduction in the occurrence of mutations (OR = 0.866, [95% CI: 0.783, 0.958], $p = 0.005$) (Fig. 2.6). Logistic regression Method B supported this finding (OR = 0.905, [95% CI: 0.828, 0.99], $p = 0.030$) (Fig. 3.7). The logistic regression with feature selection and cross validation also evidenced *BCL9* to be a negative predictor of the occurrence of *RNF43* mutations (OR = 0.906, [95% CI: 0.83, 0.989], $p = 0.028$) (Fig. 3.9).

3.3.2 The number of 20AARs in *APC* in the presence of other Wnt pathway mutations (mini-drivers)

To further investigate the just-right hypothesis within the context of *APC*'s 20AARs, I examined the number of 20AARs in each sample with two pathogenic mutations in *APC*. In addition, I investigated the impact of additional Wnt pathway mutations on the number of 20AARs. Mutations in the following genes were investigated: *AMER1*, *APC*, *AXIN1*, *AXIN2*, *BCL9*, *BCL9L*, *CTNNB1*, *FBXW7*, *JUN*, *RNF43*, *SOX9*, *TCF7L2* and *ZNRF3*. Initially, a Shapiro-Wilk test for normality was conducted on both MSS and MSI⁺ datasets separately. This confirmed that both datasets were not normally distributed ($p < 0.05$). Multiple Mann-Whitney U tests were then conducted to assess whether the occurrence of a mutation in various Wnt pathway related genes was associated with a change in 20AARs in *APC*. P-values were then corrected via Benjamini-Hochberg correction procedure. The results indicated that the presence of an *AMER1* mutation is associated with a significant increase in the number of 20AARs in *APC* in MSS CRCs (Table 3.2). No change in 20AARs was found in MSI⁺ CRCs. This suggests that the presence of an *AMER1* mutation could adjust the just-right level of Wnt signalling to account for a sub-optimal number of 20AARs (Fig. 3.10).

Table 3.3 The association between the number of 20AARs in *APC* and additional Wnt pathway mutations in primary MSS CRCs: Mann-Whitney U results. A Mann-Whitney U test was utilised to assess whether the number of 20AARs in *APC* was different depending on the presence of multiple different Wnt pathway mutations. The results show a significant change in the number of 20AARs in the presence of an *AMER1* mutation following p-value correction via Benjamini-Hochberg correction procedure.

Mutation in gene	U	p-value	Corrected p-values
<i>AMER1</i>	40856.0	0.000	0.000
<i>AXIN2</i>	7502.5	0.660	0.998
<i>BCL9</i>	11808.0	0.091	0.615
<i>BCL9L</i>	20321.0	0.829	0.999
<i>CTNNB1</i>	3501.0	0.480	0.995
<i>FBXW7</i>	53596.0	0.042	0.375
<i>JUN</i>	4550.5	0.589	0.998
<i>SOX9</i>	50979.5	0.753	0.999
<i>TCF7L2</i>	47780.5	0.395	0.989

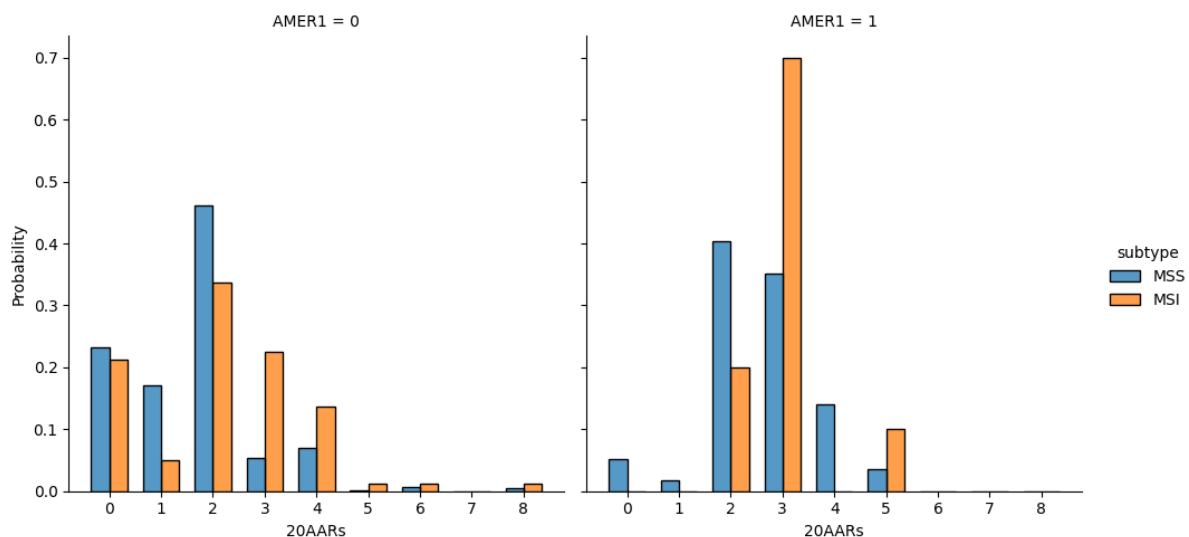


Figure 3.10 The number of 20AARs in *APC* in the presence of an *AMER1* mutation in both MSS and MSI⁺ CRCs. A Mann-Whitney U test, followed by Benjamini-Hochberg correction, determined that, in the presence of an *AMER1* mutation, the number of 20AARs in *APC* is significantly increased in MSS CRCs.

3.3.3 The number of 20AARs in *APC* in the proximal and distal colon/rectum

To assess whether the just-right model is the same in the proximal and distal colon/rectum, I examined whether the number of 20AARs in *APC* is consistent across the different locations. The data

was filtered down to only include those samples of definitive location. In total, there were 361 proximal colon samples and 507 distal/rectal samples. Notably, following a Shapiro-Wilk and Mann-Whitney U test, the results indicated that the number of 20AARs in *APC* is higher in the proximal colon than the distal colon/rectum in MSS samples ($U = 100332.0$, $p < 0.001$) (**Fig. 3.11**). In MSI⁺ samples, no significant difference was found between the number of 20AARs in the proximal and distal colon/rectum.

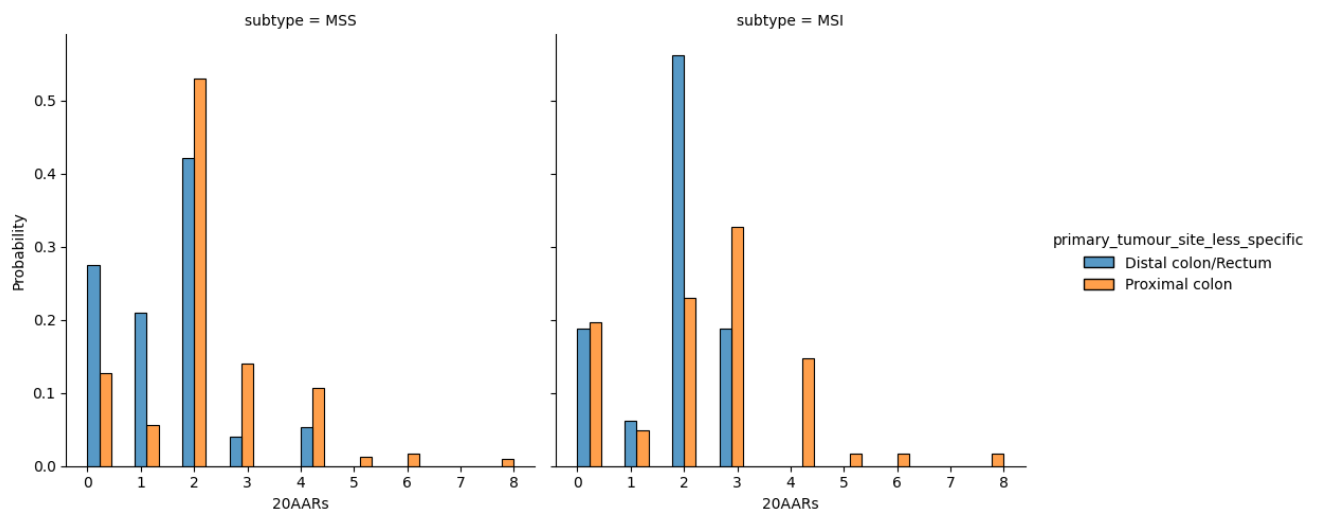


Figure 3.11 The distribution of 20AARs in *APC* in the proximal and distal colon/rectum. A significant difference was found between the number of 20AARs in *APC* between the proximal and distal colon/rectum in MSS CRCs.

I then examined if the presence of an additional Wnt pathway mutation was associated with number of 20AARs in *APC* in the proximal and distal colon in MSS and MSI⁺ CRCs. Following a Shapiro-Wilk and Mann-Whitney U test, the results showed that the presence of an *AMER1* mutation was associated with an increase in the number of 20AARs in the proximal ($U = 6232.5$, $p = 0.005$) and the distal colon/rectum ($U = 4472.5$, $p = 0.008$) (**Table 3.4, Fig. 3.12**). No statistical difference in 20AARs was found in the proximal or distal colon/rectum in MSI⁺ CRCs.

Table 3.4. The distribution of 20AARs in *APC* in the proximal and distal colon, in the presence of additional Wnt pathway mutations, in MSS CRCs: A statistical assessment. A Mann-Whitney U was utilised to examine if there is a significant difference in the number of 20AARs in the proximal and distal colon in the presence of additional Wnt pathway mutations. The results indicate that in the presence of *AMER1*, the distribution of 20AARs in *APC* is altered in the proximal colon and distal colon/rectum, in MSS CRCs.

Mutation in gene	Statistic		p-value		Corrected p-value	
	Proximal colon	Distal colon/Rectum	Proximal colon	Distal colon/Rectum	Proximal colon	Distal colon/Rectum
<i>AMER1</i>	6232.5	4472.5	0.0003	0.0005	0.005	0.008
<i>AXIN2</i>	530.0	1168.0	0.698	0.385	0.999	0.995
<i>BCL9</i>	384.0	4093.0	0.657	0.782	0.999	0.999
<i>BCL9L</i>	2275.0	4330.5	0.146	0.529	0.906	0.999
<i>CTNNB1</i>	384.0	970.5	0.657	0.300	0.999	0.986
<i>FBXW7</i>	3683.0	12627.0	0.521	0.175	0.999	0.932
<i>JUN</i>	667.5	433.5	0.636	0.200	0.999	0.945
<i>SOX9</i>	5351.0	10123.0	0.566	0.919	0.999	0.999
<i>TCF7L2</i>	3022.5	12681.5	0.051	0.873	0.565	0.999

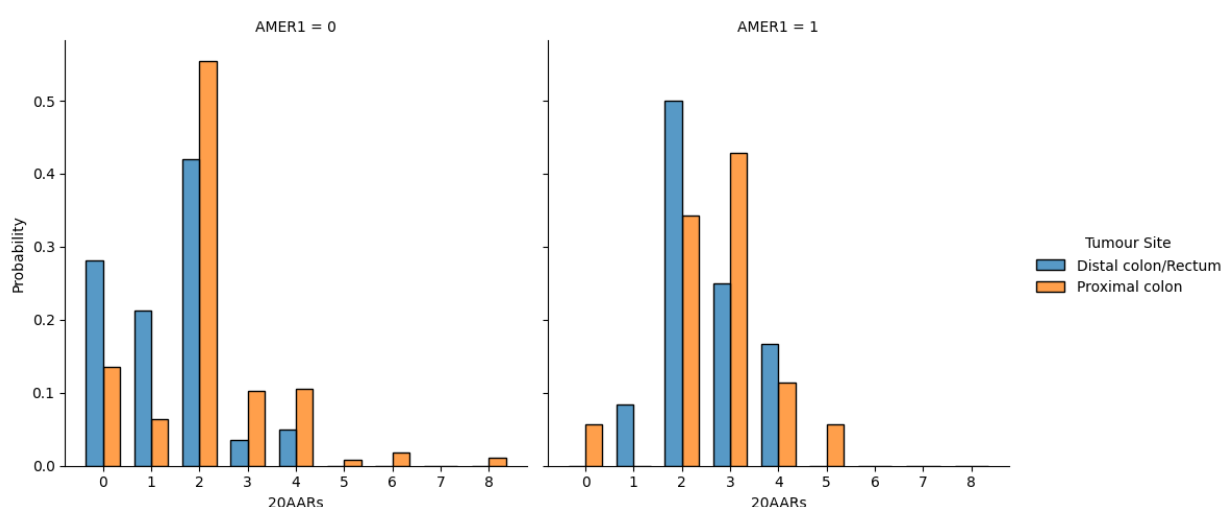


Fig. 3.12 The presence of an *AMER1* mutation in the proximal colon and distal colon/rectum is associated with an increase in the number of 20AARs in *APC*.

3.3.4 The position of *AMER1* mutations with and without a mutation in *APC*

In total, there were 83 CRCs with a mutation in *AMER1*. The location and type of these mutations can be found in Figure 3.13.

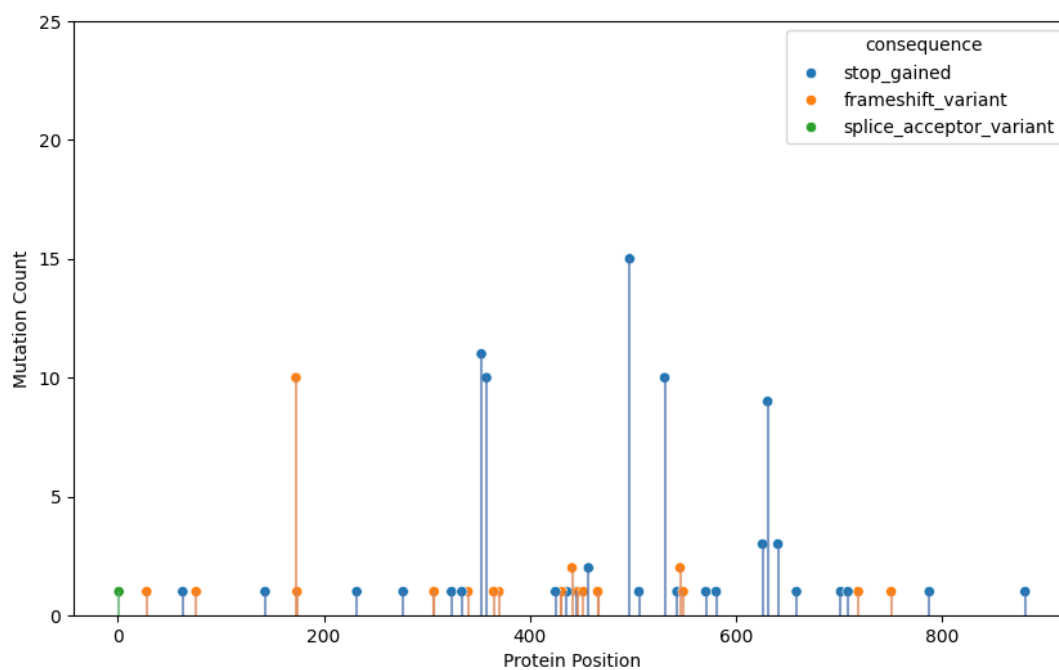


Figure 3.13 A lollipop plot representing all *AMER1* mutations in primary, treatment naïve CRCs in the 100KGP. In summary, there are six mutation sites with a count of above nine. These mutations reside in M1, A1, A2 and the REA of *AMER1*.

Over 9 cases of mutations are found in each M1, A1, A2 and the REA. M1 and M2 are responsible for membrane localisation, they are the N-terminal membrane localisation domains of *AMER1*. A1, A2 and A3 are *APC* interaction domains. The REA, standing for repeated arginine-glutamic acid-alanine motifs, interact with the armadillo repeats of β -catenin (**Fig. 3.14**). These mutations potentially have a range of effects on the resulting protein.

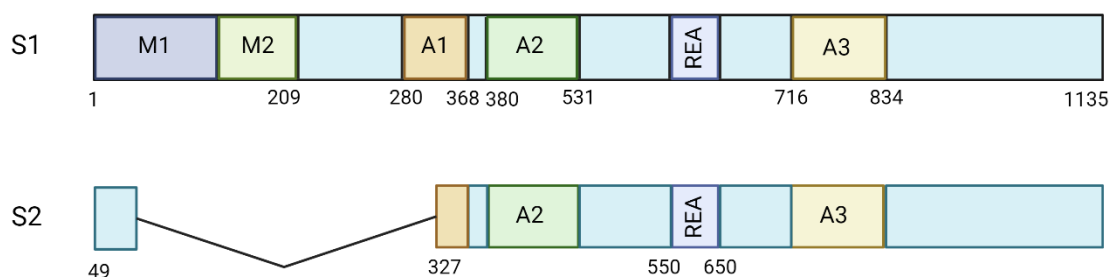


Figure 3.14 The protein structure of the two isoforms of *AMER1* (S1 and S2). Adapted from Tanneburger *et al.* (2011). This diagram shows the N-terminal membrane localisation domains (M1, M2), *APC* interaction domains (A1, A2, A3) and the REA responsible for interacting with the armadillo repeats of β -catenin.

Out of 83 *AMER1* mutations, 69 also occurred in the presence of an *APC* mutation. The mutations with the highest counts occurred in A1, A2 and the REA (**Fig. 3.15**). These mutations maintain the N-terminal membrane localisation domains of *AMER1*.

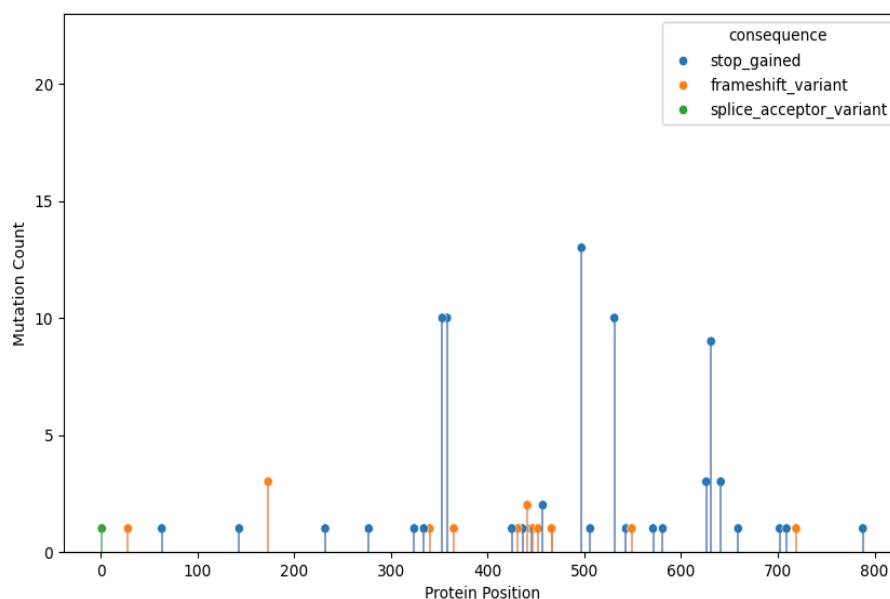


Figure 3.15 A lollipop plot representing all *AMER1* mutations also harbouring a mutation in *APC*, in primary, treatment naïve CRCs in the 100KGP. In summary, there are five mutation sites with a count of above nine. These mutations reside in A1, A2 and the REA of *AMER1*.

98% of MSS CRCs harbouring an *AMER1* mutation, also contained an *APC* mutation. The position of these mutations can be seen in Figure 3.16. Again, the N-terminal membrane localisation domains are maintained with the majority of *AMER1* mutations present in MSS CRCs.

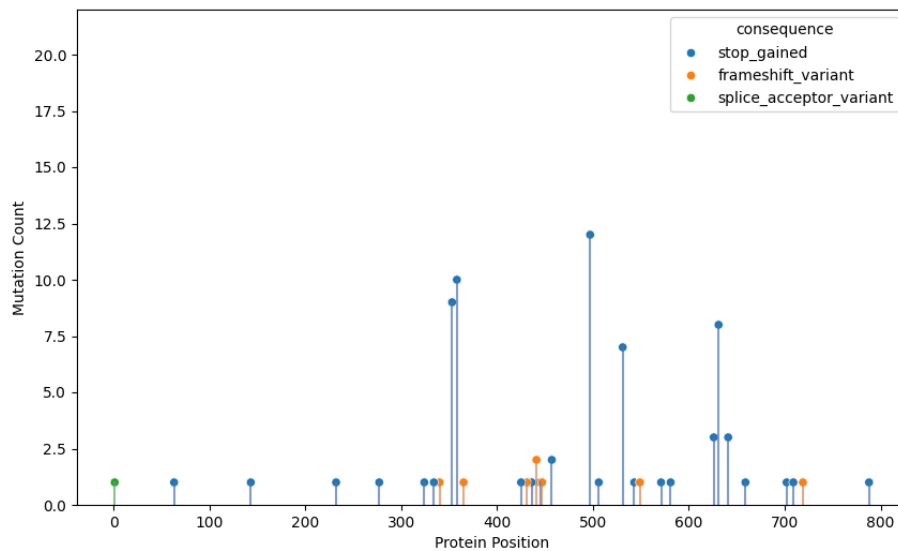


Figure 3.16 A lollipop plot representing all *AMER1* mutations also harbouring a mutation in *APC*, in primary, MSS, treatment naïve CRCs in the 100KGP. In summary, the most common mutations reside in M1, A1, A2 and the REA of *AMER1*.

In total, there were 23 MSI⁺ CRCs that harboured a mutation in *AMER1*. 43% of tumours harbouring an *AMER1* mutation also contained an *APC* mutation. As stated in section 3.3.1.5 *AMER1* and *APC* mutations were not found to correlate with one another in the context of MSI⁺ CRCs. The location of these mutations can be found in Figure 3.17.

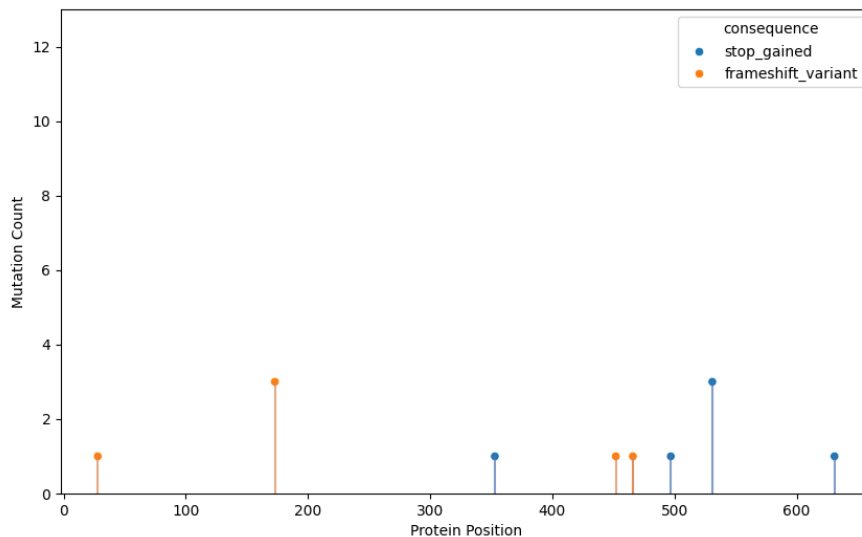


Figure 3.17 A lollipop plot representing all *AMER1* mutations also harbouring a mutation in *APC*, in primary, MSI⁺, treatment naïve CRCs in the 100KGP. In summary, the most common mutations reside in M1, A1, A2 and the REA of *AMER1*.

Figure 2.18 shows the presence of two mutations in the N-terminal region of *AMER1*, residing in M1 and M2. The other mutations are spread across A1, A2 and the C-terminal region of *AMER1*. The lack of *AMER1* mutations and their position in MSI⁺ CRCs, highlights a difference in their importance in MSI⁺ CRCs compared to MSS CRCs.

3.3.5 Age differences in MSS and MSI⁺ CRCs

In addition to differences in mutated genes, following a Shapiro-Wilk and subsequent Mann-Whitney U test, the data showed a statistical difference between the ages of those patients with MSI⁺ and MSS CRCs ($U = 261556.5$, $p < 0.001$) (**Fig. 3.18**).

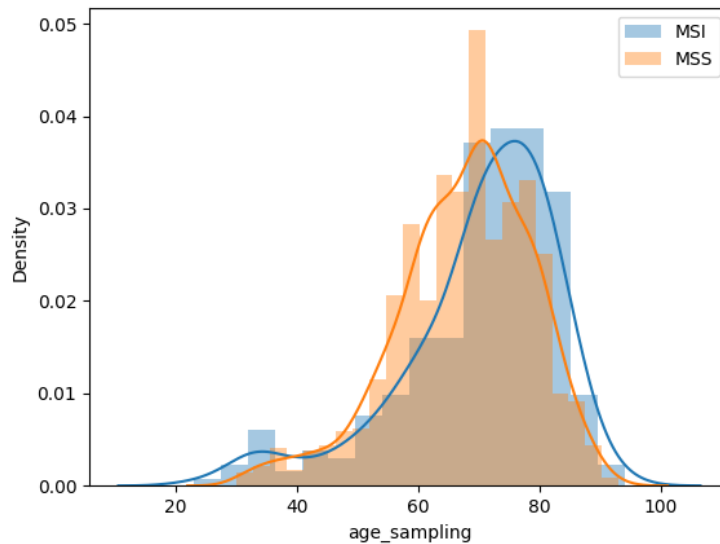


Figure 3.18 MSI⁺ CRCs have a significantly higher age at sampling than MSS CRCs.

No difference was found in the age of patients with MSS CRCs that either harboured or did not harbour an *APC* mutation ($U = 125189.5$, $p = 0.347$). However, in MSI⁺ CRCs, patients that harboured *APC* mutations were significantly younger than those without *APC* mutations ($U = 7594.5$, $p < 0.001$) (**Fig. 3.19**). No patients with diseases of Mendelian inheritance were included in this analysis.

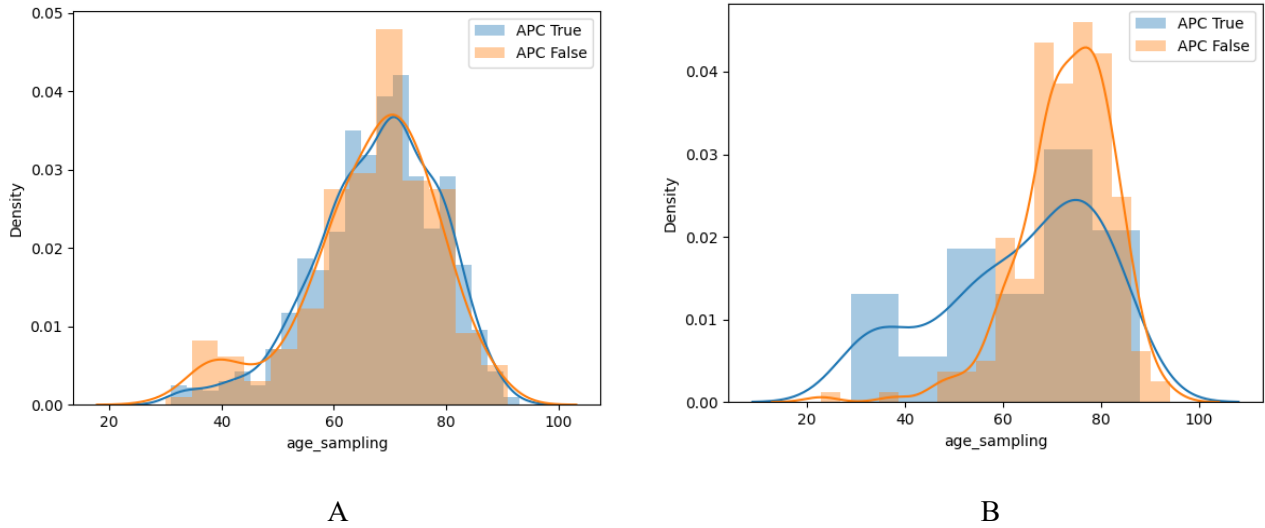


Figure 3.19 *APC* mutations are associated with a younger sampling age in MSI⁺ CRCs. A) There is no significant difference in the age of patients with or without a mutation in *APC* in MSS CRCs. B) Patients with a CRC harbouring an *APC* mutation were significantly younger than those without an *APC* mutation in MSI⁺ CRCs.

Again, following Shapiro-Wilk and Mann-Whitney U tests, no significant difference was found between the ages of those with and without a mutation in *RNF43* in MSS CRCs ($U = 21689.5$, $p = 0.926$). Conversely, *RNF43* mutations were found to be associated with an increase in the age at sampling in MSI⁺ CRCs ($U = 13983$, $p < 0.001$) (**Fig. 3.20**).

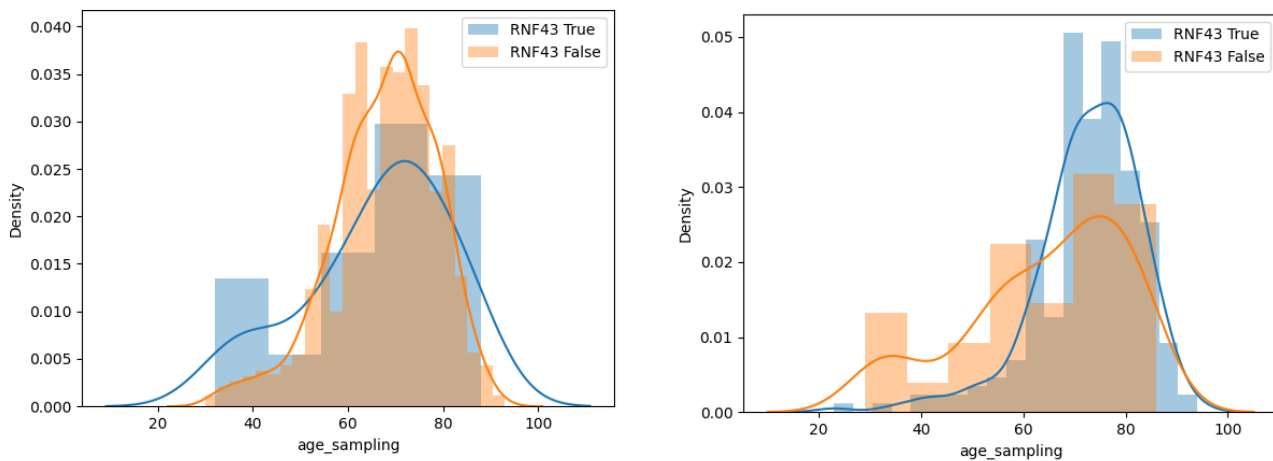


Figure 3.20 *RNF43* mutations are associated with a higher sampling age in MSI⁺ CRCs. A) There is no significant difference in the age of patients with or without a mutation in *RNF43* in MSS CRCs. B) Patients with a CRC harbouring an *RNF43* mutation were significantly younger than those without an *RNF43* mutation in MSI⁺ CRCs.

3.3.6 Exploring the just-right model in FAP

To further explore the just-right model, Xell Brunet and I analysed pathogenic Wnt-pathway mutations in multiple polyp samples from two patients. Patient 1's polyp samples were taken from the

rectum, whilst patient 2's polyp samples were taken from the colon, with a cancer sampled from the rectum. By looking at both the germline and somatic *APC* mutations, we were able to determine the remaining number of 20-AARs in *APC*, per sample.

The results reveal a variety of number of 20AARs in multiple samples from the same patient. Notably, in patient 2's cancer sample, *AMER1* is associated with an increased in 20AARs which supports the findings presented in section 2.3.2 (**Table 3.5**).

Table 3.5 An investigation of the just-right model in FAP. This table summarises the pathogenic Wnt pathway mutations harboured by multiple polyp/cancer samples from two FAP patients. The table details the positions in which germline and somatic *APC* mutations are located and thus the remaining number of 20AARs.

Patient	Sample	<i>APC</i> germline	<i>APC</i> somatic	Total 20AARs	Others
1	Polyp 1	R0	R2	2	NA
1	Polyp 2	R0	R0	0	NA
1	Polyp 3	R0	R2	2	<i>AMER1</i>
1	Polyp 4	R0	R0	0	<i>AMER1 & SOX9</i>
2	Polyp 1	R0	R2	2	<i>TCF7L2</i>
2	Polyp 2	R0	R1	1	NA
2	Polyp 3	R0	R3	3	NA
2	Polyp 4	R0	R1	1	NA
2	Polyp 5	R0	R2	2	NA
2	Cancer	R0	R3	3	<i>AMER1</i>

3.4 Discussion

In this chapter, the relationships between mutations in the Wnt pathway were thoroughly examined using a series of statistical methods. An initial Spearman's rank analysis identified broad relationships of interest. This was followed by a logistic regression that controlled for potential confounders such as location, sex and mutation burden, providing a clearer picture of the direct associations between mutations. To account for broader gene interactions, a subsequent logistic regression incorporated all genes as covariates to account for any broader interactions which may influence the relationship in question. Finally, logistic regression combined with feature selection and cross-validation was employed to further refine the model, pinpoint the most relevant gene interactions, minimise overfitting and ensure robust results. Together, these methods offer a comprehensive and nuanced understanding of the patterns of mutual exclusivity and co-occurrence among Wnt pathway mutations.

In most cases, consistent relationships were identified across all tests, indicating robust associations. However, there were some examples of differing outcomes depending on the methodology used. For example, *FBXW7* only showed a significant, positive association with *APC* via logistic regression only when feature selection and cross validation were employed. By incorporating mutations in additional genes as co-variates into the model, it is able to account for more confounding factors than when these mutations are excluded, potentially allowing for clarification of the true relationship between mutations in *FBXW7* and *APC*. Conversely, adding information about mutations in all Wnt pathway genes without feature selection and cross-validation can lead to overfitting. This occurs when the model becomes too tailored to the training dataset's specific nuances, potentially failing to capture true relationships in new, unseen data.

The mutually exclusive relationships found between mutations in Wnt pathway major-driver genes *RNF43*, *CTNNB1*, RSPO fusions and *APC* support the just-right model, suggesting that mutations in these genes result in excessively high Wnt signalling levels that are disadvantageous to the cell. These findings also validate the methodology and serve as controls, as these specific, mutually exclusive relationships have been previously documented. For example, Giannakis *et al.* (2014) analysed 222 colorectal exomes and found the same mutually exclusive relationship between *RNF43* and *APC* (Giannakis *et al.*, 2014b). As previously stated, Seshagiri *et al.* (2012) evidenced a mutually exclusive relationship between RSPO fusions and *APC* mutations (Seshagiri *et al.*, 2012). A mutually exclusive relationship between *CTNNB1* and *APC* mutations has been widely documented and theorised as evidence of the just-right model of CRC evolution (Segditsas and Tomlinson, 2006). This study therefore confirms these findings on a larger scale, and provides support for the just-right model in CRC through mutually exclusive relationships between Wnt pathway, major-driver mutations and mutations in *APC*.

This data also identifies multiple candidate Wnt pathway mini-driver genes through a large-scale analysis of co-occurrence between known major-driver mutations and other mutations in Wnt pathway genes, providing further evidence for the role of the just-right model in CRC evolution. In MSS CRCs, a pattern of co-occurrence was observed between mutations in *APC* and mutations in *AMER1*, *FBXW7*, *SOX9* and *TCF7L2*. This potentially provides supportive evidence for the just-right model, through the association between candidate Wnt pathway mini-driver mutations and major-driver mutations in *APC*.

Notably, mutual exclusivity was observed consistently between mutations in the major-driver genes (*RNF43*, *CTNNB1*, RSPO fusions) and *APC* mutations. However, mutually exclusive relationships were not consistently observed between mutations in each of the other major-driver genes. For example, a negative relationship between mutations in *RNF43* and *CTNNB1* is observed, but in MSI⁺ CRCs only. At surface level, this contradicts the just-right model as a co-occurrence in these mutations could result in a non-optimal level of Wnt signalling. However, no positive association was found either. Since *RNF43* mutations are much more prevalent in MSI⁺ CRCs, mutation frequency and distribution in different subtypes may explain why no particular relationship is shown between *RNF43* and *CTNNB1* mutations in MSS CRCs (Giannakis *et al.*, 2014a). This finding is therefore still supportive of the just-right model.

Additionally, no negative associations were observed between RSPO fusions and either *RNF43* or *CTNNB1* mutations. This lack of association could again be attributed to the low frequency of these mutations, which may not be sufficient enough in frequency to reveal any particular relationship. These relationships therefore do not definitively contradict the just-right model.

This study highlights clear distinctions in mutational relationships between MSS and MSI⁺ CRCs. This is as expected as colorectal polyps are categorised into different types based on key histological and clinical differences, these include: conventional adenomas, serrated lesions and other polyp types e.g. juvenile/inflammatory. Sporadic MSI and CpG island methylator phenotypes (CIMP) often lead to the development of serrated lesions. In these cases, inactivating mutations in *APC* are rare and do not induce carcinogenesis. Instead, serrated lesions are frequently associated with *BRAF* mutations and a high mutation burden due to mismatch repair mutation or methylation of *MLH1* (De Palma *et al.*, 2019). This high mutation burden subjects these tumours to different selective pressures, altering the significance of Wnt pathway mutations. As a result, the just-right model may manifest differently in the Wnt pathway of MSI⁺ CRCs compared to conventional, *APC*-mutated CRCs. For example, *RNF43* mutations are much more prevalent than *APC* mutations in the context of MSI (Giannakis *et al.*, 2014a). The positive association between *SOX9*, *AMER1*, *FBXW7* and *APC* mutations are no longer evidenced. Conversely, *RNF43* mutations show interesting relationships with candidate Wnt pathway mini-driver mutations in MSI⁺ CRCs. *RNF43* mutations strongly co-occurred with

mutations in *ZNRF3*, a relationship supported by previous literature (Farnhammer, Colozza and Kim, 2023). This co-occurrence may represent a major/mini-driver relationship, providing supporting evidence for the just-right model. However, a negative association was observed between *RNF43* mutations and mutations in other candidate Wnt pathway mini-driver genes such as *TCF7L2* and *BCL9*. This may contradict the just-right model, or perhaps indicate its complexity and modification based on tumour context and the importance of other pathways in carcinogenesis. Future work may look into whether the just-right model can be extended across both the Wnt and MAPK pathway, particularly in the context of MSI and serrated lesions.

Interestingly, *TCF7L2* mutations were found to co-occur robustly with *CTNNB1* mutations in both MSS and MSI⁺ tumours, potentially representing an alternative just-right combination of mutations.

Further differences between MSS and MSI⁺ CRCs are observed by examining the sampling age of CRC patients. The results show that MSI⁺ tumours are more common in older patients. Previous studies have documented a proximal shift in CRCs with increasing age. Evidence also suggests that frequency of *MLHI* promoter methylation increases with age potentially due to environmental influences (Yiu *et al.*, 2005). If *MLHI* promoter methylation drives carcinogenesis in older individuals and *RNF43* mutations are prevalent in MSI⁺ CRCs, it therefore is expected that our data shows an association between *RNF43* mutations and increased age. While *APC* mutations occur less frequently in MSI tumours, our data suggests that when they do occur, they are found in younger patients. The high mutation burden associated with MSI might explain why *APC* mutations appear earlier in these cases. The *APC* gene, being approximately 138 kb in size, is more susceptible to mutations than smaller genes (Aoki and Taketo, 2007). This may contribute to its early involvement in MSI⁺ CRCs.

The results also support the theory that mini-driver mutations can fine-tune the action of a major-driver mutation. This is shown by an increase in the optimal number of 20AARs in *APC* in the presence of an *AMER1* mutation. The FAP data also provides supporting evidence for this association between *AMER1* and an increased number of 20AARs in *APC* as this combination occurred in the only FAP-associated cancer sample.

The number of 20AARs in *APC* was found to be higher in the proximal colon of MSS CRCs than the distal colon/rectum. This suggests that the mutations selected for and the just-right number of mutations may be different depending on tumour location. It is also important to note that the tumour microenvironment influences the transcriptional activity of Wnt genes. Thus, dependent on the location of the tumour, different selective pressures may influence the optimal level of Wnt signalling (Ramakrishnan *et al.*, 2023). This variability may be a contributing factor in the observed differences in combinations of mutations between tumours in the proximal and distal colon/rectum.

An initial limitation of the mutational co-occurrence analysis is the utilisation of binary matrix on which it was based. Cornish *et al.* (2024) produced a binary matrix cataloguing the presence or absence of pathogenic mutations across all 185 candidate driver genes. A significant disadvantage of basing the analysis on this data, is its exclusion of LOH events. As outlined in section 2.1, LOH events are central to the just-right model in relation to *APC*. This may also be true of other TSGs such as *RNF43*. By utilising this data, it is not clear whether one or two-hits have occurred to all TSGs within each sample and therefore whether the TSG has truly lost its function. However, clear categorisation of all genes utilised in this study as either TSG or oncogene is not yet available in the literature. Thus, it is not yet possible to know whether all of these genes require one or two-hits to achieve LOF. In conclusion, while this analysis provides a broad overview of mutation co-occurrence, incorporating LOH events would be of implicit value to realising the precise just-right model necessary for tumourigenesis, in the Wnt signalling pathway.

A second caveat of the correlation analysis, based on the binary matrix, is the categorisation of mutations as pathogenic or non-pathogenic. To produce the binary matrix, Cross *et al.* (2022) standardised the criteria they utilised to categorise mutations as pathogenic or otherwise, across all genes. Their methodology encompasses the utilisation of variant effect predictor tool OncoKB, an assessment of missense mutation recurrence and the presumption that all genes with an ambiguous classification of neither TSG or oncogene lose their function upon acquisition of a truncation mutation. This uniform approach enables a broad overview of likely pathogenic mutation co-occurrence across all TSGs, oncogenes and genes of ambiguity. However, this co-occurrence analysis would benefit from a more precise approach to determining the pathogenicity of mutations. For example, *APC* is frequently mutated in CRC and thus, there is relatively extensive knowledge about the pathogenicity of mutations in *APC*. Previous studies have shown that truncation mutations tend to occur in the MCR of *APC*, resulting in the removal of SAMP, and therefore AXIN binding ability, 20AAR3, the CID and all domains harbouring the ability to alter β -catenin levels (Schneikert *et al.*, 2014). It could therefore be argued that any mutations downstream of the SAMP have little affect on the functionality of *APC*. Yet, in the binary matrix produced by Cross *et al.* (2022), truncation mutations, as well as any recurrent missense mutations, throughout the entirety of *APC*, would be classified as pathogenic. For this reason, further analyses in this thesis adopts a more detailed approach to assess mutation pathogenicity, whilst recognising that the depth of understanding varies significantly among genes. Nevertheless, this methodology remains useful for providing a generalised overview of mutation co-occurrence, especially when comparable insights are lacking for less frequently mutated genes.

Notably, the Wnt pathway can be dysregulated via alternative mechanisms, not explored in this chapter. For example, a study by Aguilera *et al.* (2006) evidenced the transcriptional silencing of Wnt inhibitor *DICKKOPF-1* (DKK1), in multiple colon cancer cell lines, through CpG island promoter hypermethylation. The canonical function of DKK1 is to modulate Wnt signalling by binding to LRP5/6

resulting in LRP endocytosis and therefore preventing β -catenin accumulation in the nucleus. Through the silencing of DKK1, signalling is increased. In addition to evidencing DKK1 silencing in colon cancer cell lines, Aguilera et al. (2006) also evidenced the hypermethylation of DKK1 in nine of 54 primary colorectal tumours and evidenced that DKK1 was selectively hypermethylated in advanced CRC. Theoretically, the silencing of extracellular Wnt modulators such as DKK1, secreted frizzled-related protein (SFRP1) and Wnt inhibitor factor-1 (WIF1) could significantly contribute to providing cells with a just-right level of signalling, providing a potential explanation for the selective silencing of DKK1 in advanced CRC. Further studies into the just-right model may wish to incorporate the hypermethylation status of Wnt-related extracellular modulates such as DKK1, SFRP1 and WIF1 (Aguilera *et al.*, 2006).

As described in section 1.3, AMER1 is of dual functionality. One function of AMER1 is to act as a scaffold and assemble the β -catenin destruction complex at the plasma membrane, leading to the subsequent ubiquitination and degradation of β -catenin (Tanneberger, Pfister, Kriz, *et al.*, 2011). Deletion of the membrane-binding domain of *AMER1* results in dysregulation of its negative effects on Wnt signalling. Conversely, AMER1 also functions to promote the phosphorylation of LRP6 leading to the inhibition of the destruction complex and increased β -catenin signalling (Tanneberger, Pfister, Brauburger, *et al.*, 2011). *AMER1* was identified as a candidate driver gene by Cornish *et al.* (2024) and mutations in *AMER1* have been shown to be common in CRC by previous studies (~10% of CRC cases) (Sanz-Pamplona et al., 2015; Cornish et al., 2024). This study however notes the strong correlation between mutations in *AMER1* and Wnt-pathway major-driver gene *APC*. In addition, a key novelty of this study is the identification of the correlation between *AMER1* and *APC* mutations resulting in three 20AARs, potentially providing evidence for *AMER1* mutations optimising Wnt-pathway signalling in the presence of a suboptimal number of 20AARs in *APC*. This study has also shown that in the context of *APC*-mutant, MSS CRCs, pathogenic, stop gained, *AMER1* mutations tend to occur within A1, A2 or the REA. If the mutations are occurring in AMER1-S1, this leaves both M1 and M2 N-terminal membrane-binding domains intact. Previous evidence shows that maintenance of this domain is essential for both activation and inhibition of Wnt signalling by AMER1 (Tanneberger, Pfister, Brauburger, *et al.*, 2011).

From the available data alone, it is not possible to know which isoform of AMER1 has been mutated. As stated above AMER1 exists in two isoforms due to splicing, AMER1-S1 and AMER1-S2. AMER1-S2 is lacking in membrane-binding domains and evidence suggests that knock-down of AMER1-S2 has no effect on either LRP6 phosphorylation or repression of Wnt signalling (Tanneberger, Pfister, Brauburger, *et al.*, 2011; Tanneberger, Pfister, Kriz, *et al.*, 2011). Therefore, if the evidenced *AMER1* mutations are occurring within AMER1-S2, this suggests that there is no overall effect on Wnt signalling. However, mutations deleting the REA of AMER1-S1, disrupt its interaction with β -catenin. Some AMER1 mutations occurring prior to the REA will evidently effect AMER1

binding to β -catenin. In which case, for AMER1 to have an inhibitory effect on Wnt signalling, these cells are reliant upon the association of AXIN and APC to cause an interaction between AMER1 and β -catenin. This interaction will evidently also be affected by *APC* mutations and thus functionality. Notably, there are also examples of AMER1 mutations that retain parts of the REA repeats in this study. These mutations result in a protein capable of β -catenin binding (Tanneberger, Pfister, Kriz, *et al.*, 2011). It would be of huge interest to explore whether the maintenance or absence of the REA repeats is associated with a particular number of 20AARs in *APC*. This would involve analysing the co-occurrence of specific mutations in both *AMER1* and *APC*.

Following a deeper exploration into the co-occurrence of specific mutations in *AMER1* and *APC*, various mutations could be replicated both in organoids and in mouse models. Hypothetically, according to the just-right/mini-driver model, *AMER1* mutations on their own act as mini-drivers and cannot cause carcinogenesis. However, once a specific mutation in *AMER1* is paired with a specific mutation in *APC*, this may result in a just-right level of β -catenin inhibition/LRP6 phosphorylation and may result in an increased rate of cell growth in organoids, or an increased tumour count in the proximal colon of a mouse model. It is possible that by interchanging the position of both the *AMER1* and *APC* mutation, the rate of cell growth and tumour count could be decreased by a the just-right level of Wnt signalling not being achieved. This experiment could also be repeated with other candidate mini-driver/major-driver mutations.

The dual functionality of genes is not limited to AMER1. AXIN has also been shown to have dual functionality (Sanz-Pamplona *et al.*, 2015b). In addition, the activity of SOX9 is highly context dependent. SOX9 has been evidenced to downregulate β -catenin protein levels, thus repressing the transcription of target Wnt genes. More recent evidence has shown the ability of SOX9 to bind to TCF resulting in transcriptional activation of Wnt target genes (Ramakrishnan *et al.*, 2023). Thus, different mutations in different genes may result in an unknown increase or decrease in Wnt signalling. It is therefore difficult to determine a just-right level of Wnt signalling with a lack of knowledge of the consequence of different mutations on the level of Wnt. It is therefore of great importance to the just-right model to gain further understanding of the specific consequences of mutations in different genes on the level of Wnt signalling. This particular study lacks this information and so provides a much broader insight into the just-right model which may be useful in providing an initial gauge of the just-right model.

An additional caveat of this study is that I did not determine the clonality of the mutations in the samples. Consequently, while different combinations of mutations have been observed within the same sample, it is not definitively established that they are present in the same cell. To determine the clonality of a mutation, the variant allele frequency (VAF) is calculated and adjusted for tumour purity. A lower-than-expected VAF suggests that the mutation is subclonal. Reassuringly, Cross *et al.* (2018)

reported stabilising selection in CRCs, which exhibit lower levels of ITH compared to CRAs. This finding suggests that the mutations identified in this study are likely to be clonal.

It could also be argued that these combinations of mutations occur because of mutational processes in the colon, rather than as a result of selection. Utilising the same samples, Brunet *et al.* (*) showed no significant difference between the mutation signatures in the normal tissue vs the tumour samples, indicating that the results found are indeed due to selection rather than mutational process.

While this study focuses on primary, treatment-naïve tumours, it may be beneficial to investigate mutational relationships in Wnt pathway genes in patient samples to identify patterns of co-occurrence and mutual exclusivity that could be linked to treatment resistance. Additionally, examining various mutation combinations and their associations with changes in prognosis or survival could offer important insight into how the just-right model might influence this. Exploring the influence of these major-mini-driver relationships in mouse models would be of valuable insight to further elucidate the influence of the just-right model on tumour evolution in different locations in the intestine.

Altogether this chapter provides supportive evidence for the just-right model in revealing mutual exclusivity in between mutations in Wnt pathway major-driver genes and patterns of co-occurrence between Wnt pathway major-driver genes and potential mini-driver genes. *AMER1* was identified as a strong candidate mini-driver gene through its association with an increased number of 20AARs in *APC*, suggesting mutations in *AMER1* could adjust the level of Wnt signalling to account for a sub-optimal number of 20AARs. Although this work also suggests that the just-right model may act differently depending on tumour location and tumour context e.g. MSI status.

Chapter 4

An exploration of *TCF7L2* as a CRC risk gene

4.1 Background

To further elucidate the just-right model, it is essential to differentiate between functional and non-functional intra-tumour heterogeneity (ITH). According to Castro-Giner, Ratcliffe and Tomlinson (2015), the current dichotomous classification of mutations into either driver or passenger lacks accuracy. They propose that there are mutations of low frequency and small overall effect, unlikely to cause tumorigenesis. These mini-driver mutations are not critical for tumour development; tumours can develop without a mutation in these genes. They should be involved in cellular processes such as gene expression regulation, allowing for the adaptation and modulation of expression of other genes, and should display parallel or convergent evolution, suggesting that their role is in tumour optimisation rather than initiation (Castro-Giner, Ratcliffe and Tomlinson, 2015). Through the identification of mini-driver mutations and their integration into the just-right model, we can differentiate between functional and non-functional ITH and in doing so, better grasp how subtle genetic changes influence tumour progression and therapeutic responses.

TCF7L2 is a transcription factor in the Wnt pathway, it therefore plays a role in the modulation of expression of other genes in the Wnt pathway. Mutations in *TCF7L2* are present in approximately 10% of CRCs, tumours can develop without *TCF7L2* mutations and are non-essential for maintenance of CRC cell lines (Fröhlich, Rose and Hecht, 2023). Interestingly, Chapter 2 shows the presence of *TCF7L2* mutations on both trunks and branches of CRAs, suggesting that they could play a role in tumour initiation and optimisation. Chapter 3 shows that mutations in *TCF7L2* do not display a mutually exclusive relationship with other major-drivers in the Wnt pathway (*APC*, *RNF43* and *CTNNB1*) (Kim and Jeong, 2019; Fröhlich, Rose and Hecht, 2023). In fact, they show a pattern of co-occurrence with mutations in *APC* and *CTNNB1*. According to the just-right model, this suggests that mutations in *TCF7L2* could provide a moderate overall effect on Wnt signalling, able to modulate expression of other major-drivers but without the capacity of driving tumorigenesis in isolation. Altogether, these results suggest that *TCF7L2* could act as a Wnt pathway mini-driver gene. I therefore sort to further investigate the role of *TCF7L2* as a CRC risk gene to provide further support of its potential contribution to cancer development, and further clarify evidence for and against its categorisation as a mini-driver gene in CRC.

4.1.1 GWAS a brief history

The study of single-nucleotide polymorphisms (SNPs), or inherited variations in the DNA sequence at the single nucleotide level, has proven invaluable for the understanding of genetic susceptibility to many complex phenotypic traits and diseases. As highlighted in chapter 1.1.3, family-based, genetic linkage studies have outlined some highly penetrative variants or mutations in several genes that cause familial CRC syndromes such as Lynch syndrome and FAP. However, the development of sporadic CRC, constituting a much larger proportion of cases, is complex and

influenced by a variety of genetic and environmental factors. It is therefore more difficult to identify variants or mutations linked to sporadic CRC predisposition (Huyghe *et al.*, 2019b).

In the early 2000's, The International HapMap project played a prominent role in the identification of many SNPs and genetic variations across multiple populations worldwide (International HapMap Consortium, 2003). In a review of GWAS, Hirschhorn and Daly (2005) discussed the potential role of GWAS in the identification of statistical associations between SNPs and various traits and diseases, highlighting a theoretical shift from candidate gene studies to hypothesis-free genetic association studies. GWAS, unlike candidate-gene studies, would be uninhibited by a lack of data and knowledge of the genetic and molecular function of target genes (Hirschhorn and Daly, 2005).

Subsequent GWAS were therefore able to utilise sequencing technologies to genotype an array of previously identified SNPs in thousands of individuals, in order to identify any overrepresentations in those with a particular trait or disease compared with unaffected controls (Gallagher and Chen-Plotkin, 2018). In 2007, Tomlinson *et al* (2007) demonstrated the ability of GWAS to identify novel SNP associations with CRC. This study involved the genotyping of 550,163 SNPs in 930 individuals with familial colorectal neoplasia (620 with CRC and 310 with a high-risk adenoma) and 960 unaffected individuals. As a result, rs6983267, at the 8q24.21 region, was found to be significantly associated with the development of CRC (Tomlinson *et al.*, 2007). Subsequently, a multiphase analysis observed a previously unidentified significant association of rs3802842 on 11q23 with CRC risk (Tenesa *et al.*, 2009).

Recently, GWAS meta-analyses have utilised and compiled previous GWAS data to provide an even more comprehensive analysis of CRC predisposition variants. Complemented with transcriptome and methylome-wide (TWAS & MWAS) association analyses, Fernandez-Rozadilla *et al* (2023) identified 205 genome-wide significant SNPs, 155 of these in agreement with previous findings, 37 SNPs at novel loci and 13 independent risk SNPs in conditional analysis, including rs4616575 at *ZNRF3* (Fernandez-Rozadilla *et al.*, 2023a). According to the US National Human Genome Research Institute Catalog of human genome-wide association studies there are 6545 publications and 541660 top SNP associations to date (Burdett, 2017).

4.1.2 GWAS: The process

GWAS utilise statistical methods to identify SNPs that are significantly associated with specific traits and disease. Genotyping via SNP arrays or WGS is required to assess genetic variation within thousands of affected and unaffected individuals within a clearly defined population. A substantial number of individuals is required to achieve statistical power. A number of quality control (QC) steps are also required to reduce the likelihood of false positive statistical associations. SNPs with low

genotype calls may indicate unreliable or low-quality genotype data, potentially introducing bias and leading to false associations. Therefore, SNPs with low genotype calls and individuals with high rates of missing genotypes are excluded from the dataset. Another QC step involves filtering the data to only include those SNPs with a minor allele frequency (MAF) above a certain threshold, for example $> 5\%$. SNPs with a low MAF are infrequent and prone to genotyping errors. This in turn reduces the statistical power of any association found between a low frequency SNP and a particular phenotype and increases the likelihood of a false positive result. In addition, it is recommended to exclude individuals with high or low rates of heterogeneity in comparison to the rest of the population e.g. $> 65\%$. This could be indicative of sample contamination or genotyping errors (Marees *et al.*, 2018; Fernandez-Rozadilla *et al.*, 2023a). Appropriate statistical methods can then be utilised to make associations between SNPs and a target phenotype or disease. A p-value threshold of 5×10^{-8} is often utilised for a SNP to be deemed genome-wide significant (Fernandez-Rozadilla *et al.*, 2023a).

4.1.3 Conditional analysis

Whilst GWAS have aided our understanding of aetiology, for many complex traits, GWAS is not fully able to elucidate the inheritance of every complex trait. For example, human height has been linked to 40 different loci, which only explains approximately 5% of human phenotypic variance (Manolio *et al.*, 2009). Linkage disequilibrium (LD) is an inherent weakness of GWAS. LD describes the non-random association of alleles at different loci that may be related through proximity to each other on the same chromosomal segment. LD is harnessed during the GWAS process to save both time and expense. Instead of analysing each individual SNP in a person's genome, tagging SNPs (tagSNPs) can be utilised. A tagSNP is a representation of a group of SNPs or haplotype block. Through the utilisation of tagSNPs, a large proportion of variation within an individual's genome can be genotyped without the requirement of analysing the entire genome. However, rather than having a biological function themselves, the most strongly associated genome-wide significant variants, often referred to as "leading SNPs", are likely to be in LD with the causal variant (Wall and Pritchard, 2003; Spain and Barrett, 2015a). LD can be measured via an r^2 or D' metric ranging from zero to one. A score of zero indicates that the SNPs were inherited independently of one another and conversely a score of one indicates perfect genetic linkage (Huyghe *et al.*, 2019). Yang *et al.* (2011) therefore developed a software tool named Genome-wide complex trait analysis (GCTA) to address the issue of missing heritability and therefore LD. GCTA re-analyses GWAS data using mixed linear models and genetic relationship matrices to estimate heritability, taking into account all SNPs simultaneously, rather than focussing on single SNP associations (Yang *et al.*, 2013a). By performing conditional analysis in this way, it is possible to determine if there are additional genetic variants influencing the phenotype independently of the "lead" SNP. It can also therefore validate the independent action of the "lead" SNP and reduce the number of independent signals at a particular locus (Spain and Barrett, 2015a).

4.1.4 Post-GWAS

As stated above, a positive attribute of GWAS is its lack of restriction to previously studied target genes, of which the biological and molecular roles are already known. However, in the case of a large number of studies, many genome-wide significant SNPs have been positioned outside of coding regions within the genome. The function of these SNPs is elusive, highlighting the necessity of further studies to understand the biological mechanism of action of non-coding SNPs.

Expression quantitative trait locus (eQTL) studies are utilised to assess the functional consequences of a GWAS variant by looking into its potential effect on gene expression. A variant is described as an eQTL if it is associated with altered gene expression. By integrating GWAS and eQTL studies it is possible to help identify which genes may be involved in disease risk.

4.1.5 *TCF7L2* in CRC GWAS

As stated in chapter 1, *TCF7L2* acts as a transcription factor in the Wnt pathway. *TCF7L2* mutations are present in approximately 10% of CRCs. In 2014, Zhang et al. (2014) identified a genome-wide significant SNP (rs11196172) at the *TCF7L2* locus in an East-Asian cohort. Rs11196172 is located in intron 4 of the *TCF7L2* gene. They found an association between the risk allele at rs11196172 and increased expression of *TCF7L2* in colon tumour tissue using TCGA (Tumour Cancer Genome Atlas) data. Notably, Zhang *et al.*, (2014) also found the risk allele at rs11196172 to be associated with increased Vehicle transport through interaction with T-SNAREs 1A (*VTI1A*) expression, a gene located approximately 131kb upstream of *TCF7L2*.

VTI1A functions to mediate the fusion of intracellular vesicles within the golgi complex. The role of *VTI1A* is important insulin-stimulated glucose transport (Tang *et al.*, 2022). Notably, rs7903146 at the *TCF7L2* locus was found to have a significant association with type 2 diabetes mellitus. Mitroi et al., (2022) found that the risk allele at rs7903146 was associated with type 2 diabetes and CRC (Mitroi *et al.*, 2022).

Using a European cohort, a GWAS by Law et al. (2019) identified 31 new risk loci for CRC. They evidenced genome-wide significance at both the rs11196172 and rs11196171 at the 10q25.2 locus. This locus incorporates both *TCF7L2* and *VTI1A*. In this instance, they identified the lead SNP to be rs12255141, located in an active enhancer region of *VTI1A*. No evidence for any SNPs, at the 10q25.2 locus, as eQTLs was shown in this study when exploring expression data from normal colonic mucosa (INTERMPHEN) and Genotype-Tissue expression (GTEx) data from the transverse colon.

Interestingly, even though no SNPs were identified as eQTLs by Law et al., (2019), through the *in-situ* promoter capture Hi-C(Chi-C) of LoVo and HT29 cells, they identified a looping interaction

between an active enhancer region located in *VTIIA*, and *TCF7L2*. They concluded that *TCF7L2* was therefore implicated as the target gene of rs12255141.

Fernandez-Rozadilla *et al.*, (2023) performed a GWAS utilising European and East Asian ancestry. They also identified rs11196172 as a genome-wide significant SNP. In this study they identified the strongest signals across the entire genome. This chapter aims to zoom into *TCF7L2* and use a more targeted approach, with increased statistical power to detect weaker associations that may not be apparent when exploring the entire genome.

4.2 Chapter aims

Altogether, SNPs at the 10q25.2 locus at both *VTIIA* loci and *TCF7L2* loci are significantly associated with CRC. It is currently unclear whether these SNPs are independently associated with CRC due to potentially being in LD with one another. This chapter aims to firstly carry out a conditional analysis using meta-analysis data from the study by Fernandez-Rosadilla *et al.*, (2023) to determine if there is more than one independent signal at the 10q25.2 region.

Interestingly, *VTIIA* locus variants have been suggested to be correlated with altered *TCF7L2* expression. This chapter aims to further assess whether there is an association between any genome-wide significant SNPs in the region and altered expression of either *TCF7L2* or *VTIIA*. This will provide further clarity as to whether *TCF7L2* and or *VTIIA* are candidate CRC predisposition genes.

4.3 Methods

Methods used for the analysis of *TCF7L2* (chromosome 10) as a CRC predisposition gene are modelled upon the conditional analysis carried out by Joe Ward as part of his PhD project, looking at the 4q24 locus of chromosome 4 and its relation to CRC.

4.3.1 Conditional analysis

Meta-analysis data from a previous CRC GWAS study, which included 78,473 cases and 107,143 controls, across 17 cohorts of European descent were utilised to further analyse the role of *TCF7L2* SNPs in CRC predisposition (Fernandez-Rozadilla *et al.*, 2023a). Each cohort is detailed in Table 4.1 and Table 4.2.

1Mb upstream of the *TCF7L2* transcription start site and 1Mb downstream of the *TCF7L2* transcription termination site were selected as limits to define a finalised region on which to perform these analyses (10q25.2). The final selected region was chr10:11371009-115927436 (hg19). All variants with an I^2 above 75 % or $p_{(\text{heterogeneity})} < 0.01$ were excluded from the analysis. Conditional analysis was conducted using Genome-Wide Complex Trait Analysis (GCTA, v1.91.4beta) (Yang *et al.*, 2012), leveraging LD reference data to compute conditional p-values. LD reference data was derived from European individuals within the 1,000 Genomes Project data and compiled in a bedfile using PLINK (v1.90) (Purcell *et al.*, 2007). For a SNP to be considered genome-wide significant, a $p_{(\text{conditional})}$ threshold was set at 5×10^{-7} . LocusZoom plots were generated using LocusZoom (Pruim *et al.*, 2011).

Table 4.1 A summary of all cohorts utilised in the European GWAS Meta-Analysis. Adapted from a table by Joe Ward as part of a GWAS PhD project, this table shows the numbers of cases and controls in each study along with the total number of patients in each cohort and their countries of origin. Altogether, these 17 cohorts were utilised in the European GWAS meta-analysis. Abbreviations: Combination therapy with or without cetuximab as first-line therapy in treating patients with metastatic colorectal cancer (COIN); Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO); Colorectal Cancer Transdisciplinary Study, Europeans (CORECT-EUR); Colorectal Cancer Study of Austria (CORS-1); Darmkrebs: Chancen der Verhütung durch Screening (DACHS-4); Finnish GWAS (FIN2); National Study of Colorectal Cancer Genetics (NSCCG); Short-Course Oncology Treatment (SCOT); Study of Colorectal Cancer in Scotland (cases) and Generation Scotland (controls) (SCOT-GS); Study of Colorectal Cancer in Scotland (cases) and Lothian Birth Cohort (controls) (SCOT-LBC); Colorectal Tumour Gene Identity Study (UK-1); VICTOR & QUASAR2 Trials (cases) and 1958 Birth Cohort (controls) (VQ58); United Kingdom (UK); United States of America (USA).

Cohort	Country/countries of origin	Case No.	Control No.	Total
COIN	UK	1950	2162	4112
GECCO OmniExpress Exome	Germany/The Netherlands/USA	4439	4115	8554
CORECT-EUR	Australia/Canada/Finland/Germany/Israel/Spain/Sweden/UK/USA	19948	12124	32072
CORS-1	Austria	1460	774	2234
Croatia	Croatia	689	441	1130
GECCO Oncoarray Custom iSelect	Austria/Czech Republic/Spain/Sweden/UK/USA	11835	11856	23691
DACHS-4	Germany	1028	661	1689
FIN2	Finland	1760	14132	15892
GECCO Stage I Pooled	Australia/Canada/France/Germany/USA	11895	14659	26554
NSCCG	UK	6596	7205	13801
SCOT	UK	2910	4095	7005
Scotland Phase I	Scotland	932	943	1875
SOCCS-GS	Scotland	4551	8804	13355
SOCCS-LBC	Scotland	996	1297	2293
UK-1	UK	890	900	1790
UK Biobank	UK	4800	20289	25089
VQ58	UK	1794	2686	4480

4.3.2 eQTL analysis

To determine if both *TCF7L2* and *VTG1A* show evidence of being eQTLs, and thus whether the risk alleles at different SNP loci identified through conditional analysis influence gene expression, *TCF7L2* and *VTG1A* SNP genotype data were analysed alongside Trimmed Mean of M-Values (TMMs) gene expression data from 107 individuals of European descent, as provided by the INTERMPHEN study (Fernandez-Rozadilla et al., 2023). This data includes expression levels of both *TCF7L2* and *VTG1A* in the proximal colon, distal colon and rectum. Using this data, the Jonckheere-Terpsta (JT) test was employed to determine if the presence of risk alleles at specific SNP loci significantly affects gene expression. The JT test was carried out using R-package Clinfun (Seshan and Karissa, 2023).

4.3 Results

4.3.1 Conditional Analysis

Meta-analysis data provided by Fernandez-Rozadilla *et al.* (2023) was utilised in a conditional analysis to further explore *TCF7L2* SNPs and their role in CRC predisposition. This data included 78,472 cases and 107,143 controls across 17 cohorts of European descent and 21,731 cases and 47,444 controls from 14 cohorts of East-Asian descent.

Within the selected region of chr10:11371009-115927436 (hg19), a total of 8,127 SNPs were detected in the European cohort and 7,607 in the East-Asian cohort. In the European cohort, multiple SNPs identified in the region of *TCF7L2* had no rsID. These SNPs are rare, have not been validated and may be less well-characterised. To ensure data quality and reliability, any SNPs without an rsID were excluded from further analysis ($n = 273$).

In the European cohort, 53 SNPs reached genome-wide significance ($p < 5 \times 10^{-8}$) at the 10q25.2 locus. Since a region including 1Mb both upstream and downstream of *TCF7L2* was taken into consideration, the lead SNP for the entire region was identified as rs56188434 ($p = 3.46 \times 10^{-16}$), located at the *VTI1A* locus (**Fig. 4.1**).

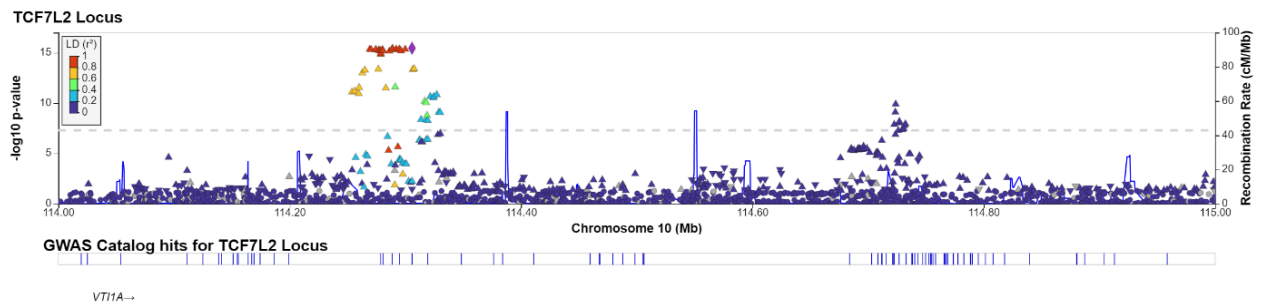


Figure 4.1 Genome-wide significant SNPs identified at the 10q25.2 locus, present in European, CRC meta-analysis data. Data from a CRC GWAS meta-analysis by Fernandez-Rozadilla *et al.* (2023) was utilised to explore SNPs from 1Mb upstream to 1Mb downstream of *TCF7L2* (chr10:11371009-115927436). This identified 53 genome-wide significant SNPs, with the lead SNP identified as rs56188434 at the *VTI1A* locus. The x-axis represents the genomic position on chromosome 10 in megabases (Mb), while the left y-axis shows the $-\log_{10}$ (p-value) of the association between each SNP and CRC. The right y-axis indicates the recombination rate in centiMorgans per megabase (cM/Mb). Each point represents a SNP with the colour indicating the degree of LD (r^2) with the lead SNP rs56188434. The horizontal dashed line marks the genome-wide significance threshold ($p < 5 \times 10^{-8}$).

When looking at the *TCF7L2* locus specifically, the lead SNP was identified as rs2296782 with a p-value of 1.25×10^{-10} (**Fig. 4.2**)

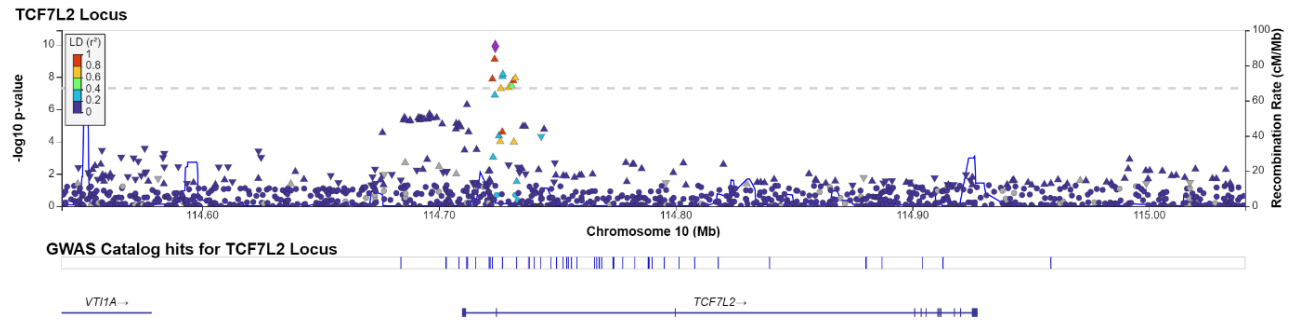


Figure 4.2 SNPs identified at the *TCF7L2* locus, present in European, CRC meta-analysis data. Rs2296782 was identified as the lead SNP at the *TCF7L2* locus.

Following the identification of 53 genome-wide significant SNPs in the 10q25.2 region, a conditional analysis was conducted to determine how many independent GWAS signals are present by accounting for LD between these 53 genome-wide significant SNPs. To do this, SNPs were conditioned on the lead SNP in the whole region, rs56188434 (**Table 4.2**).

Table 4.2 A conditional analysis of genome-wide significant SNPs at the 10q25.2 locus, conditioning on lead SNP rs56188434. This table presents the 53 SNPs identified at the 10q25.2 locus, that reached genome-wide significance ($p < 5 \times 10^{-8}$). The SNP ID, chromosome number and position (hg19) are listed. For each SNP, the p-values from the initial GWAS and the conditional analysis are provided along with the r^2 indicating linkage disequilibrium with the leading *VTG1A* locus SNP, rs56188434. SNPs that did not produce a p-value in the conditional analysis are listed as NA, these SNPs are in strong LD with the lead SNP.

SNP ID	Chromosome #	Position (hg19)	p(GWAS)	p(Conditional)	r^2 with rs56188434
rs80109592	10	114,253,877	7.85965e-12	0.665568	0.568
rs142307889	10	114,256,281	6.34676e-12	0.724385	0.583
rs77705687	10	114,259,714	1.12984e-11	0.844523	0.568
rs58936340	10	114,260,084	2.82935e-12	0.596814	0.557
rs111936498	10	114,262,773	1.0027e-13	0.811955	0.641
rs77802795	10	114,265,071	5.17734e-14	0.689559	0.628
rs12263620	10	114,269,239	4.36824e-16	NA	0.838
rs10509964	10	114,270,474	4.76585e-16	NA	0.865
rs78142102	10	114,274,269	5.12004e-16	0.722676	* is not biallelic
rs34139797	10	114,274,559	4.60052e-16	NA	0.838
rs76880073	10	114,276,661	4.20204e-14	0.779379	0.838
rs7894915	10	114,277,039	5.67207e-16	NA	0.836
rs7895362	10	114,277,173	5.14111e-16	NA	0.852
rs10082356	10	114,278,181	6.43511e-16	0.859217	0.836
rs4554811	10	114,278,734	1.40182e-15	0.809828	0.836
rs4414150	10	114,278,740	1.0357e-15	0.778336	0.836
rs4554812	10	114,278,871	5.80084e-16	0.850024	0.852
rs17129834	10	114,280,318	4.82766e-16	NA	0.838

rs12241008	10	114,280,702	6.08734e-16	NA	0.838
rs17129837	10	114,283,116	3.17683e-12	0.937593	0.555
rs12240720	10	114,286,188	6.81344e-16	NA	0.867
rs12246635	10	114,288,619	3.88498e-16	NA	0.897
rs11195986	10	114,289,017	3.72408e-16	NA	0.896
rs4434930	10	114,291,241	2.48317e-12	Not in CA	0.522
rs17129851	10	114,291,787	4.9925e-16	NA	0.896
rs4439449	10	114,292,303	4.58265e-16	NA	0.896
rs12255141	10	114,294,892	4.1189e-16	NA	0.897
rs73365049	10	114,296,308	6.42397e-16	NA	0.882
rs58575880	10	114,299,884	4.41248e-16	NA	0.913
rs56188434	10	114,305,761	3.46048e-16	*NA	1
rs80207046	10	114,306,441	4.6644e-14	0.538183	0.629
rs74236144	10	114,307,546	3.80437e-14	0.640613	0.656
rs11195993	10	114,313,556	4.28458e-9	0.225245	0.457
rs2028783	10	114,316,652	6.57216e-11	0.125361	0.531
rs12266418	10	114,318,004	9.04299e-11	0.159785	0.543
rs2242205	10	114,318,065	4.63633e-9	0.227135	0.444
rs11815469	10	114,318,714	1.67691e-9	0.349698	0.529
rs77373153	10	114,319,358	5.27763e-9	0.270279	0.457
rs17129862	10	114,321,528	2.19459e-11	0.0662511	0.502
rs7100296	10	114,324,035	2.95964e-11	0.0539785	0.477
rs12254971	10	114,324,710	2.23816e-11	0.0440146	0.477
rs12259387	10	114,327,211	1.44353e-11	0.0264545	0.477

rs11195996	10	114,328,887	8.17128e-10	0.0285201	0.407
rs12250914	10	114,330,160	8.69884e-10	0.0488135	0.403
rs11196170	10	114,722,621	1.30413e-8	7.21E-10	0.004
rs2296784	10	114,723,559	7.84112e-10	2.39E-11	0.004
rs2296782	10	114,723,857	1.25498e-10	2.50E-11	0.001
rs11196172	10	114,726,843	9.18031e-9	1.22E-06	0.001
rs11196173	10	114,727,067	6.53676e-9	7.96E-07	0.001
rs61875109	10	114,729,563	4.43764e-8	2.32E-08	0.003
rs12573128	10	114,730,797	3.45572e-8	3.25E-06	0.001
rs61875110	10	114,731,483	1.68998e-8	1.64E-08	0
rs7897438	10	114,732,307	1.15008e-8	5.38E-09	0.003

When conditioning on the lead SNP rs56188434, 44 of 53 SNPs became non-significant, likely due to being in LD with the lead SNP. Conversely, nine SNPs remained significant ($p < 5 \times 10^{-7}$) (**Fig 4.3**). This suggests that these nine SNPs are likely not in LD with the lead SNP and represent at least one more independent signal. Interestingly, unlike rs56188434 which is positioned in the *VTI1A* locus, these nine significant SNPs are positioned in the *TCF7L2* locus.

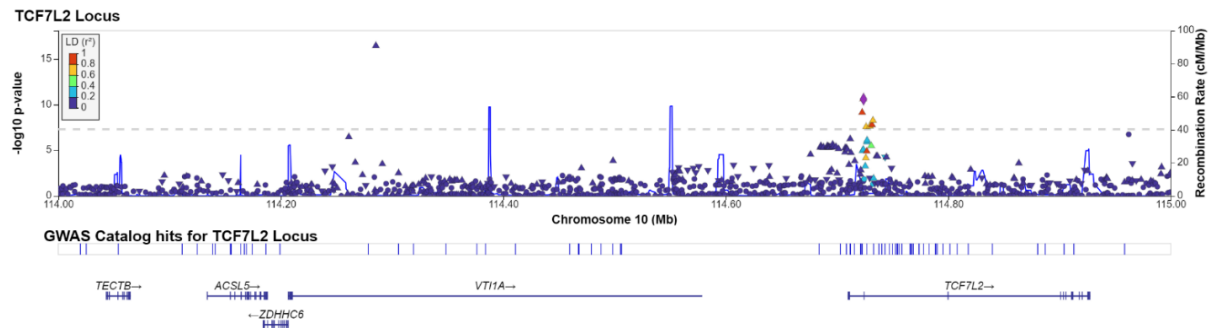


Figure 4.3 Genome-wide significant SNPs at the 10q25.2 locus when conditioned on rs56188434, using European, CRC meta-analysis data. A conditional analysis, utilising data from a CRC, GWAS by Fernandez-Rozadilla *et al.* (2023) identified nine significant SNPs when conditioned on the lead SNP, rs56188434.

To establish if there are additional independent signals within the group of nine significant SNPs, an additional conditional analysis was carried out conditioning on both lead SNP rs56188434 and the lead SNP at the *TCF7L2* locus which itself remained significant when conditioning on just rs56188434 (**Table 4.3**).

Table 4.3 Conditional analysis of genome-wide significant SNPs in the region around *TCF7L2* region, conditioning on both rs2996782 and rs56188434. This table presents the 53 SNPs identified in the *TCF7L2* region that reached genome-wide significance ($p < 5 \times 10^{-8}$). The SNP ID, chromosome number and position (hg19) are listed. For each SNP, the p-values from the initial GWAS and the conditional analysis are provided along with the r^2 indicating linkage disequilibrium with the leading *TCF7L2* locus SNP, rs2996782 and the leading *VTG1A* locus SNP rs56188434. SNPs that did not produce a p-value in the conditional analysis are listed as NA, these SNPs are in strong LD with the lead SNPs.

SNP ID	Chr #	Position (hg19)	p(GWAS)	p(conditional)	r^2 with rs2996782	r^2 with rs56188434
rs80109592	10	114,253,877	7.85965e-12	0.694918	0	0.568
rs142307889	10	114,256,281	6.34676e-12	0.797725	0	0.583
rs77705687	10	114,259,714	1.12984e-11	0.898351	0	0.568
rs58936340	10	114,260,084	2.82935e-12	0.699572	0	0.557
rs111936498	10	114,262,773	1.0027e-13	0.85492	0	0.641
rs77802795	10	114,265,071	5.17734e-14	0.75087	0	0.628
rs12263620	10	114,269,239	4.36824e-16	NA	0.001	0.838
rs10509964	10	114,270,474	4.76585e-16	NA	0.001	0.865
rs78142102	10	114,274,269	5.12004e-16	0.617739	*is not biallelic	* is not biallelic
rs34139797	10	114,274,559	4.60052e-16	NA	0.001	0.838
rs76880073	10	114,276,661	4.20204e-14	0.821922	0	0.838
rs7894915	10	114,277,039	5.67207e-16	NA	0.001	0.836
rs7895362	10	114,277,173	5.14111e-16	NA	0.001	0.852
rs10082356	10	114,278,181	6.43511e-16	0.784725	0.001	0.836
rs4554811	10	114,278,734	1.40182e-15	0.718527	0.001	0.836
rs4414150	10	114,278,740	1.0357e-15	0.688219	0.001	0.836
rs4554812	10	114,278,871	5.80084e-16	0.775732	0.001	0.852

rs17129834	10	114,280,318	4.82766e-16	NA	0.001	0.838
rs12241008	10	114,280,702	6.08734e-16	NA	0.001	0.838
rs17129837	10	114,283,116	3.17683e-12	0.967355	0	0.555
rs12240720	10	114,286,188	6.81344e-16	NA	0.001	0.867
rs12246635	10	114,288,619	3.88498e-16	NA	0.001	0.897
rs11195986	10	114,289,017	3.72408e-16	NA	0.001	0.896
rs4434930	10	114,291,241	2.48317e-12	NA	0.002	0.522
rs17129851	10	114,291,787	4.9925e-16	NA	0.001	0.896
rs4439449	10	114,292,303	4.58265e-16	NA	0.001	0.896
rs12255141	10	114,294,892	4.1189e-16	NA	0.001	0.897
rs73365049	10	114,296,308	6.42397e-16	NA	0.001	0.882
rs58575880	10	114,299,884	4.41248e-16	NA	0.001	0.913
rs56188434	10	114,305,761	3.46048e-16	*NA	0.001	1
rs80207046	10	114,306,441	4.6644e-14	0.592183	0	0.629
rs74236144	10	114,307,546	3.80437e-14	0.659544	0	0.656
rs11195993	10	114,313,556	4.28458e-9	0.114326	0.007	0.457
rs2028783	10	114,316,652	6.57216e-11	0.0596347	0.006	0.531
rs12266418	10	114,318,004	9.04299e-11	0.0741079	0.008	0.543
rs2242205	10	114,318,065	4.63633e-9	0.0835934	0.01	0.444
rs11815469	10	114,318,714	1.67691e-9	0.160427	0.01	0.529
rs77373153	10	114,319,358	5.27763e-9	0.110261	0.009	0.457
rs17129862	10	114,321,528	2.19459e-11	0.0201871	0.007	0.502
rs7100296	10	114,324,035	2.95964e-11	0.0158351	0.007	0.477
rs12254971	10	114,324,710	2.23816e-11	0.0124316	0.007	0.477

rs12259387	10	114,327,211	1.44353e-11	0.00682633	0.006	0.477
rs11195996	10	114,328,887	8.17128e-10	0.00864557	0.872	0.407
rs12250914	10	114,330,160	8.69884e-10	0.0153947	0.006	0.403
rs11196170	10	114,722,621	1.30413e-8	0.928098	0.872	0.004
rs2296784	10	114,723,559	7.84112e-10	0.668567	0.872	0.004
rs2296782	10	114,723,857	1.25498e-10	*NA	1	0.001
rs11196172	10	114,726,843	9.18031e-9	0.631699	0.372	0.001
rs11196173	10	114,727,067	6.53676e-9	0.581474	0.377	0.001
rs61875109	10	114,729,563	4.43764e-8	0.8429	0.743	0.003
rs12573128	10	114,730,797	3.45572e-8	0.866741	0.391	0.001
rs61875110	10	114,731,483	1.68998e-8	0.612327	0.846	0
rs7897438	10	114y,732,307	1.15008e-8	0.99649	0.761	0.003

The results of this conditional analysis showed that none of the 53 previously genome-wide significant SNPs reached significance. This is likely because all of these SNPs are in strong LD with either rs56188434 or rs2296782. This shows that there are likely two independent signals in the region of *TCF7L2*, one at the *VTI1A* locus and another at the *TCF7L2* locus (**Fig. 4.4**).

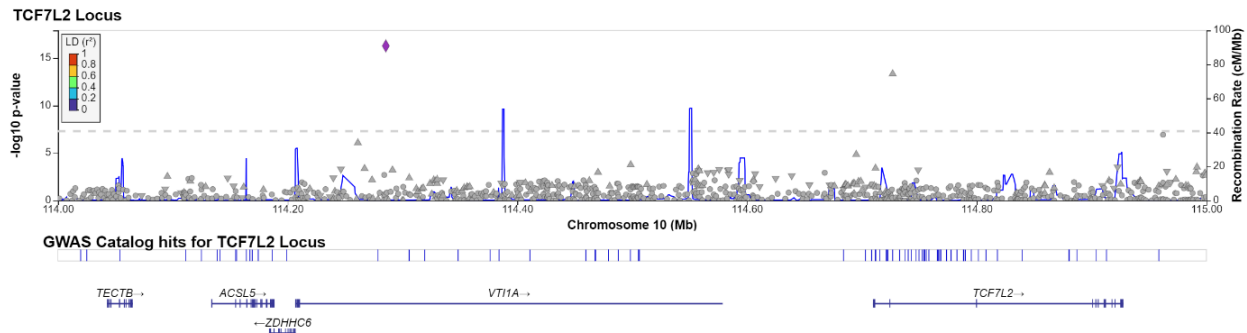


Figure 4.4 Genome-wide significant SNPs when conditioned on both rs56188434 and rs2296782. A conditional analysis, utilising data from a CRC, GWAS by Fernandez-Rozadilla *et al.* (2023) identified no significant SNPs when all genome-wide significant (p_{GWAS}) SNPs were conditioned on both the lead SNP at the *TCF7L2* locus (rs2296782) and lead SNP at the *VTI1A* locus (rs56188434).

4.3.2 eQTL analysis

Using genotype-expression data from samples of 107 individuals of European descent, enrolled in the INTERMPHEN study, both *TCF7L2* and *VTI1A* expression levels in the proximal colon, distal colon and rectum were analysed to determine if the presence of the effect allele at each of the lead SNPs at both the *TCF7L2* and *VTI1A* locus was associated with a change in *TCF7L2* or *VTI1A* expression. To examine trends across the three different genotypes (e.g. AA, AG, GG), multiple JT tests were employed.

Initially, rather than focussing on lead SNPs alone, each of the 53 genome-wide significant SNPs were tested for an association for changes in gene expression. However, these results were heavily affected by LD and therefore multiple corrections testing. Due to LD, many of the SNPs had identical genotypes, leading to replicated statistical results, as the risk allele was present to the same extent regardless of the SNP. Due to LD, and therefore the production of replicated results, the main results presented in this chapter are limited to the associations between gene expression and genotype for only lead SNPs at both the *TCF7L2* locus and *VTI1A* locus. Although caveats of this will also be discussed. An LD matrix for visualisation of LD between all 53 genome-wide SNPs can be seen in Figure 4.5.

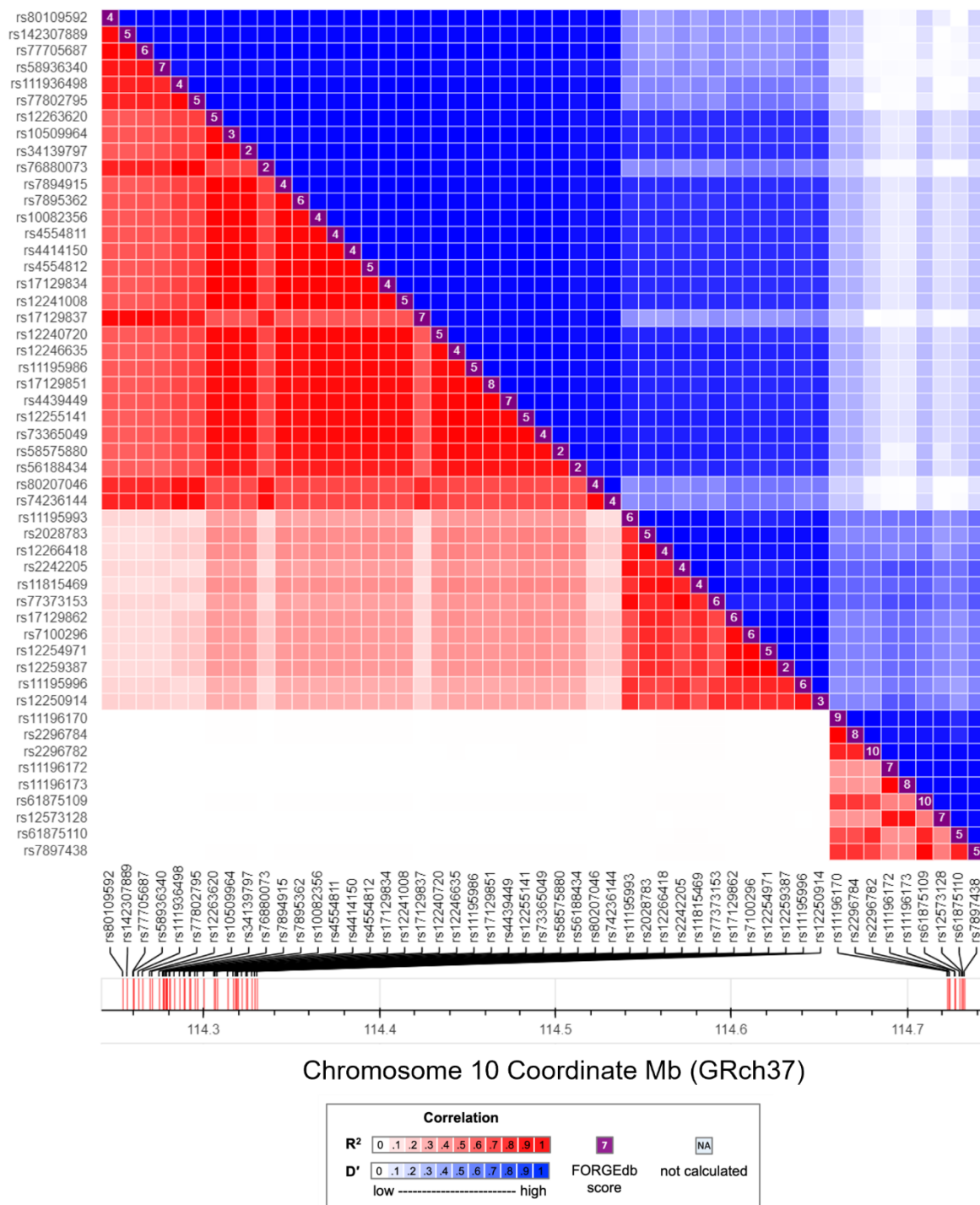


Figure 4.5 LD matrix for 53 genome-wide significant SNPs at the 10q25.2 locus. Made using LDlink. Red represents the r^2 LD score. Blue represents the D' LD score. The FORGEdb score is an evaluation of the functional important of genetic variants.

Firstly, both the risk allele at the lead SNP at the *TCF7L2* locus, rs2296782, and risk allele at the *VTG1A* locus, rs56188434, were assessed for an association with *TCF7L2* expression in the proximal colon, the distal colon and the rectum (**Table 4.4**).

Table 4.4 The effect of risk alleles at lead SNPs at both the *TCF7L2* locus (rs2296782) and at the *VTIIA* locus (rs56188434) on *TCF7L2* expression in the proximal colon, distal colon and rectum. Shown is the effect allele for each SNP, along with their effect sizes (β) reported in the GWAS. Additionally, standard errors of the effect sizes are also presented along with raw and corrected Jonckheere-Terpstra (JT) p-values, which assess the trend in gene expression across different genotypes.

SNP	Effect allele	$\beta_{(GWAS)}$	$SE_{(GWAS)}$	$p_{(GWAS)}$	$p_{(JT)}$	$p_{(JT)}$	$p_{(JT)}$
					proximal colon	distal colon	rectum
rs2296782	G	0.0555383	0.00863398	1.25498e-10	0.764677	0.845009	0.006437
rs56188434	A	0.102209	0.0125316	3.46048e-16	0.190543	0.684382	0.991074

Interestingly, presence of the risk allele at the lead SNP at the *TCF7L2* locus (rs2296782) was not found to have an association with *TCF7L2* expression in the proximal or distal colon, as found by Law *et al.*, (2019). Conversely, the risk allele at rs2296782 was associated with increased *TCF7L2* expression in the rectum. This data suggests that either rs2296782 is an eQTL associated with increased *TCF7L2* expression in the rectum, or is in LD with the eQTL. The risk allele at rs56188434, the lead SNP at the *VTIIA* locus, was not associated with a change in *TCF7L2* expression in the proximal colon, distal colon or rectum (**Fig. 4.6**).

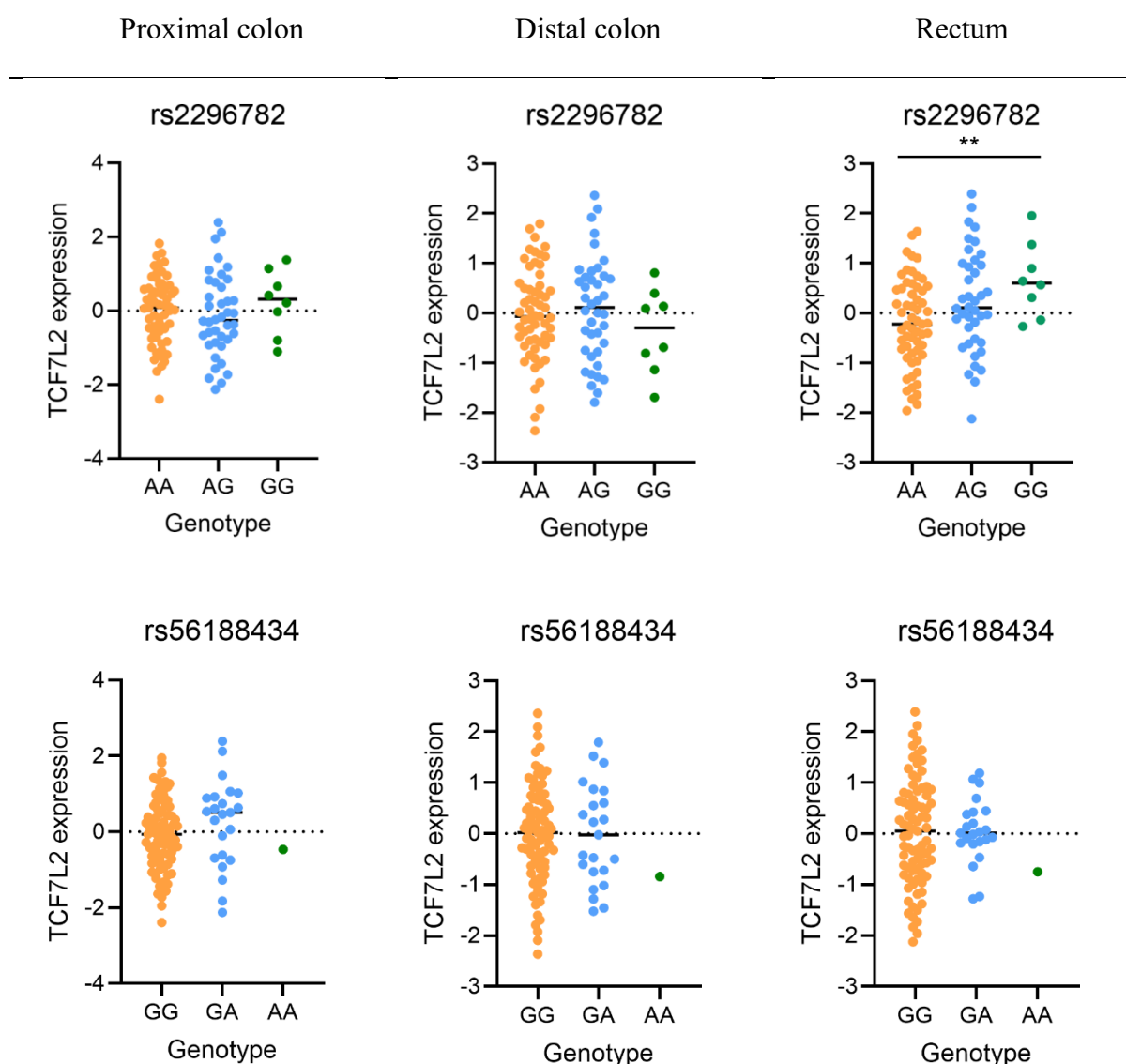


Figure 4.6. The association of the risk allele at rs2296782 and rs56188434 and TCF7L2 expression in the proximal colon, distal colon and rectum of 107 individuals in the INTERMPHEN study. A Jonkheere-Terpstra (JT) test was utilised to test for associations between the risk allele at the lead SNP at the *TCF7L2* locus (rs2296782) and at the lead SNP at the *VTG1A* locus (rs56188434) and TCF7L2 expression in the proximal colon, distal colon and rectum of 107 individuals in the INTERMPHEN study.

Figure 4.6 shows that at the rs56188434 locus there is only one individual homozygous for the risk allele. This could have influenced the results due to lack of data points in this category.

When each of the 53 genome-wide significant SNPs were tested for an association with TCF7L2 expression, seven additional SNPs at the *TCF7L2* locus were found to be significantly associated with increased TCF7L2 expression in the rectum. However, this significance was not maintained following Benjamini-Hochberg correction. P-values were also replicated across SNPs as they were in high LD with one another, and therefore SNPs were likely to share the same genotype. These results suggest that when looking at SNPs in LD, it cannot be deciphered which of the SNPs is truly associated with gene expression.

Following this analysis, the risks alleles at the both lead SNPs at the *TCF7L2* and *VTI1A* locus were tested for associations with *VTI1A* expression in the proximal colon, distal colon and rectum (Table 4.5).

Table 4.5 The effect of risk alleles at both the lead *TCF7L2* locus SNP (rs2296782) and the lead *VTI1A* locus SNP (rs56188434) on *VTI1A* expression in the proximal colon, distal colon and rectum. Shown is the effect allele for each SNP, along with their effect sizes (β) reported in the GWAS. Additionally, standard errors of the effect sizes are also presented along with raw and corrected Jonckheere-Terpstra (JT) p-values, which assess the trend in gene expression across different genotypes.

SNP	Effect allele	$\beta_{(GWAS)}$	$SE_{(GWAS)}$	$p_{(GWAS)}$	$p_{(JT)}$	$p_{(JT)}$	$p_{(JT)}$
					proximal colon	distal colon	rectum
rs2296782	G	0.0555383	0.00863398	1.25498e-10	0.448735	0.016085	0.140943
rs56188434	A	0.102209	0.0125316	3.46048e-16	0.258489	0.231269	0.985123

The risk allele at rs2296782, the lead SNP at the *TCF7L2* locus, was significantly associated with decreased expression of *VTI1A* in the distal colon. The risk allele at rs56188434 however, was not associated with a change in *VTI1A* expression in any tissue types (Fig. 4.7).

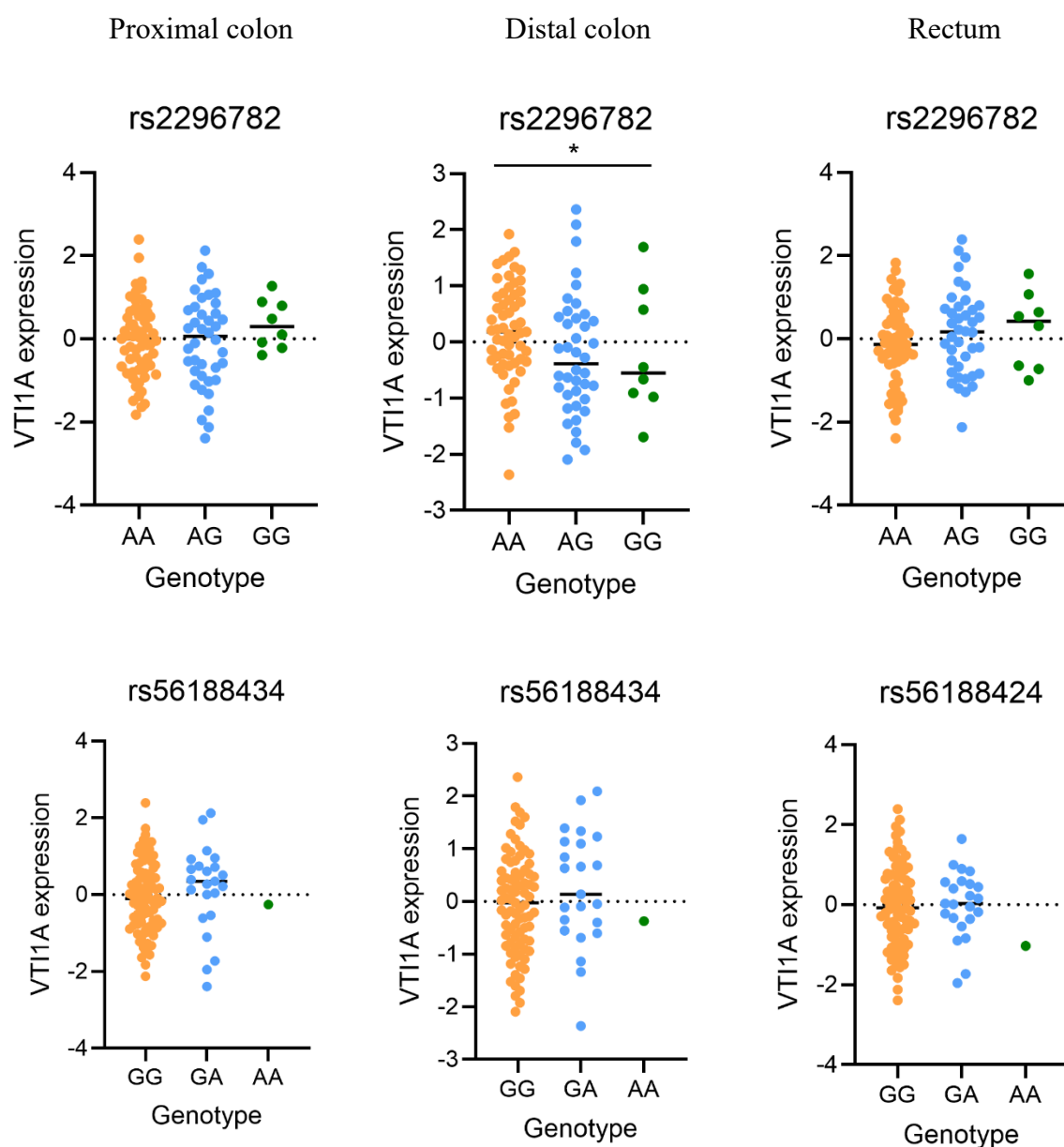


Figure 4.7 The association of the risk allele at rs2296782 and rs56188434 and VTI1A expression in the proximal colon, distal colon and rectum of 107 individuals in the INTERMPHEN study. A Jonkheere-Terpstra (JT) test was utilised to test for associations between the risk allele at the lead SNP at the *TCF7L2* locus (rs2296782) and at the lead SNP at the *VTI1A* locus (rs56188434) and VTI1A expression in the proximal colon, distal colon and rectum of 107 individuals in the INTERMPHEN study.

The risk allele at rs56188434 was not associated with a change in VTI1A expression in any tissue types. This could be affected by only one individual in the dataset being homozygous at this locus. Looking at these results in isolation suggests that an eQTL, associated with a change in *VTI1A* expression is present in *TCF7L2*, but not *VTI1A*.

A caveat of this result was identified through the analysis of all 53 genome-wide significant SNPs. As stated above, the risk allele at the lead SNP, at the *VTI1A* locus, was not associated with a change in *VTI1A* expression. However, 10 risk alleles at different *VTI1A* SNP loci were associated with

an increase in *VTIIA* expression in the distal colon, despite being in LD with the lead, prior to Benjamini-Hochberg correction. None of these SNPs maintained significance following post-hoc correction. These SNPs displayed identical genotypes, indicating a high degree of LD between them. This resulted in repeated expression distribution across genotypes, and repeated statistical results (**Table 4.6**). This highlights a caveat of looking at the association of gene expression and lead SNP alone, despite being in LD with the other SNPs.

When testing all 53 genome-wide significant SNPs for an association for *VTIIA*, much like the lead SNP at the *TCF7L2* locus, an additional five risk alleles at SNPs at the *TCF7L2* locus were associated with decreased *VTIIA* expression in the distal colon, prior to Benjamini-Hochberg correction. These were all in LD with the lead SNP at the *TCF7L2* locus.

Table 4.6 The effect of risk alleles at ten different *VTIIA* SNP loci significantly associated with increased *VTIIA* expression in the distal colon, expression prior to Benjamini-Hochberg correction. Shown is the effect allele for each SNP, along with their effect sizes (β) reported in the GWAS. SNPs in the *TCF7L2* locus are in the upper section of this table, SNPs at the *VTIIA* locus are in the lower section. Additionally, standard errors of the effect sizes are also presented along with raw and corrected Jonckheere-Terpstra (JT) p-values, which assess the trend in gene expression across different genotypes. In addition, r^2 measures for LD with the lead SNP at the *VTIIA* locus are shown, 1 indicating high LD and 0 indicating low LD.

SNP	Effect allele	$\beta_{(GWAS)}$	$SE_{(GWAS)}$	$p_{(GWAS)}$	r^2 with rs56188434	$p_{(JT)}$ proximal colon	$p_{(JT)}$ proximal colon corrected	$p_{(JT)}$ distal colon	$p_{(JT)}$ distal colon corrected	$p_{(JT)}$ rectum	$p_{(JT)}$ rectum corrected
rs11195993	A	0.0304204	0.114521	4.28458e-9	0.457	0.931767	0.931767	0.046514	0.145014	0.748149	0.997025
rs2028783	G	0.102724	0.0157308	6.57216e-11	0.531	0.931767	0.931767	0.046514	0.145014	0.748149	0.997025
rs12266418	A	0.102193	0.0157653	9.04299e-11	0.543	0.931767	0.931767	0.046514	0.145014	0.748149	0.997025
rs2242205	A	0.0984573	0.0168024	4.63633e-9	0.444	0.931767	0.931767	0.046514	0.145014	0.748149	0.997025
rs11815469	G	0.0937769	0.0155611	1.67691e-9	0.529	0.931767	0.931767	0.046514	0.145014	0.748149	0.997025
rs77373153	T	0.0983604	0.0168478	5.27763e-9	0.457	0.931767	0.931767	0.046514	0.145014	0.748149	0.997025
rs17129862	C	0.108684	0.0162397	2.19459e-11	0.502	0.931767	0.931767	0.046514	0.145014	0.748149	0.997025
rs7100296	A	0.108377	0.0163009	2.95964e-11	0.477	0.931767	0.931767	0.046514	0.145014	0.748149	0.997025
rs12254971	G	0.109699	0.0163985	2.23816e-11	0.477	0.931767	0.931767	0.046514	0.145014	0.748149	0.997025
rs12259387	G	0.112824	0.0167061	1.44353e-11	0.477	0.931767	0.931767	0.046514	0.145014	0.748149	0.997025

These results suggest that an eQTL is likely in LD with the lead SNP at the *VTI1A* locus that is associated with increased expression of *VTI1A* in the distal colon. The genotype distribution of two of these 10 SNPs is shown below. The genotype distribution was identical for all 10 SNPs (**Fig. 4.8**).

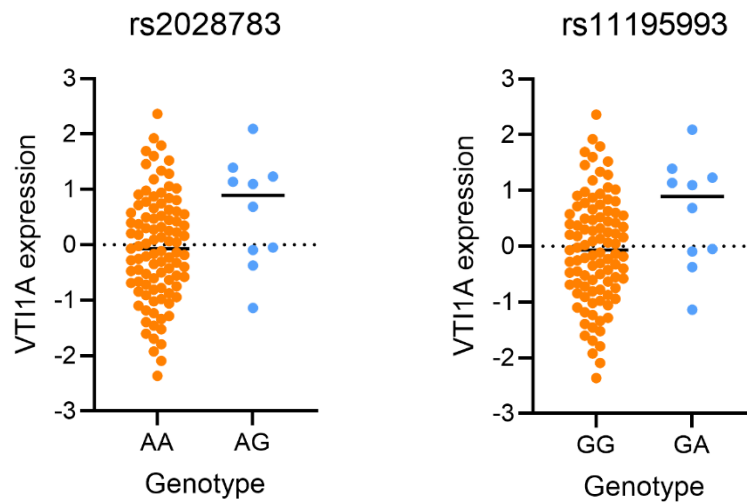


Figure 4.8 The association of risk alleles at two *VTI1A* SNP loci and *VTI1A* expression in the distal colon. The association of risk alleles at 10 different *VTI1A* SNP loci were significantly associated with an increase in *VTI1A* expression prior to Benjamini-Hochberg correction. The other nine SNPs are not presented here as the same genotype was found for each SNP across participants, and therefore this resulted in the same expression distribution across genotypes.

The relationship between *TCF7L2* and *VTI1A* expression in the proximal colon, distal colon and rectum was also analysed irrespective of associations with potentially causative SNPs. This showed a significant, positive relationship between expression in both genes in both the proximal colon ($F = 131.8$, $p < 0.001$) and rectum ($F = 91.43$, $p < 0.001$). No association was shown between *TCF7L2* and *VTI1A* expression in the distal colon ($F = 0.04$, $p = 0.842$) (**Fig. 4.9**).

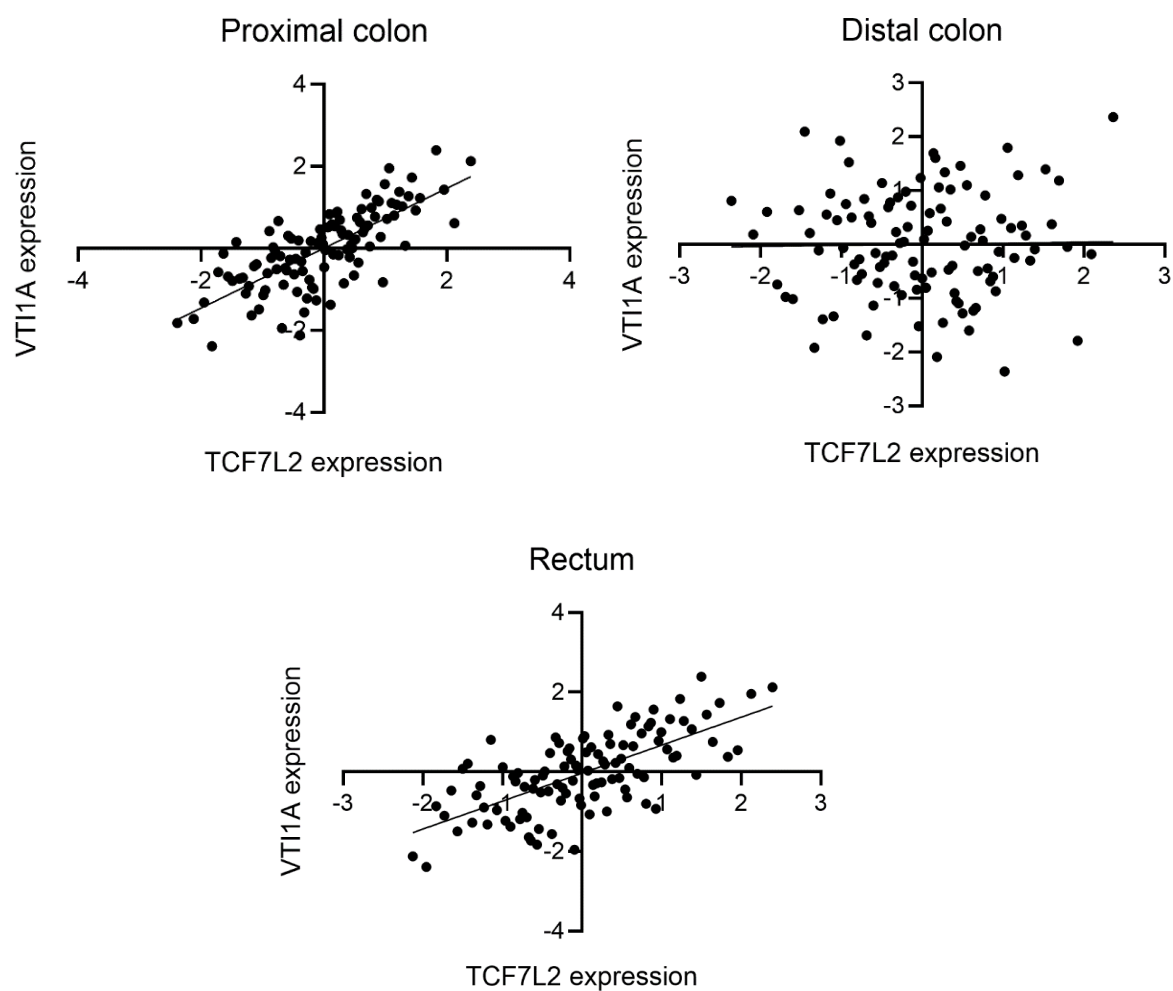


Figure 4.9. The relationship between *TCF7L2* and *VT11A* expression in the proximal colon, distal colon and rectum.

4.4 Discussion

Through the analysis of GWAS meta-analysis data from the study by Fernandez-Rozadilla *et al.*, (2023), the results identified 53 genome-wide significant SNPs at the 10q25.2 locus. This region incorporates both *TCF7L2* and *VTIIA* genes. Thus, the initial question addressed by this analysis was whether the GWAS signal for *TCF7L2* is likely independent of that for *VTIIA* given their shared location within the same region.

Previously, GWAS have identified rs12255131 as the lead SNP in the 10q25.2 region in a European cohort (Law *et al.*, 2019) and, when looking in a larger dataset, Fernandez-Rozadilla *et al.*, (2023) identified rs11196172. Law *et al.*, (2019) did not identify any 10q25.2 eQTLs through the exploration of normal colonic mucosa using the INTERMPHEN dataset, suggesting that neither *VTIIA* nor *TCF7L2* SNPs are associated with changes in gene expression.

In this chapter, new analysis identified the lead SNP for the 10q25.2 region as rs56188434 ($p = 3.46 \times 10^{-16}$) located at the *VTIIA* locus. The lead SNP is likely different from those found in previous data due to an increase in sample size and this study focusing on the 10q25.2 region alone. When looking specifically at the *TCF7L2* locus, no longer including the *VTIIA* region, another major SNP was identified, rs2296782 ($p = 1.25 \times 10^{-10}$). Neither of these SNPs have been identified by previous GWAS according to the GWAS Catalogue. Altogether, 53 SNPs were identified as genome-wide significant across the entire region. It is unlikely that all 53 SNPs represent independent associations with CRC due to the occurrence of LD. A conditional analysis was therefore performed incorporating LD into the statistical model. When conditioning on the lead SNP of the entire region, rs56188434 (*VTIIA* locus), 44 SNPs were no longer associated with CRC and were therefore likely in LD with the lead SNP. Despite 44 SNPs no longer being significantly associated with CRC, nine SNPs remained significant. These nine significant SNPs were at the *TCF7L2* locus. These significant SNPs represent an association with CRC independent of the lead SNP, rs56188434. To check if these SNPs represented nine independent associations with CRC, the major SNP at the *TCF7L2* locus, rs2296782 was added into the conditional analysis so that p-values could be recalculated accounting for LD with this SNP. None of the SNPs remained significantly associated with CRC, indicating that all nine SNPs at the *TCF7L2* locus were in LD with the lead in this region (rs2296782). In conclusion, this signifies the existence of two independent GWAS signals significantly associated with CRC at the 10q25.2 locus, specifically, one at the *VTIIA* locus and another at the *TCF7L2* locus. These suggest that SNPs at the *TCF7L2* locus are associated with CRC risk, independently of *VTIIA*.

This conditional analysis has allowed for the identification of two independent GWAS signals at the 10q25.2 locus, originating from two separate clusters of SNPs. Importantly however, it has not provided information regarding which of these SNPs are likely causal, due to LD. An additional caveat of this methodology is that it relies solely on a p-value threshold to identify SNPs of statistical

significance. The outcome can therefore be affected by study-specific factors such as power. Thus, results are not easily comparable between studies (Spain and Barrett, 2015).

A benefit of this study is that it uses a more lenient p-value threshold than other studies ($p_{\text{(conditional)}} = 5 \times 10^{-7}$). For example, Law *et al.*, (2019) used a p-value threshold of 5×10^{-8} . This helps to identify weaker secondary signals that are potentially independent of the lead SNP, allowing for a more in-depth evaluation of a larger number of signals at the locus.

A different methodology, able to further assist in the identification of causal SNPs, is Bayesian fine-mapping. Bayesian fine-mapping utilises statistical models to assess which variants are most likely to be causal. Assessing the association between SNPs and a given phenotype, such as CRC, relies upon the computation of Bayes Factor. Bayes Factor gauges the strength of evidence for a SNP being causal by comparing the marginal likelihood of the data, or how well the generic data fits the statistical model, under varying prior probability distributions. It outputs a credible set of SNPs with the highest likelihood of being causal. The larger the Bayes factor, the stronger the evidence for a SNP being causal (Hutchinson, Asimit and Wallace, 2020). A pitfall of early Bayesian based fine-mapping strategies was their assumption that there is only one causal SNP at any particular locus. Although this approach only requires marginal association statistics, according to Yang *et al.*, (2011), this may be an oversimplification, as traits can be linked to multiple SNPs (Yang *et al.*, 2013b; Kichaev *et al.*, 2014). Another limitation of early Bayesian fine-mapping methods was that they did not integrate functional annotations of variants or gene expression data to calculate posterior probabilities. This information can help to prioritise causal variants. Without this information, the model may prioritise variants that are statistically associated with a trait but have no biological relevance in the tissue of interest (Kichaev *et al.*, 2014). PAINTOR is a fine-mapping tool, developed in 2014, that accounts for multiple causal variants and incorporates functional annotation data to assist in the prioritisation of causal variants. This helps to avoid situations where the model prioritises statistically associated variants that lack functional importance (Kichaev *et al.*, 2014). It may therefore be valuable to re-run this analysis using PAINTOR to gain deeper insight into specific variants and their association with CRC in the context of both LD and functional annotation of SNPs. Conducting this analysis across multiple cohorts of different ethnicities could help to determine whether SNPs in the 10q25.2 region are consistently associated with CRC across diverse genetic backgrounds.

The results of the eQTL analysis in this chapter show that the presence of risk alleles at the lead SNP, at the 10q25.2 locus (rs56188434) and major SNP at the *TCF7L2* locus (rs2296782) were not significantly associated with a change in *TCF7L2* expression in the proximal or the distal colon. This matches the results found by both Law *et al.* (2019) and Fernandez-Rozadilla *et al.* (2023). In addition, the risk allele at rs56188434 was also not associated with a change in *TCF7L2* expression in the rectum. Interestingly, the risk allele at rs2296782 was significantly associated with an increase in

TCF7L2 expression in the rectum. When risk alleles at all 53 genome-wide significant SNPs were tested for an association with TCF7L2 expression, risk alleles at seven additional SNPs at the *TCF7L2* locus were found to be significantly associated with an increase in TCF7L2 expression in the rectum, but nothing remained significant following post-hoc correction. This result is likely to be heavily influenced by LD, it is therefore unclear which SNP may represent an eQTL associated with increased TCF7L2 expression. Interestingly, this association being identified in the rectum only suggests that genetic variation at these SNPs may affect the regulation of TCF7L2 differently depending on the tissue type. This has been reported previously. Vaughan-Shaw et al. (2021) reported differing site-specific eQTL effects of genetic variants at the 11q23.1 locus. Altogether, the association of these SNPs with both CRC and increased TCF7L2 expression suggests that *TCF7L2* could be a candidate risk gene of CRC, specifically in the rectum, subject to confirmation via further functional studies.

The presence of risk alleles at both rs56188434 and rs2296782 were then tested for an association with changes in VTI1A expression in the proximal colon, distal colon and rectum. This interestingly identified a significant association between risk alleles at rs2296782, at the *TCFL2* locus and decreased expression of VTI1A in the distal colon. When testing the presence of risk alleles at all 53 SNPs for an association, five additional risk alleles at *TCF7L2* SNP loci were also associated with decreased expression of VTI1A in the distal colon. Although, significance was not maintained following Benjamini-Hochberg correction, likely influenced by LD. This result suggests that there is an eQTL at the *TCF7L2* locus associated with decreased VTI1A expression, but further studies are required to elucidate which SNP is causative.

Risk alleles at the lead SNP for the entire region, rs56188434 (*VTI1A* locus) was not significantly associated with changes in VTI1A expression in any of the different tissue types. Notably, when testing risk alleles at all 53 genome-wide significant SNPs for an association with changes in VTI1A expression, ten risk alleles at different *VTI1A* SNP loci were associated with increased VTI1A expression in the distal colon, despite being in LD with the lead. However, significance was not maintained following Benjamini-Hochberg correction. Each of these ten results displayed exactly the same genotype distribution indicating a high level of LD. This study would greatly benefit from the implementation of Bayesian fine-mapping to reduce the number of credible, genome-wide significant SNPs. This, in combination with a larger dataset, may reduce the heavy influence of post-hoc correction on the results. These results suggest that there could be an eQTL at the *VTI1A* locus associated with an increase in the expression of VTI1A in the distal colon, therefore identifying *VTI1A* as a candidate risk gene of CRC, but this interpretation must be approached with caution due to heavy influences of LD and sample size on the resulting statistics. Candidate causal SNPs can also be validated using both CRISPR/Cas9 and mouse models.

Interestingly, Fernandez-Rozadilla *et al.*, (2023) also performed a transcriptome-wide association study (TWAS) to determine if different gene expression is linked to CRC. TWAS integrates gene expression data with known GWAS variants, predicts gene expression as a result of each variant and determines if gene expression is significantly linked to a trait (Mai *et al.*, 2023). Fernandez-Rozadilla *et al.*, (2023) did not find *TCF7L2* expression to be linked to CRC, except when looking at expression specifically in immune cells e.g. lymphocytes and spleen. Since some of the variants in this study have not been identified by previous GWAS, it may be of value to reanalyse the TWAS data to investigate whether *TCF7L2* or *VTI1A* expression is significantly linked to CRC via the above identified genome-wide significant SNPs.

Once eQTLs are validated by further statistical analysis and functional studies, it may be of interest to test the response in different tissues. For example, multiple SNPs were associated with a change in *TCF7L2* expression in the rectum, thus a CRISPR/Cas9 knock-in of a *TCF7L2* mutation in a cell line of rectal origin could be compared to the same mutation in a cell line originating from the colon. Differences in growth, survival and morphology could be analysed between cell lines. In addition, the results showed a positive correlation between *TCF7L2* and *VTI1A* expression in both the proximal colon and rectum. Knocking a *VTI1A* mutation into a *TCF7L2*-mutated cell line could also theoretically alter the growth, survival and morphology of the cell line. This could also be analysed using cell lines originating from different locations in the colorectum. In addition, if mutations in both genes resulted in increased tumour formation in mouse models, this could be a potential case of mini-driver cooccurrence. Mutations in these genes can be replicated in mouse models, specifically in gut epithelium with the use of villin-cre (vil-Cre) and vil-Cre-ERT2 mice. Vil-Cre-ERT2 mice cause gene recombination in the gut epithelium following tamoxifen induction (El Marjou *et al.*, 2004a).

In this analysis, SNPs and their association with gene expression has only been tested on nearby genes, *VTI1A* and *TCF7L2*. SNPs associated with a change in expression of nearby genes are called cis-eQTLs. Conversely, trans-eQTLs can regulate the expression of genes up to 5Mb away from the SNP locus (Võsa *et al.*, 2021). Capture Hi-C (CHi-C) can be utilised to identify long-range chromatin interactions, helping to map SNPs to their long-distance target genes. This is particularly useful to map SNPs in non-coding regions to their target genes (Law *et al.*, 2019). Chi-C could be used in this instance on the genome-wide significant SNPs identified in this study to identify if they act as trans-eQTLs. It is via this method that Law *et al.*, (2019) identified a looping interaction between an enhancer region in *VTI1A* and *TCF7L2*. They concluded that this made *TCF7L2* the target gene of the rs12255141 SNP. This is potentially a slightly premature conclusion being as their eQTL analysis did not identify rs12255141 as associated with altered *TCF7L2* expression. Again, knock-in experiments would ideally be performed to assess *TCF7L2* expression in the presence of this *VTI1A* mutation, as although the looping interaction could represent the ability of *VTI1A* to alter *TCF7L2* expression the opposite is also possible. Interestingly, the eQTL analysis also showed an association of six *TCF7L2* SNPs with reduced

VTI1A expression in the distal colon, the looping interaction may therefore be linked to the ability of *TCF7L2* to downregulate *VTI1A* expression. A *TCF7L2* knock-out CRISPR model would be of interest to analyse the effect on *VTI1A* expression.

In conclusion, this chapter confirms that there are two independent signals at the 10q25.2 locus, one at the *VTI1A* locus and another at the *TCF7L2* locus. Risk alleles at multiple SNPs at the *TCF7L2* locus are associated with increased *TCF7L2* expression in the rectum suggesting their categorisation as eQTLs. Risk alleles at SNPs at the *VTI1A* locus were associated with increased *VTI1A* expression in the distal colon prior to Benjamini-Hochberg correction. This suggests that *VTI1A* may also be associated with CRC risk but this is difficult to properly interpret with the current data. Through both the use of Bayesian fine-mapping and a larger expression dataset, this may further elucidate candidate causal SNPs at the *VTI1A* locus. The implementation of TWAS may also provide further support of SNPs in both genes being categorised as eQTLs. Nonetheless, these results suggest that *TCF7L2* is a candidate CRC risk gene in its own right, through the influence of both *TCF7L2* and *VTI1A* expression, subject to further statistical and functional studies.

Notably, risk alleles at multiple SNPs at *TCF7L2* loci were associated with decreased *VTI1A* expression suggesting that *TCF7L2* could play a role in the regulation and fine-tuning of *VTI1A* expression in the distal colon. In addition, *TCF7L2* and *VTI1A* expression were found to be positively correlated in both the proximal colon and rectum. This relationship between *TCF7L2* and *VTI1A* could be further explored in future studies to determine with this pattern of cooccurrence represents a mini-driver relationship, which would be further evidence for the just-right model.

Through the recognition of *TCF7L2* as a potential CRC risk gene, its influence on CRC evolution, and its categorisation as a mini-driver as CRC, can be further examined through the production of a mouse model. This is carried out in the following chapter.

Chapter 5

An *in-vivo* analysis of *TCF7L2* as a mini-driver of CRC

5.1 Background

As stated in the previous chapter, *TCF7L2* is a Wnt pathway transcription factor, able to form a nuclear complex with β -catenin. *TCF7L2* is highly expressed in white adipose tissue, in which its role is to regulate gene expression of genes involved in lipid and glucose metabolism (Geoghegan *et al.*, 2019). Interestingly, *TCF7L2* is the genetic locus most strongly linked to type 2 diabetes risk, and the first to be reliably identified through GWAS (Gloyn, Braun and Rorsman, 2009). *TCF7L2* also performs multiple roles in brain development including regulating the terminal selection of thalamic neurons and aiding oligodendrocyte maturation (Szewczyk *et al.*, 2024). In the intestines, *TCF7L2* is essential for intestinal epithelium regulation (Wenzel *et al.*, 2020). *TCF7L2* mutations are present in approximately 10% of CRCs (Fröhlich, Rose and Hecht, 2023). The structure of *TCF7L2*, with its 17 exons can be found in Figure 5.1.

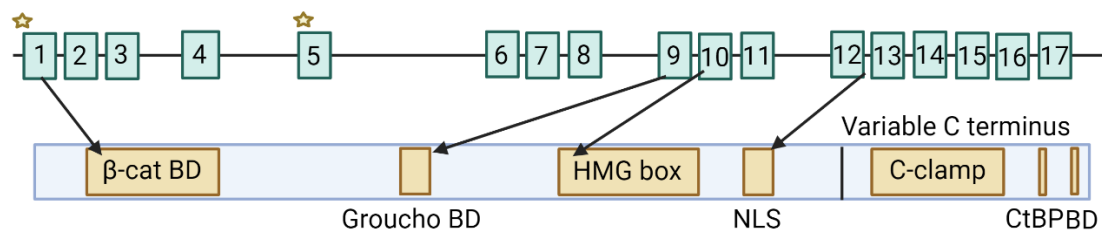


Figure 5.1 The structure of *TCF7L2*. Upper chart: gene structure showing 17 exons (green boxes) of *TCF7L2* are displayed. Asterisks mark start codons. Lower chart: protein structure showing several domains (brown boxes): BD = binding domain. β -cat = β -catenin. HMG = High mobility group domain. NLS = nuclear localisation sequence. CtBP = C-terminal binding protein. Adapted from (Wenzel *et al.*, 2020).

As shown in chapter 3, in CRC, *TCF7L2* mutations co-occur with mutations in *APC* in the context of MSS CRCs, in the 100KGP, providing supporting evidence for the just-right model. Interestingly, Angus-Hill *et al.* (2011) developed multiple edited murine alleles of *Tcf7l2* to test the function of *Tcf7l2* *in-vivo*. In this paper they refer to *Tcf7l2* as *Tcf4*. Angus-Hill *et al.*, (2011) produced multiple alleles to study the function of *Tcf4*. Firstly, to generate a null allele, they replaced exon 1 of *Tcf4* using an EGFP recombinase fusion (*Tcf4^{Cre}*). They targeted exon 1 specifically as it contains the β -catenin binding domain, and they stated that exon 1 is targeted with both nonsense and frameshift mutations in human CRC. In addition, they also flanked exon 1 with loxp sites, allowing for conditional disruption of *Tcf4* (*Tcf4^{Lox}*). They then bred *Tcf4^{Lox}* to *Hrpt^{cre}* to generate a null allele (*Tcf4^Δ*).

Using all of these alleles, Angus-Hill *et al.*, (2011) were able to establish that mice with complete loss of both alleles of *Tcf4* were not viable at birth, using the alleles *Tcf4^{Cre/Cre}*, *Tcf4^{Cre/Lox}*, and *Tcf4^{Cre/Δ}*. Following this, they aimed to better control the extent of *Tcf4* deletion to better establish the effect of *Tcf4* loss. For this purpose, they created a *Tcf4^{CreNeo/Lox}* (*Tcf4^{NeoNull}*) line. The purpose of the *Neo* gene is to limit the number of cells expressing Cre, therefore reducing the number of cells where

Tcf4 is deleted completely. This therefore produced a hypomorphic allele in which gene function is deleted in some cells but not lost completely in all cells. In these *Tcf4^{NeoNull}* mice intestinal crypts were enlarged but they observed a lack of tumours. They concluded that hypomorphic *Tcf4* loss was not adequate by itself for tumourigenesis. This is supportive evidence for *TCF7L2* as a mini-driver as mini-drivers are hypothetically unable to induce tumourigenesis in isolation (Castro-Giner, Ratcliffe and Tomlinson, 2015). Thus, they predicted that additional genetic mutations would be necessary for tumourigenesis. To further explore this hypothesis, Angus-Hill *et al.*, (2011), induced a heterozygous *Tcf4* knock-out allele (*Tcf4^{Het}*) onto an *Apc^{Min}* background. They found that *Tcf4^{Het};Apc^{Min}* mice produced a significantly higher number of tumours in the colon than *Tcf4^{WT};Apc^{Min}* mice. These results suggest that *Tcf7l2* may be capable of fine-tuning the Wnt signalling provided by an *Apc* mutation in the colon of heterozygous knock-out mice. *Apc^(Min/+)* mice harbour a p.Arg850Ter mutation. This mutation is a major-driver mutation, particularly potent in driving the formation of small intestinal tumours, and a small number of colonic tumours. These mice show a high mortality rate with increased age (Bürtin, Mullins and Linnebacher, 2020).

As stated above, *TCF7L2* mutations are present in 10% of CRCs, therefore CRCs are capable of occurring without a *TCF7L2* mutation. *TCF7L2* is also a transcription factor, with the ability to modulate the expression of other genes in the Wnt pathway. In chapter 4, *TCF7L2* was shown to potentially contribute to CRC risk, subject to confirmation via further functional and statistical studies. In Chapter 3, a pattern of co-occurrence between mutations in *TCF7L2* and *APC* was identified in a large CRC dataset. So far, this evidence is supportive of the categorisation of *TCF7L2* as a Wnt-pathway mini-driver. This chapter aims to validate and further explore the findings of Angus-Hill *et al.*, (2011) to establish whether *TCF7L2* mutations are capable of driving tumourigenesis in isolation, whether more tumours are produced in combination with an *APC* mutation, and whether there are any tissue-specific effects of a *TCF7L2* mutation throughout the small and large intestines.

Notably, the modification of other Wnt pathway genes in combination with *Apc* mutations in mouse models have significantly influenced tumour development. For instance, β -catenin knock-out mutations have been shown to cause complete elimination of tumours when crossed onto an *Apc^{1638N}* background, highlighting the dependence of this *Apc* mutation on β -catenin for tumour formation (Bakker *et al.*, 2013). In addition, deletion of *Bcl9/9l* has been shown to increase adenoma formation in the proximal small intestine while suppressing growth in the colon (Gay *et al.*, 2019).

In this chapter a *Tcf7l2^(fl/fl)* line was crossed with both constitutive *Pgk1-cre* and *Vill-Cre* lines. In the *Pgk1-cre* line, the Cre recombinase enzyme is driven by the *Pgk1* promoter which is active in all organs. A *Tcf7l2* knock-out would therefore occur in all organs throughout the body (Lonai, 1998). Conversely, Cre recombinase driven by the *Vill-cre* promoter is highly specific to the intestinal epithelium (Rutlin *et al.*, 2020).

5.2 Aims

This chapter aims to further investigate the work of Angus-Hill *et al.*, (2011) and test whether *Tcf7l2* can act as a candidate mini-driver of CRC. Murine model intestinal tumour analysis is carried out to determine whether a heterozygous conditional knock-out mouse model of *Tcf7l2* can induce intestinal tumourigenesis and therefore act as a driver of CRC or, whether a knock-out of *Tcf7l2* is beneficial for tumourigenesis in the presence of major-driver *Apc* mutation. This study utilises both *Pgkl-cre* and *Vill-cre* mice to produce constitutive heterozygous knock-outs of *Tcf7l2* across all organs and specifically in the intestinal epithelium, with the purpose of exploring tumour number and volume in both the small and large intestines to validate or dispute the findings of Angus-Hill *et al.*, (2011), therefore providing either supporting or contradictory evidence for both the mini-driver model and the just-right model.

5.3 Methods

5.3.1 *Tcf7l2* mouse strain generation

The *Tcf7l2*^(fl/fl) conditional knock-out mouse, originally designed and generated by Angus-Hill *et al.*, (2011), was obtained from the Jackson Laboratory (catalogue #031436). This line was backcrossed onto a C57BL/6J background for three consecutive generations. The *Tcf7l2*^(fl/fl) conditional mouse construct designed by Angus-Hill *et al.*, (2011), at the stage we received it, is set out in the figure below (Fig. 5.2).

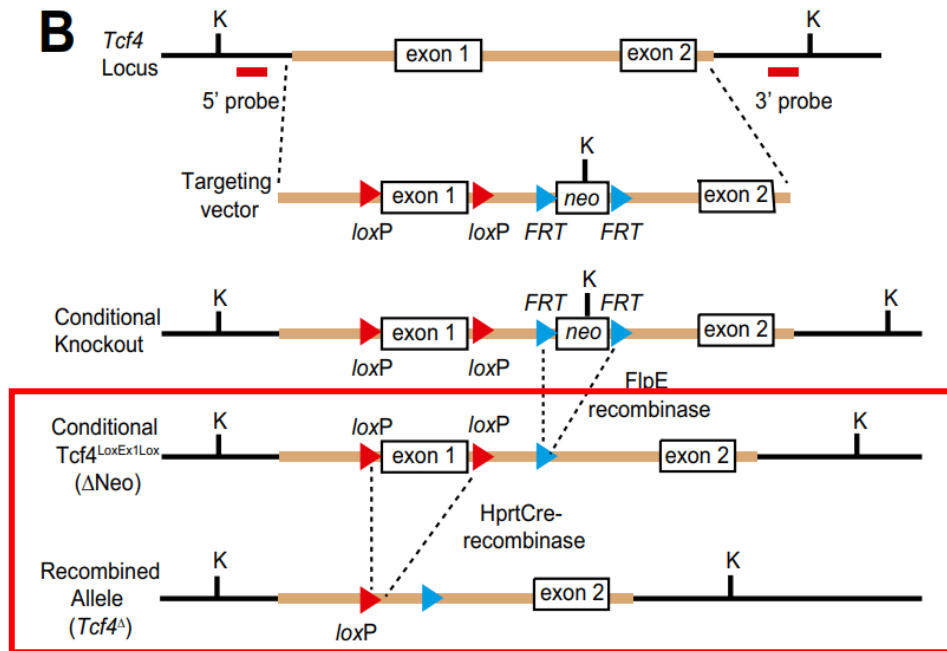


Figure 5.2 The conditional knock-out *Tcf7l2* murine allele with *loxP* sites flanking exon 1. The murine *Tcf7l2* allele locus, also known as *Tcf4*, containing exon 1 and exon 2, is shown in the diagram, with targeting vector construct and conditional knock-out allele structure, as received from the Jackson Laboratory. They are represented diagrammatically, together with the allele structure following recombination by either FlpE recombinase to remove the *neo* cassette or Cre recombinase to excise exon1. Figure by Angus-Hill *et al.*, (2011).

Pgkl-cre (catalogue #020811) and *Vill-cre* (catalogue #035595) were also obtained from the Jackson Laboratory and backcrossed with C57BL/6J mice or three generations. *Apc*^(Min/+) mice were kindly gifted by Professor Malcolm Dunlop's research group.

To derive experimental mice containing the appropriate alleles, *Pgkl-cre* mice were crossed with *Tcf7l2*^(fl/fl) knock-out mice. Resulting *Tcf7l2*^{(+/-);Pgkl-cre}(pos) offspring were then crossed with *Apc*^(Min/+) mice to produce a double-mutant murine line. Strains are detailed in Table 4.1. *Tcf7l2*^(fl/fl) knock-out mice were also crossed with *Vill-cre* mice to produce a heterozygous *Tcf7l2* strain: *Tcf7l2*^{(+/-);Vill-cre}(pos). No instances of *Tcf7l2*^{(fl/fl);Vill-cre}(pos) were observed in over 20 crosses

providing supportive evidence for the conclusion made by Angus-Hill *et al.*, (2011) that a homozygous knock-out of *Tcf7l2* is likely to be lethal.

Unfortunately, within the time constraints of this PhD project, it was not possible to cross *Tcf7l2*^(+/fl); *Vill-cre*^(pos) mice with the *Apc*^(Min/+) line.

Table 5.1 Summary of the mouse strains utilised in the study of *Tcf7l2* as a mini-driver of intestinal tumourigenesis. This table describes the mouse strains utilised in this study of *Tcf7l2*, including the construct's target, mouse genome informatics (MGI) number, scientific nomenclature and notes on the application in intestinal tumourigenesis research.

Strain	Targeted construct	MGI	Nomenclature	Notes
<i>Tcf7l2</i> ^(fl/fl)	LoxP sites flank exon 1 of <i>Tcf7l2</i> gene	170088	B6.Cg- <i>Tcf7l2</i> ^{tm2.1MRc} /MahllJ	Model for candidate mini-driver of intestinal tumourigenesis
<i>Pgk1-cre</i>	Transgene insertion 1	2178050	Tg(PGK-cre)1Lni	Cre driver line for ubiquitous expression of Cre
<i>Vill-cre</i>	Transgene insertion 20	92295	Tg(Vill-cre)20Syr	Cre driver line for intestinal expression of Cre
<i>Apc</i> ^(Min/+)	ENU-induced point mutation altering nucleotide 2549 (T>A), causing a nonsense mutation at codon 850	1856318	<i>Apc</i> ^{Min}	Model for intestinal tumourigenesis

5.3.2 Genotyping

Ear-snips were biopsied from each mouse at 14 days of age and stored at -20°C. DNA was extracted from ear-snips through the addition of 50µl of Hotshot Buffer (25mM NaOH, 0.2mM disodium ethylenediaminetetraacetic acid, pH 12). The mix was then incubated for 60 minutes at 95°C. 50µl of Neutralising Buffer (40mM Tris-HCl, pH 5) was added following incubation.

The following primers were utilised for genotyping different strains of mice by PCR (**Table 5.2**).

Table 5.2 A description of the PCR primers utilised to genotype the strains utilised in this study. Primers include those for *Tcf7l2*^(fl/fl), general *Cre*, *Apc*^(Min/+) and those utilised to assess recombination of the *Tcf7l2*^(+/-) allele in the mice previously designed to assess recombination by Mohan *et al.*, (2011). Also presented is the primer located within Exon 2 of *Tcf7l2*, utilised to confirm position of primers by Mohan *et al.*, (2011).

Strain	Primers
<i>Tcf7l2</i> ^(fl/fl)	Forward: 5' CAAAACAGCTCCTCCGATTC 3' Reverse: 5' CCGGGATTTATCTCGGAAAC 3'
<i>Cre (Pgk1/Vil1)</i>	Forward: 5' TTACCGGTCGATGCAACGAG 3' Reverse: 5' CCACCGTCAGTACGTGAGAT 3'
<i>Apc</i> ^(Min/+)	MAPC: 5' TGAGAAAGACAGAAGTTA 3' MAPC 15: 5' TTCCACTTTGGCATAAGGC 3' MAPC 9: 5' GCCATCCCTTCACGTTAG 3'
<i>Tcf7l2</i> ^(+/-)	Forward: 5' CCTCCTTCATTTTTCTTCCAAAATTGCTGCTGCTGG 3' Reverse: 5' GCCCGACCCGCCTCGCTCGCTCG 3' Exon 2 reverse: 5' AAACCTTTCGGAGCGAGGCGGAGGCCGTCTT 3'

To genotype both *Tcf7l2*^(fl/fl) or *Tcf7l2*^(fl/+) mice, polymerase chain reaction (PCR) steps included an initial denaturing step was performed at 94°C for two minutes. A three-step cycle was then repeated 37 times, this included an initial 94°C step for 2 minutes, followed by 30 seconds at 55°C for annealing and a 3-minute extension time at 72°C. A final extension at 72°C was then carried out for 7 minutes. DNA was held at 10°C before being stored at -20°C.

To genotype both *Pgk1-cre* and *Vil-cre* strains, an initial denaturing step was performed at 98°C for 30 seconds, followed by a 65°C step performed for 45 seconds. These two steps were cycled nine times. A three-step cycle was then repeated 27 times: an initial step at 94°C for 7 seconds, followed by a 60°C annealing step for 30 seconds and a 72°C extension step for 30 seconds. A final extension was performed at 72°C for 2 minutes. DNA was held indefinitely at 10°C before being stored at -20°C.

To genotype *Apc*^(Min/+) and *Apc*^(+/-) mice, an initial denaturing step was performed at 94°C for 1 minute. The next three steps were repeated 30 times. Firstly, another step at 94°C was performed, followed by an annealing temperature at 54°C for two minutes. An extension temperature of 72°C was performed for 3 minutes. A final extension temperature of 72°C was then carried out for 10 minutes before DNA was held indefinitely at 10°C.

5.3.3 Temperature optimisation and gel extraction

Temperature optimisation of primers was performed as follows: an initial denaturation step at 95°C was performed for 5 minutes. The next three steps were repeated 35 times. Firstly, a 95°C step

was performed for 30 seconds, followed by an annealing step. NEB Tm calculator was utilised to calculate the optimal annealing temperature of primers (tmcalculator.neb.com). I then tested this temperature along with two lower temperatures and two higher temperatures. For example, if the optimal temperature was 60°C I would also test 58°C, 59°C, 61°C and 62°C. The annealing step was followed by an extension at 72°C for 30 seconds. A final extension step was performed at 72°C for 10 minutes.

To do a gel extraction, I would put 500ng of DNA into the initial PCR and run the entire PCR on a gel. I would then use a blue light box and scalpel to cut out the bands and utilise the Qiaquick gel extraction kit from Qiagen (#28704) to extract DNA. DNA was sequenced via the Sanger sequencing service at the IGC.

5.3.4 Tissue retrieval, processing and analysis

The main experimental mice used, $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and the two cohorts of control mice $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/-)};Apc^{(+/+)}$ were all aged to approximately 100 days before being culled and analysed. Additional controls were aged to over one year of age. Details can be found in Table 4.3. The weight of each mouse was recorded prior to being culled.

Table 5.3 A summary of the numbers and ages of mice analysed from each mouse strain.

Strain	Number of mice	Age
$Tcf7l2^{(fl/fl)}$	3	1 year 19 days
$Tcf7l2^{(+/fl)}$	3	1 year 5 days
WT: $Tcf7l2^{(+/+)};Apc^{(+/+)}$	4	105 days
$Tcf7l2^{(+/+)};Pgkl-cre^{(pos)}$	2	1 year 5 days
$Tcf7l2^{(+/-)};Apc^{(Min/+)}$	6	75-107 days*
$Tcf7l2^{(+/+)};Apc^{(Min/+)}$	5	100-105 days
$Tcf7l2^{(+/-)};Apc^{(+/+)}$	5	100-105 days
$Tcf7l2^{(+/fl)};Vill-cre^{(pos)}$	3	1 year 5 days

Two $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ mice had to be culled before the scheduled cull and tissue analysis date due to the development of significant symptoms e.g. pale paws and ears and bleeding from anus. These mice were culled at 75 and 103 days*.

The liver, spleen, small and large intestines were retrieved from each mouse at necropsy. The length and width of the liver and spleen were recorded. Intestines were cleaned with phosphate-buffered

saline (PBS). The small intestine and colon were cut open longitudinally and the length of each was measured. A mix of 300mL methanol, 150mL chloroform and 50 mL glacial acetic acid was made prior to the procedure. The tissue was covered in methacarn and incubated at room temperature for 10 minutes. The small intestine was divided (and labelled) into 5cm sections and the large intestine was divided (and labelled) into 1cm sections. Tumours within each section were identified macroscopically, measured and counted. The tissue from the colon was then coiled according to the standard Swiss-roll technique. This was also carried out for three segments of the small intestine. All Swiss-rolls were submerged in 4% formaldehyde solution, buffered, pH6.9 (#1004965000) for 24 hours before this solution was replaced with 100% ethanol. The samples were then submitted to the histology technical services at the IGC to be processed and embedded into paraffin blocks and for cutting into 5µm sections and haematoxylin and eosin (H&E) staining.

5.3.5 DNA/RNA extraction

To extract DNA/RNA from spleen, liver, small intestine and colon samples, both the Qiagen DNeasy blood and tissue kit (#69506), and the AllPrep DNA/RNA Minikit (#80204) were utilised according to the protocols.

5.3.6 qPCR

Quantitative PCRs (qPCRs) were carried out via two methods depending on the probes utilised, and their effectiveness with differing fluorescent dyes.

Firstly, a one-step method was performed in which both reverse transcription and qPCR were conducted in the same plate. A TaqMan, FAMTM dye-labelled MGB probe from ThermoFisher (#4331182), spanning the boundary between exons 5 and 6 of *Tcf7l2*, was used. Additionally, a Qiagen QuantiTect *Gapdh* probe (#QT01658692) was utilised as a control. The Applied Biosystems QuantStudio5 384 Real-Time qPCR system was then used to perform the one-step qPCR during which cDNA is produced and amplified before fluorescence is assessed. Fluorescence is assessed via a threshold cycle (Ct) value which refers to the cycle number at which the fluorescence level is significantly higher than a background threshold.

Additional qPCRs were also performed using differing probes to optimise the analysis of *Tcf7l2* expression. Firstly, a qPCR was performed using the same probes as used to genotype the recombined allele of *Tcf7l2* as shown in Table 4.4. These probes span the deleted region of exon 1 of *Tcf7l2*. Following this, another qPCR was performed using a forward probe located inside of the deleted region of Exon 1, coupled with the reverse probe utilised in the previous qPCR. The sequences of these probes can be found in the table below (Table 5.4). Again, the QuantiTect *Gapdh* probe was used as a control.

Table 5.4 Probes used for amplifying and assessing the expression of *Tcf7l2* via qPCR. Presented are the probes flanking the deleted region of Exon 1 of *Tcf7l2*. Additionally, a forward primer designed to sit inside the deleted region of Exon 1 is included.

Gene target	Sequence
<i>Tcf7l2</i> (outer deleted region)	Forward: 5' CCTCCTTCATTTTCTTCCAAAATTGCTGCTGCTGG 3' Reverse: 5' GCCCGACCCGCCTCGCCTCGCTCG 3'
<i>Tcf7l2</i> (inner deleted region)	Forward: 5' AAACCTTTCGGAGCGAGGCGGAGGCCGTCTT 3' Reverse: 5' GCCCGACCCGCCTCGCCTCGCTCG 3'

The *Tcf7l2* probes shown in Table 5.4 were utilised alongside the SYBR Green PCR Master Mix (Applied Biosystems) in a two-step protocol. In this instance, synthesis of cDNA was carried out using the Quantitect Reverse Transcription kit from Qiagen (#205311) according to the protocol. The Applied Biosystems QuantStudio5 384 Real-Time qPCR system amplified the cDNA and assessed levels of fluorescence.

5.3.7 Western blot

To extract protein from liver samples, 500µL of complete radioimmunoprecipitation assay (RIPA) buffer, supplemented with protease and phosphatase inhibitors, was added to approximately 1cm³ of frozen liver tissue (-80°C). The tissue was then homogenised for 1-5 minutes using the Qiagen TissueLyser LT (#85600). Samples were placed on ice for 30 minutes, following which they were centrifuged at 20,000xg for 10 minutes. The supernatant was aliquoted into a new tube and stored at -80°C. Following extraction, protein concentration was quantified through BCA assay using the Pierce™ BCA protein assay kit from ThermoFisher (#23225) according to the protocol.

40µg of protein sample was added to 13µL dH₂O, 5µL NuPage SDS sample buffer and 2µL NuPage antioxidant. The mix was subsequently loaded into an NuPage 3-8% Tris-AcetateProtein gel and 120V applied for one hour. Proteins were transferred to a nitrocellulose membrane via electrotransfer using the Trans-Blot Turbo Midi 0.2µM Nitrocellulose membranes. The resulting membrane was then blocked for one hour at room temperature using 5% BSA in 0.1% TBS Tween-20 (TBS-T). The membrane was then incubated overnight at 4°C with the primary recombinant rabbit monoclonal *Tcf7l2* antibody (SC06-90 #MA5-32240), binding to aa 1-49, diluted in BSA/TBS-T. The membrane was then washed for 15 minutes in TBS-T. The wash was repeated three times, following which the membrane was incubated with the horseradish peroxidase (HRP) conjugated secondary, Gapdh, goat, anti-rabbit antibody (#P1-1000) for one hour at room temperature. The membrane was

again washed three times with TBS-T and imaged using the Bio-Rad image Lab. Analysis was then performed using Fiji.

5.4 Results

5.4.1 Construct confirmation

Primer sequences able to test for recombinational deletion of exon 1 in the *Tcf7l2* conditional knock-out mouse model were provided by Mohan *et al.*, (2022). Mohan *et al.*, (2022) utilised the *Tcf7l2* knock-out mouse produced by Angus-Hill *et al.*, (2011) to study osteoblast function.

Notably, a sequence map of the construct was not made available in previous literature. Thus, the first aim of this project was to develop a sequence map, confirming both positions of the loxP sites, flippase recognition target (FRT) site and primers utilised to test for recombination. This would allow confirmation of whether the primers could reliably be used to assess recombination at loxP sites and therefore deletion of exon 1. A schematic of the general construct, as described in the Angus-Hill *et al.*, (2011) paper, can be found below (**Fig. 5.3**).

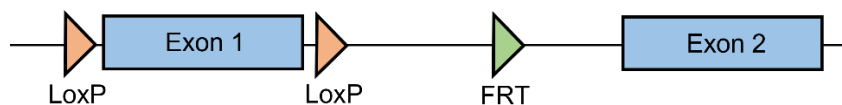


Figure 5.3 A schematic of the *Tcf7l2* knock-out allele. This shows loxP sites flanking exon 1 of the *Tcf7l2* gene with an FRT site in the adjacent intron. Adapted from Angus-Hill *et al.*, (2011).

Using the *Tcf7l2* mouse sequence on UCSC Genome Browser, the position of both the forward and reverse primers designed by Mohan *et al.*, (2022) were located. The forward primer was designed to be positioned in exon 1, and the reverse primer is positioned in intron 1 of *Tcf7l2*. The position of the forward primer indicates the possibility that the entirety of exon 1 is not deleted as this PCR would not work in a recombined mouse. It was therefore of even greater importance to create a detailed sequence map to clarify which part of exon 1 was deleted by recombination.

To develop a comprehensive construct map, a reverse primer was positioned inside of exon 2. The sequence of this primer can be found in Table 4.2. DNA was extracted from the splenic tissue of a *Tcf7l2*^(fl/fl) mouse, a WT mouse and a *Tcf7l2*^(?); *Pgk1-Cre*^(Pos) mouse with unknown recombination status. *Tcf7l2*^(fl/fl) mice were crossed with *Pgk1-Cre*^(Pos) mice to create a null allele.

For the WT mouse, a band at approximately 450bp was expected. This would cover the region between the forward primer inside exon 1 and reverse primer inside exon 2 without the presence of any loxP or FRT sites. The *Tcf7l2*^(fl/fl) mouse would produce a larger band, due to the presence of two loxP sites and an FRT site: a band of approximately 600bp was expected. If recombined, the DNA from this mouse would produce two bands, a WT band and a smaller band due to the deletion of a region in exon 1 (**Fig. 5.4**). At this stage, it was not known which region of exon 1 had been deleted and so it was difficult to predict a band size. This would be clarified through sequencing of PCR products.

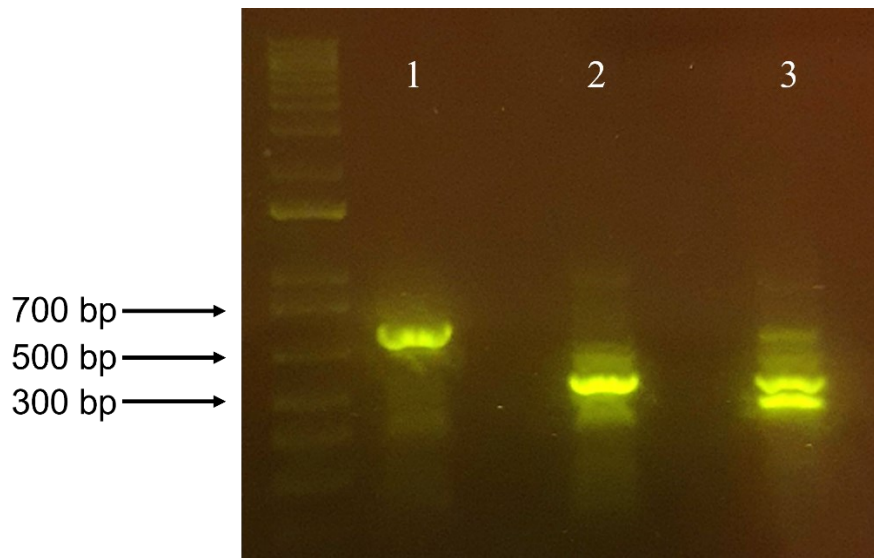


Figure 5.4 Splenic DNA from three different mice used to determine a sequence map of the *Tcf7l2* allele before and after recombination. 1) *Tcf7l2*^(fl/fl) 2) *Tcf7l2*^(WT) and 3) *Tcf7l2*^(*loxP*);*Pgk1-cre*^(Pos) tested with a forward primer inside of exon 1 and a reverse primer inside of exon 2 of *Tcf7l2*.

The results of this PCR revealed WT band at the expected length (450bp), a larger *Tcf7l2*^(fl/fl) band to account for additional loxp and FRT sites, and two bands in the *Tcf7l2*^(*loxP*);*Pgk1-cre*^(pos) mouse that had potentially undergone recombination, one band at the same height as the WT band and a smaller band, representing a partially deleted allele.

Each band was then cut out of the gel, the DNA was extracted and sequenced. Sequencing of the *Tcf7l2*^(fl/fl) DNA (lane 1) revealed the position of both loxp sites and an FRT site. Notably, the loxp sites did not flank the entirety of exon 1, but only a 249bp region of exon 1 (**Fig. 5.5**). This region was identified as part of exon 1 that included the translated region of exon 1 on analysis by the UCSC Genome Browser. Notably, the deleted region, flanked by the two loxp sites, incorporates the ATG start site of this translated region.

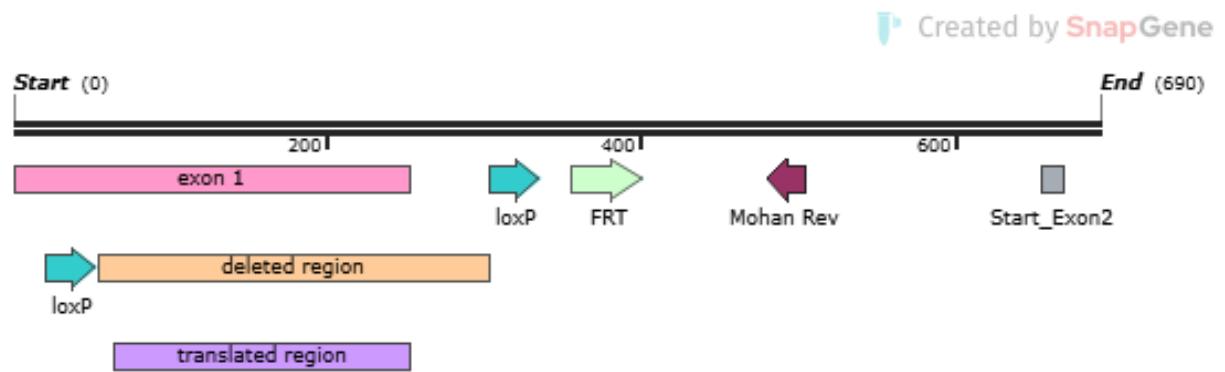
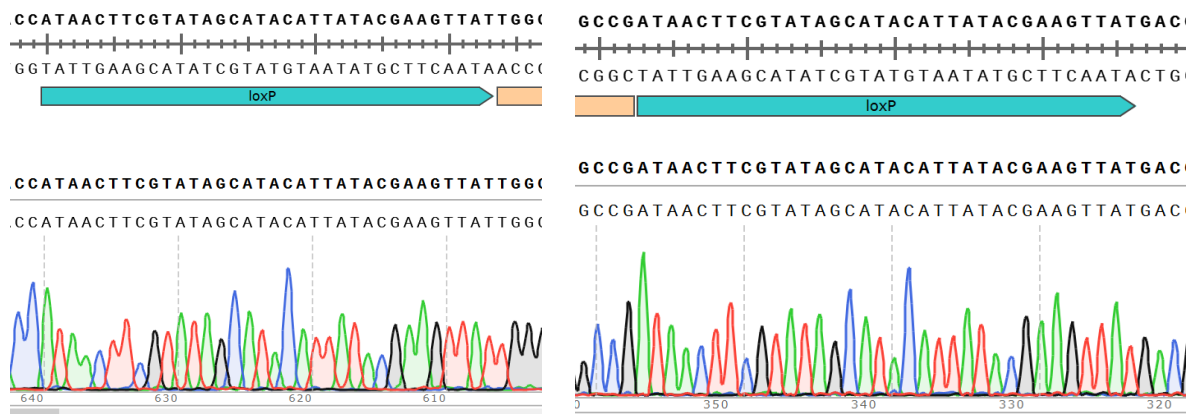


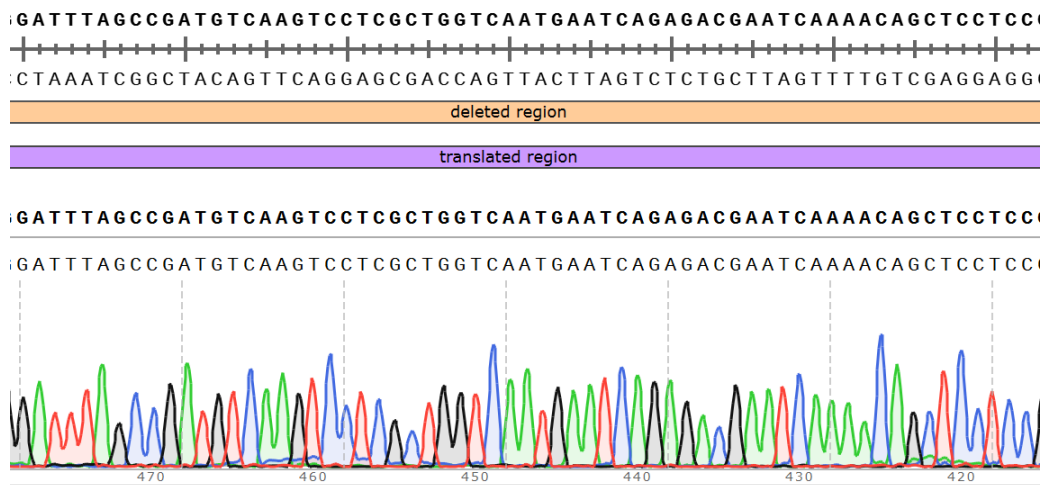
Figure 5.5 *Tcf7l2*^{fl/fl} allele schematic following sequencing of splenic DNA PCR products. This schematic is created using the DNA sequence produced using splenic DNA PCR products from a *Tcf7l2*^{fl/fl} mouse. It shows the 249bp region between the loxp sites that is deleted following recombination by Cre and the translated region (purple) of exon 1. It also shows the position of the reverse primer used by Mohan *et al.*, (2022) used to assess recombination. Created using SnapGene.

DNA sequencing electropherograms of both loxp sites, a section of the deleted/translated region and the FRT site can be seen below (**Fig. 5.6**).

A



B



C

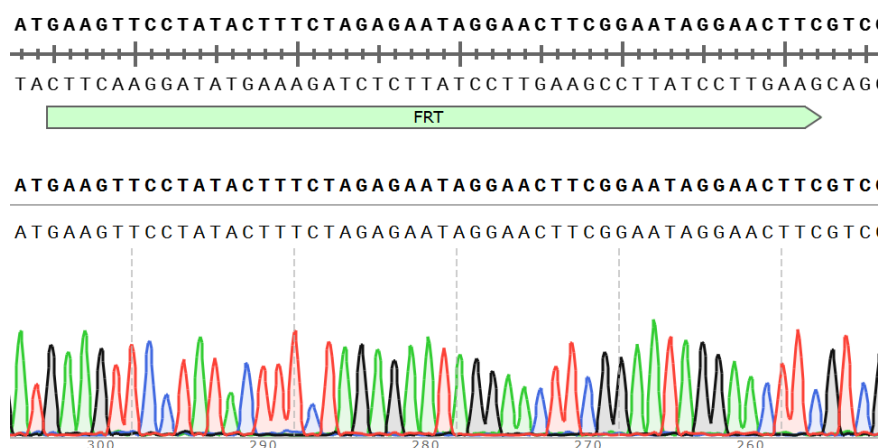


Figure 5.6 DNA sequence electropherograms of A) both loxp sites B) a section of the deleted/translated region of exon 1, and C) FRT site of *Tcf7l2*^(f/f) splenic DNA. Created using SnapGene.

The DNA sequence for the *Tcf7l2*^(fl/fl) mouse from the first loxp site to the start of exon 2 can be found below (Fig. 5.7).

ATAACTTCGTATAGCATACATTATACGAAGTTAT TGGGTGAAAAATGCCGCAGTGAACG
GCGGTGGAGGAGATGACCTAGGCGCTAACGACGAGCTGATCTCCTTCAAAGACGAAGGC
GAGCAGGAGGAGAAGAAGCTCGGAAAACTCCTCGGCGGAAAGGGATT TAGCCGATGTCA
AGTCCTCGCTGGTCAATGAATCAGAGACGAATCAAAACAGCTCCTCCGATTCCGAGGTA
GGAAGAGCCCCTCGGGCTGGCGGGTGGCTCCAGCCGGTTCCAGGGCCGATAACTTCGTA
TAGCATACATTATACGAAGTTATGACGTCCTCACAATCGATGAAGTTCCTATACTTTCTA
GAGAAATAGGAACTTCGGAATAGGAACTTCGTCTGACTCGACAGCGCGGCCGCGCCCTCA
CACTCGACCTTAAGTGTGAGGGCGCGCCGGAGAGCGGGGCTGGACGGCGAGCGAGGCG
AGGCGGGTCGGGCTTGCGGGCGGCGCGGACGGTGGCCCCGGGCTGCTGGCTGCGAAAAC
TGTAACCGTGTTTTTCTGCACCCCTTTCTCCGCCGACCCTCCCTTCCCGCACCCCCCTCGGC
GCGGCTATCGCCCCCTCGCCCTCCCCACCTCGCCCCCTCCGGGAA

Figure 5.7 *Tcf7l2*^(fl/fl) allele DNA sequence from first loxp site to the start of exon 2. Red font represents loxp sites. Blue font represents the deleted region. Green font represents the FRT site.

This sequencing provided clarification of the section of DNA that is deleted and also the positioning of the primers utilised by Mohan *et al.*, (2022) to test for recombination (Fig. 5.8).

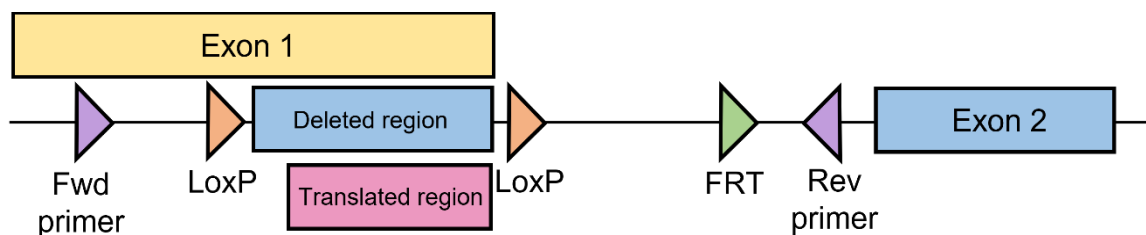


Figure 5.8 A more detailed schematic of the *Tcf7l2*^(fl/fl) allele clarified using sequenced splenic DNA PCR product.

Both the upper and lower bands of lane 3 in Figure 5.4 were also sequenced. The lower/smaller band of lane three showed an absence of the deleted region and only one loxp site as shown in Figures 5.5 and 5.9. The allele schematic, sequence electropherograms and DNA sequence of the recombined allele can be found below (Fig 5.9).

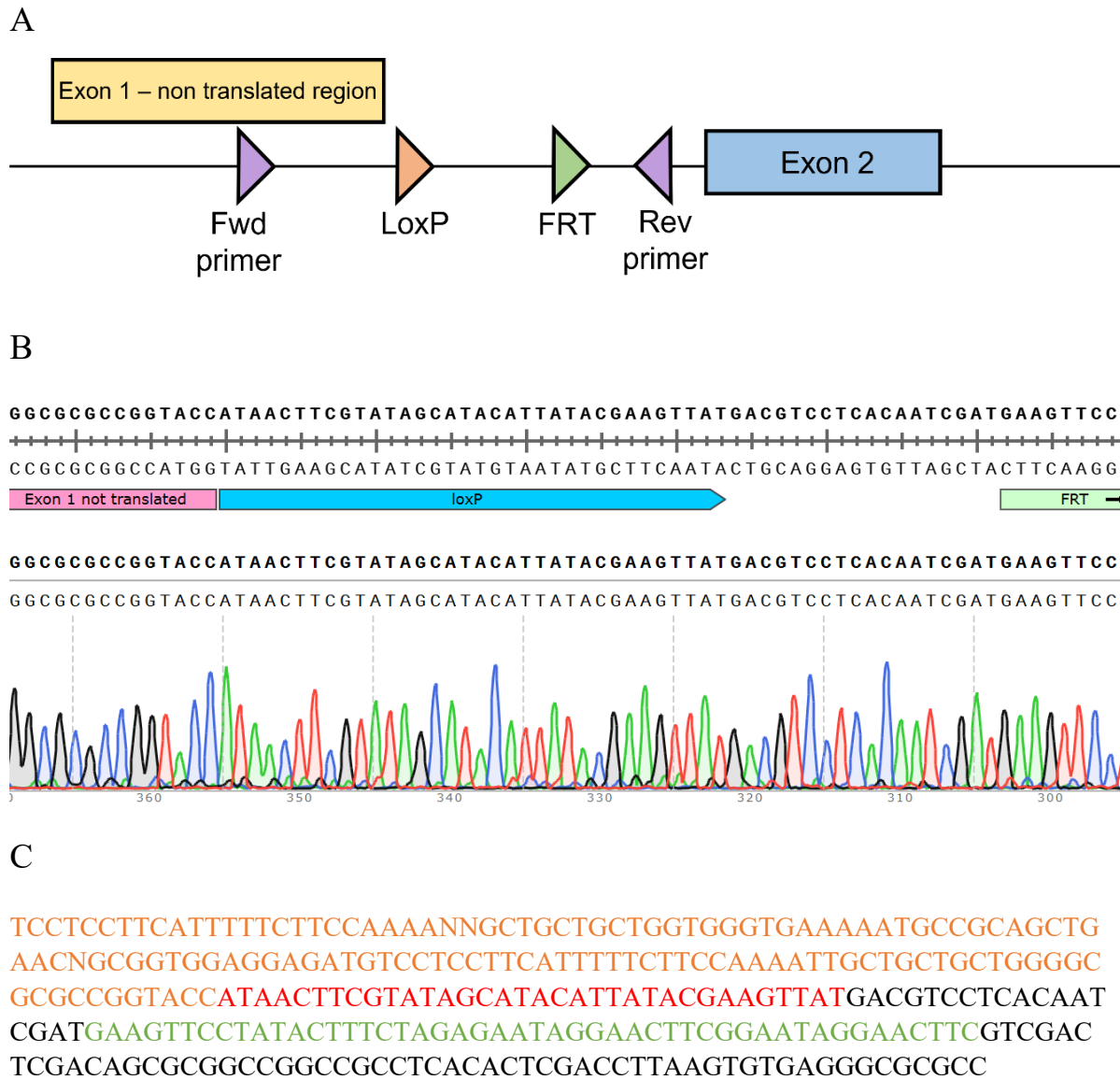


Figure 5.9 Structure of the *Tcf7l2*^(+/-);*Pgk1-cre*^(Pos) recombined allele based on sequencing of splenic DNA PCR products. A) Shows the *Tcf7l2* allele schematic with the absence of the deleted region. B) Shows the electropherogram spanning the non-deleted region of exon 1, the only remaining loxP site and the FRT site. C) DNA sequence: orange font represents the non-translated region of exon 1 that is not deleted. Red font represents the single remaining loxP site and green font is the FRT site.

Based on these results, it was possible to utilise the primers designed by Mohan *et al.*, (2022) to test all further mice for recombination. A WT *Tcf7l2*^(+/+) mouse would have a band at 361bp and a *Tcf7l2*^(+/-) would have two bands, one at 361bp and another at 227bp, representing the recombined allele (Figure 5.10).

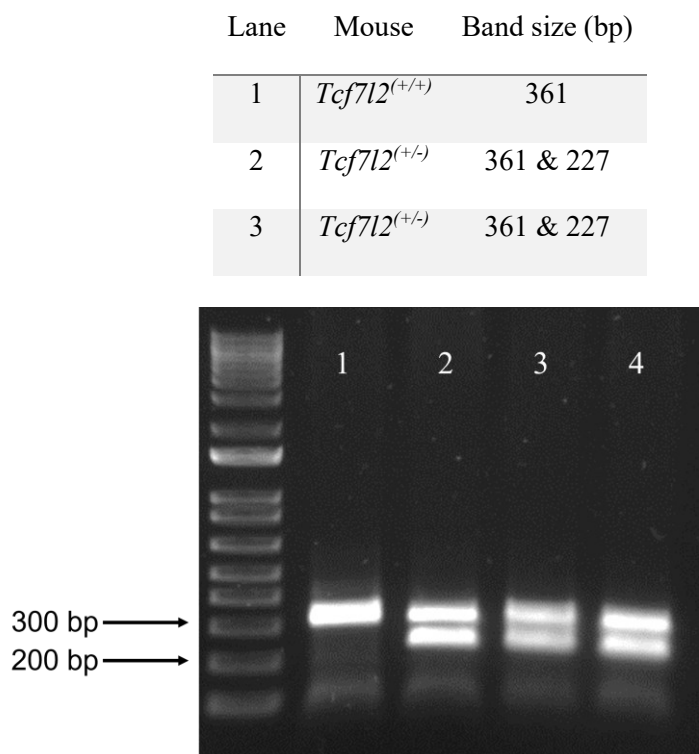


Figure 5.10 *Tcf7l2* allele genotyping assay for both *Tcf7l2*^(+/+) and *Tcf7l2*^(+/-) mice. Lane 1) *Tcf7l2*^(+/+) Lane 2) *Tcf7l2*^(+/-) Lane 3) *Tcf7l2*^(+/-) Lane 4) *Tcf7l2*^(+/-). Heterozygous knock-out mice present two bands, one with a length of 361 bp and another with a length of 227 bp.

5.4.2 Intestinal analysis

The number of tumours macroscopically identified in the small intestine and colorectum were assessed in several strains of mice. Tissue from control mice, including *Tcf7l2*^(fl/fl) mice, *Tcf7l2*^(+/fl), *Tcf7l2*^(+/+); *Pgkl-cre*^(pos) and *Tcf7l2*^(+/+); *Vill-cre*^(pos) were analysed after mice were aged to approximately one year. The strains, age and number of mice can be found in Table 5.3. Consistently, no macroscopic intestinal or other organ-specific tumours were identified in any of these mice. Lack of intestinal tumours in these tissues was confirmed through histopathological analysis. As previously found by Angus-Hill *et al.*, (2011) a homozygous knock-out of *Tcf7l2*, which in this case was *Tcf7l2*^(fl/fl); *Vill-cre*^(pos), appears to be lethal as over 20 crosses between *Tcf7l2*^(fl/+); *Vill-cre*^(pos) parent mice were carried out and a *Tcf7l2*^(fl/fl); *Vill-cre*^(pos) genotype was never obtained. 25% of the offspring should have had this genotype.

Mice with an *Apc*^(Min/+) (“multiple intestinal neoplasia”, p.Arg850Ter mutation) genotype commonly produce numerous adenomas within the intestinal tract and die after approximately 120 days (Moser, Pitot and Dove, 1990). Thus double-mutant mice, *Tcf7l2*^(+/-); *Apc*^(Min/+) with a heterozygous null allele for *Tcf7l2*, and control single-mutant *Tcf7l2*^(+/+); *Apc*^(Min/+) were aged to only approximately 100 days to avoid the generation of a severe phenotype (intestinal haemorrhage and obstruction due to multiple large intestinal tumours. Secondary controls were also aged to approximately 100 days

snips confirmed that the *Apc*^(Min/+) genotype of both mice was initially incorrect, most likely due to errors involving handlings of the earsnips. This re-categorised the two potential outlier mice into their correct genotypes. (**Fig. 5.12 & 5.13**).

A

<i>Apc</i> ^(Min/+)						<i>Tcf7l2</i> , recombination			
Lane	Mouse no. & tissue	Original genotype	Exp. band size (bp)	Corrected genotype	Obs. band size (bp)	Lane	Mouse no. & tissue	Band size (bp)	Genotype
1	#7 spleen	<i>Apc</i> ^(+/+)	550	<i>Apc</i> ^(Min/+)	300 & 550	13	7 spleen	227 & 361	<i>Tcf7l2</i> ^(+/-)
2	#7 LI	<i>Apc</i> ^(+/+)	550	<i>Apc</i> ^(Min/+)	300 & 550	14	7 LI	227 & 361	<i>Tcf7l2</i> ^(+/-)
3	#7 SI	<i>Apc</i> ^(+/+)	550	<i>Apc</i> ^(Min/+)	300 & 550	15	7 SI	227 & 361	<i>Tcf7l2</i> ^(+/-)
4	#20 spleen	<i>Apc</i> ^(Min/+)	300 & 550	<i>Apc</i> ^(+/+)	550	16	20 spleen	227 & 361	<i>Tcf7l2</i> ^(+/-)
5	#20 LI	<i>Apc</i> ^(Min/+)	300 & 550	<i>Apc</i> ^(+/+)	550	17	20 LI	227 & 361	<i>Tcf7l2</i> ^(+/-)
6	#20 SI	<i>Apc</i> ^(Min/+)	300 & 550	<i>Apc</i> ^(+/+)	550	18	20 SI	227 & 361	<i>Tcf7l2</i> ^(+/-)
7	#8 spleen	<i>Apc</i> ^(+/+)	550	<i>Apc</i> ^(+/+)	550	19	8 spleen	361	<i>Tcf7l2</i> ^(+/+)
8	#8 LI	<i>Apc</i> ^(+/+)	550	<i>Apc</i> ^(+/+)	550	20	8 LI	361	<i>Tcf7l2</i> ^(+/+)
9	#8 SI	<i>Apc</i> ^(+/+)	550	<i>Apc</i> ^(+/+)	550	21	8 SI	361	<i>Tcf7l2</i> ^(+/+)
10	#5 spleen	<i>Apc</i> ^(Min/+)	300 & 550	<i>Apc</i> ^(Min/+)	300 & 550	22	5 spleen	227 & 361	<i>Tcf7l2</i> ^(+/-)
11	#5 LI	<i>Apc</i> ^(Min/+)	300 & 550	<i>Apc</i> ^(Min/+)	300 & 550	23	5 LI	227 & 361	<i>Tcf7l2</i> ^(+/-)
12	#5 SI	<i>Apc</i> ^(Min/+)	300 & 550	<i>Apc</i> ^(Min/+)	300 & 550	24	5 SI	227 & 361	<i>Tcf7l2</i> ^(+/-)

B

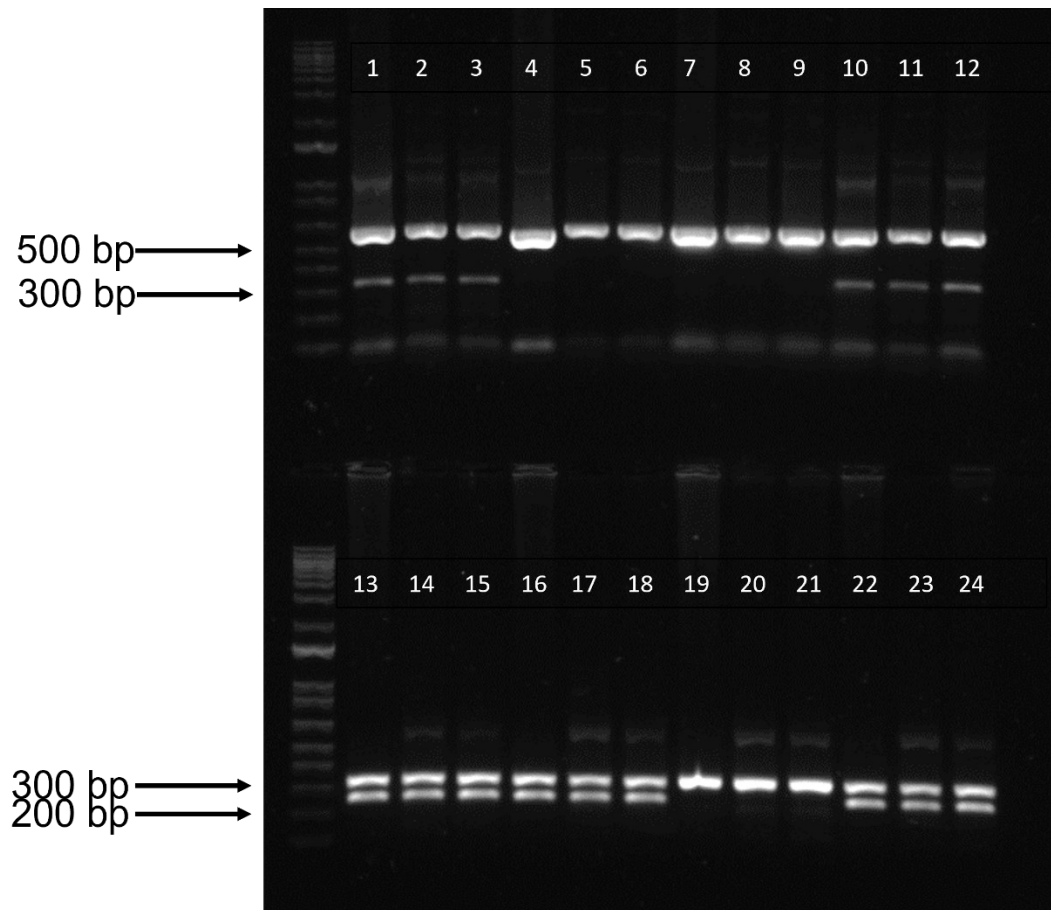


Figure 5.12 Genotyping assay results for the two potential outlier mice and controls using DNA from the spleen, large intestine and small intestine. A) Shows detailed information on sample, expected and observed band sizes and resultant murine genotypes. B) Shows the associated electrophoretic gel in which lanes 1-12 show the *Apc* genotype and lanes 13-24 show the *Tcf7l2* genotype. The results indicate that the initial *Apc*^(Min/+) genotypes of mouse 7 and mouse 20 were incorrect. These were corrected by genotyping assay PCR product band sizes in the electrophoretic gel above, recorded in the above table.

After determining the correct genotypes of the two incorrectly genotyped mice, the numbers of intestinal tumours harboured by each mouse were plotted (**Fig. 5.13**).

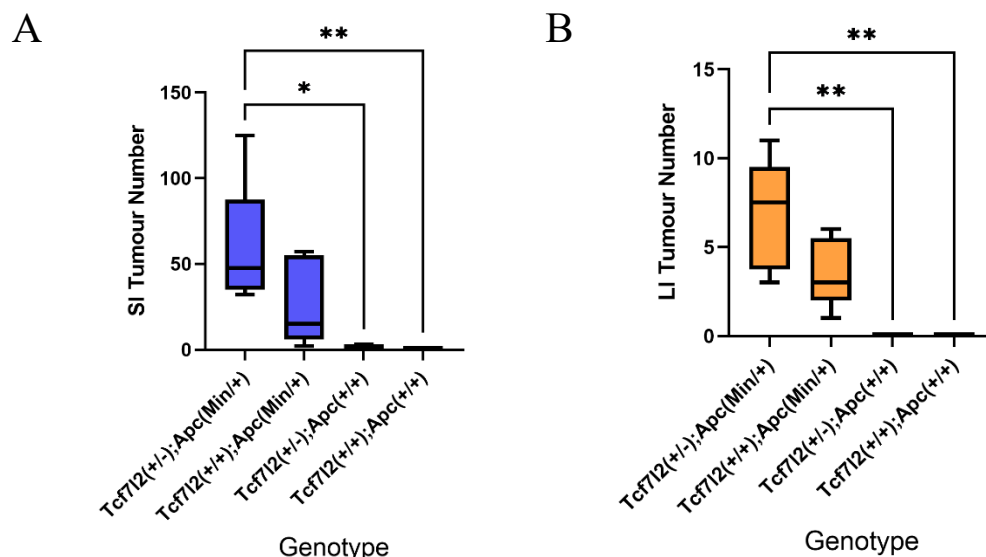
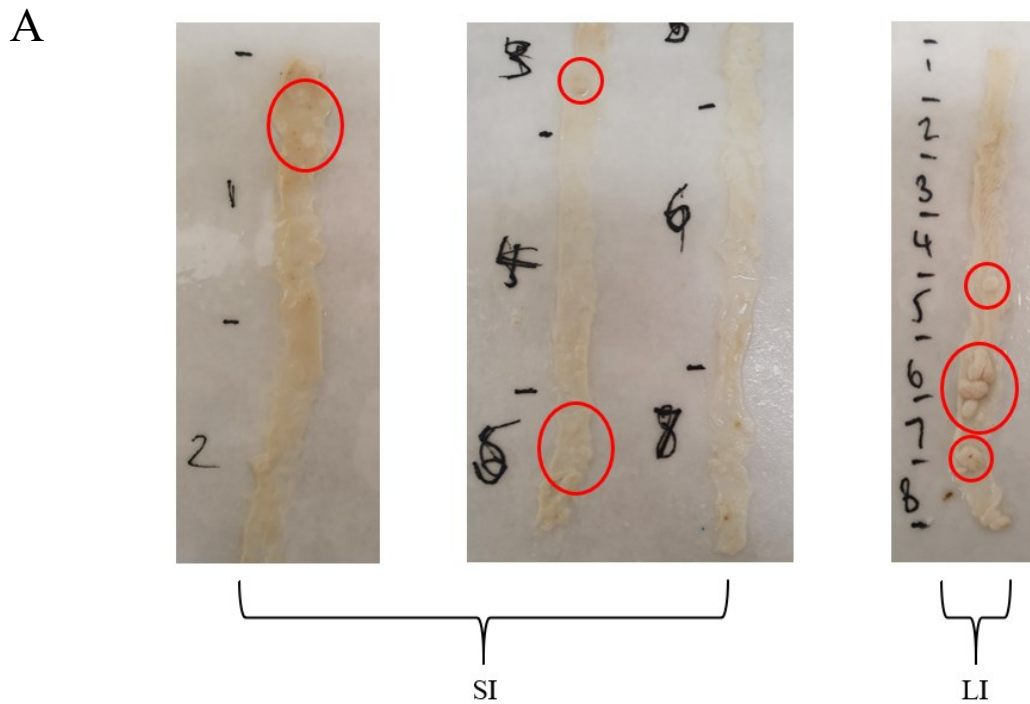


Figure 5.13 Box and whisker plots showing the number of tumours in the small and large intestines of *Tcf7l2*;*Apc* test and control mouse strains, and their statistical differences. A) Graph shows a significant difference between the number of small intestinal tumours between *Tcf7l2*^(+/-);*Apc*^(Min/+) mice and both *Tcf7l2*^(+/-);*Apc*^(+/+) and *Tcf7l2*^(+/-);*Apc*^(+/+). B) Graph shows a significant difference between the number of large intestinal tumours between *Tcf7l2*^(+/-);*Apc*^(Min/+) mice and both *Tcf7l2*^(+/-);*Apc*^(+/+) and *Tcf7l2*^(+/+);*Apc*^(+/+) (the outer limits of the boxes indicate the interquartile range. The median is shown by the horizontal lines and the whiskers represent the confidence intervals). * = $p < 0.05$, ** = $p < 0.01$. $n = 20$.

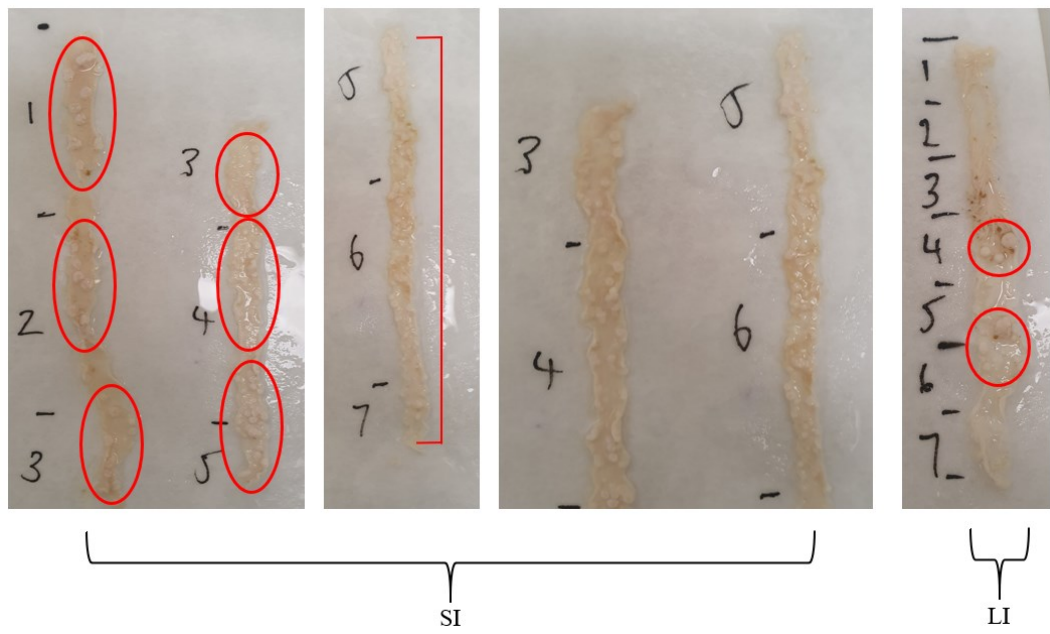
In both the small and the large intestine, the number of tumours in each of the four different genotypes of mice were not normally distributed ($P < 0.05$). Kruskal-Wallis tests followed by Dunn's multiple comparisons tests were therefore carried out to determine if there is a significant difference in tumour in both the small and the large intestine between groups.

A Kruskal-Wallis test indicated that there were significant differences in the numbers of tumours in the small intestine between genotypes ($H = 14.71$, $p = 0.002$). Similarly, a Kruskal-Wallis test indicated a significant difference in the numbers of tumours in the large intestine between genotypes ($H = 16.75$, $p < 0.001$). A Dunn's multiple comparison test then clarified which genotypes were significantly different in tumour number. In the small intestine, there was no difference between the main test mice *Tcf7l2*^(+/-);*Apc*^(Min/+) and control *Tcf7l2*^(+/+);*Apc*^(Min/+) mice ($p > 0.999$). Similarly, in the large intestine, there was also no difference between the main test mice *Tcf7l2*^(+/-);*Apc*^(Min/+) and control *Tcf7l2*^(+/+);*Apc*^(Min/+) mice ($p > 0.999$). A significant difference was identified in the number of tumours in the small intestine between both *Tcf7l2*^(+/-);*Apc*^(Min/+) and *Tcf7l2*^(+/-);*Apc*^(+/+) genotypes ($p = 0.029$). A significant difference was also found between *Tcf7l2*^(+/-);*Apc*^(Min/+) and *Tcf7l2*^(+/+);*Apc*^(+/+) ($p = 0.005$). Significant differences were also found in the large intestine: a significant difference was identified between the numbers of tumours in the large intestine between *Tcf7l2*^(+/-);*Apc*^(Min/+) and *Tcf7l2*^(+/-);*Apc*^(+/+) mice ($p = 0.004$) and also between *Tcf7l2*^(+/-);*Apc*^(Min/+) and *Tcf7l2*^(+/+);*Apc*^(+/+) mice ($p = 0.008$).

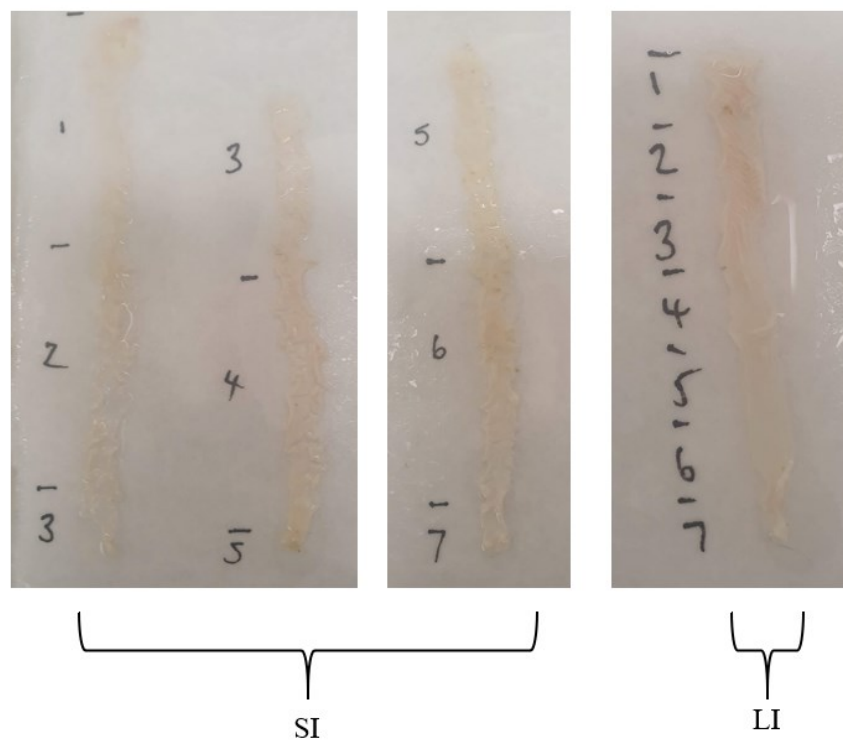
A T-test was also carried out between $Tcf712^{(+/-)};Apc^{(Min/+)}$ and $Tcf712^{(+/-)};Apc^{(Min/+)}$ mice alone, following a Shapiro-Wilk confirmation of normal distribution ($p > 0.05$), to determine if there was a significant difference between the number of tumours in the entirety of the small intestine when only the two main test and control groups were incorporated into the analysis. The results indicated that there was no significant difference in tumour number in the small intestine between $Tcf712^{(+/-)};Apc^{(Min/+)}$ and $Tcf712^{(+/-)};Apc^{(Min/+)}$ mice ($T = 1.735$, $p = 0.117$). Similarly, when the same test was carried out to investigate tumour number in the entirety of the large intestine, following a Shapiro-Wilk test for normality ($p > 0.05$). No significant difference in tumour number was found between $Tcf712^{(+/-)};Apc^{(Min/+)}$ and $Tcf712^{(+/-)};Apc^{(Min/+)}$, although this bordered on significance ($T = 2.153$, $p = 0.0597$) (**Fig. 5.13**). Images of the SI and LI of different strains of mice can be seen in (**Fig. 5.14**).



B



C



D

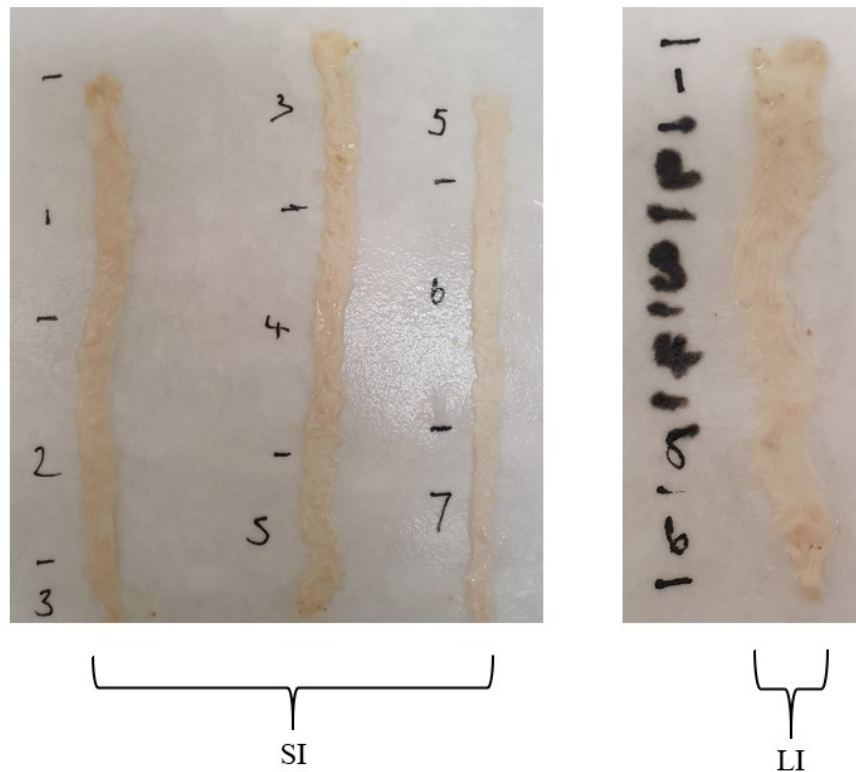


Figure 5.14 Images of intestinal tumours in different experimental strains of mice. A) $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ B) $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ C) $Tcf7l2^{(+/-)};Apc^{(+/+)}$ D) $Tcf7l2^{(+/+)};Apc^{(+/+)}$. SI = small intestine. LI = Large intestine. Red indicates the presence of tumours. Sections in the SI are 5cm in length. Sections in the LI are 1cm in length.

The distribution of tumours across sections of the small intestines (SI) between each genotype were then analysed. The SI of each mouse was divided and labelled into 5cm sections, section 1 at the proximal end of the SI and section 8 at the distal end (**Fig. 5.15**).

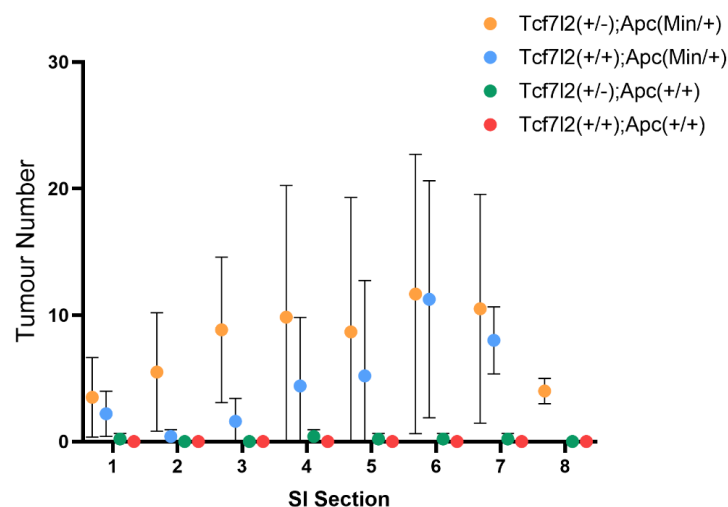
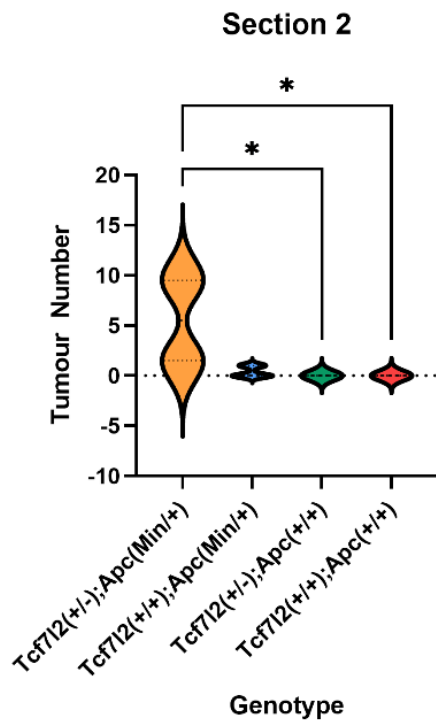


Figure 5.15 Tumour number by SI section. Shows the mean with standard deviation of tumour number per 5cm section of SI.

Initially, A Kruskal-Wallis test was carried out. This determined that there was a significant difference between the number of tumours in section 1 of the small intestine depending on the genotype ($H = 10.39$, $p = 0.0155$). A Dunn's multiple comparisons test, in which the mean rank of each group is compared with mean rank of each of the other groups, with adjusted p-values, then contradicted this initial finding by suggesting no significant difference in tumour number between each of the groups in section 1. Following a Shapiro-Wilk test for normality ($p > 0.05$), the difference in number of tumours in section 1 between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ alone were investigated via T-test. The results also suggested that there was no significant difference in tumour number in section 1 of the small intestine between the main test and control mice, $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ ($T = 0.816$, $p = 0.436$).

In section 2, an initial Kruskal-Wallis identified a significant difference between the number of tumours in different genotypes ($H = 12.00$, $p = 0.007$). However, a Dunn's test identified a significant difference in tumour number in section 2 of the small intestine between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and both $Tcf7l2^{(+/-)};Apc^{(+/+)}$ ($p = 0.016$) and $Tcf7l2^{(+/+)};Apc^{(+/+)}$ ($p = 0.029$). When subsequently carrying out a Mann-Whitney U test, following Shapiro-Wilk, a significant difference was identified between the number of tumours in section 2 of the small intestine between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice. $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ had significantly more tumours in section 2 of the small intestine than $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ controls (**Fig. 5.16**).

A



B

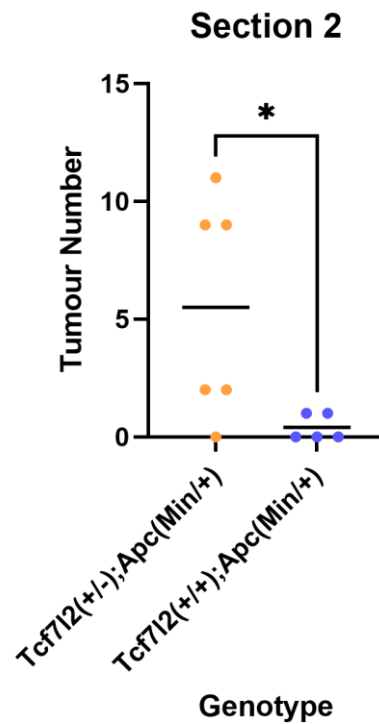
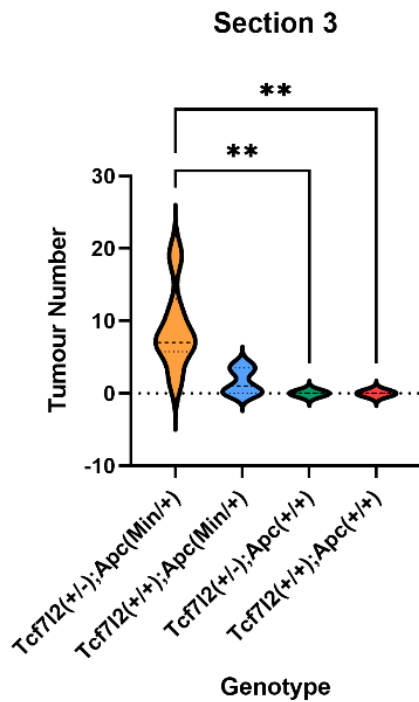


Figure 5.16 Graphs showing the number of tumours in section 2 of the small intestine according to genotype. A) Graph shows the result of the Dunn's multiple comparisons test between all genotypes. B) Shows the result of the Mann-Whitney U test between the main test mice and control mice, when only incorporating the main test and control mice: test mice $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and control mice $Tcf7l2^{(+/+)};Apc^{(Min/+)}$. Horizontal bar = Mean. * = $p < 0.05$. $n = 20$.

When investigating the number of tumours in section 3 of the small intestine, A Kruskal-Wallis plus Dunn's multiple comparison test indicated a significant difference in tumour number between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and both $Tcf7l2^{(+/-)};Apc^{(+/+)}$ ($p = 0.004$) and $Tcf7l2^{(+/+)};Apc^{(+/+)}$ genotypes ($p = 0.008$). Following this, Shapiro-Wilk test and subsequent T-test were carried out utilising only the main test mice and control mice, $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$. The results indicated a significantly increased number of tumours in section 3 of the small intestine of the test $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ mice than the $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ controls (**Fig. 5.17**).

A



B

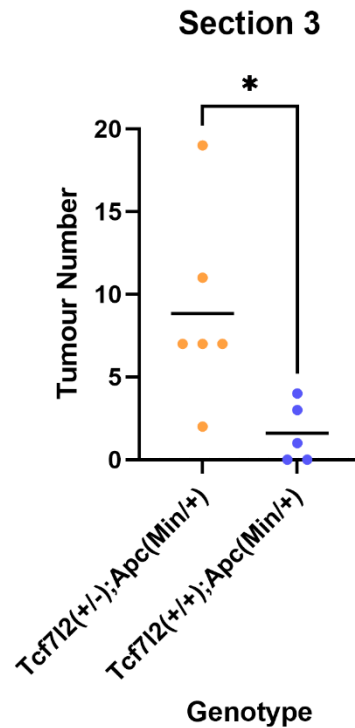


Figure 5.17 Graphs showing the number of tumours in section 3 of the small intestine according to genotype. A) Graph shows the result of the Dunn's multiple comparisons test between all genotypes. B) Graph shows the result of the Mann-Whitney U test between the main test mice and control mice, when only incorporating the main test and control mice: $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$. Horizontal bar = Mean. * = $p < 0.05$, ** = $p < 0.01$. $n = 20$.

The only groups with significantly different numbers of tumours in section 4 were $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(+/+)}$ (Fig. 5.18). A Shapiro-Wilk followed by Mann-Whitney U test indicated no significant difference between the number of tumours between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ than the $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice for section 4 of the small intestine.

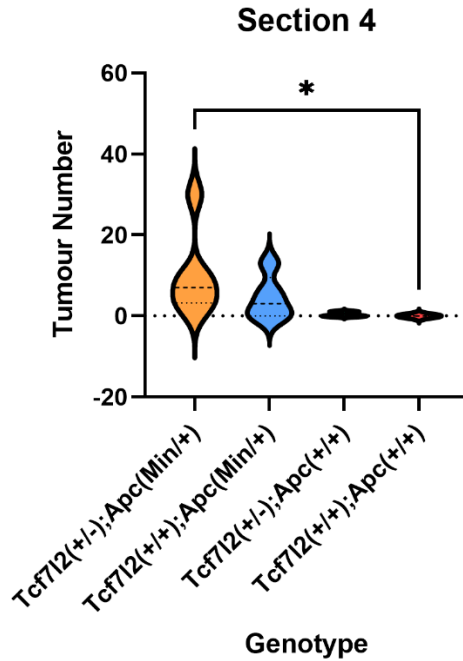


Figure 5.18 Graph showing the number of tumours in section 2 of the small intestine according to genotype. Graph shows the result of the Dunn's multiple comparisons test between all genotypes. * = $p < 0.05$. $n = 20$.

No significant difference in tumour number in section 5 of the small intestine was identified between any groups via Kruskal-Wallis test ($H = 6.398$, $p = 0.094$). This was validated via Shapiro-Wilk test and subsequent Mann-Whitney U test when investigating differences in tumour number between *Tcf7l2*^(+/-);*Apc*^(Min/+) and *Tcf7l2*^(+/+);*Apc*^(Min/+) mice ($U = 14.5$, $p = 0.987$).

The Kruskal-Wallis test identified a significant difference in tumour number between groups in section 6 of the small intestine ($H = 12.01$, $p = 0.007$), but the Dunn's multiple comparison's test showed no significant difference between the groups. This was also validated by a Shapiro-Wilk test followed by a T-test which showed no significant difference between the number of tumours in *Tcf7l2*^(+/-);*Apc*^(Min/+) and *Tcf7l2*^(+/+);*Apc*^(Min/+) mice ($T = 0.062$, $p = 0.952$) for section 6 of the small intestine.

A Kruskal-Wallis test and Dunn's multiple comparison test identified a significant difference in tumour number in section 7 of the small intestine between *Tcf7l2*^(+/-);*Apc*^(Min/+) and both *Tcf7l2*^(+/-);*Apc*^(+/+) ($p = 0.042$) and *Tcf7l2*^(+/+);*Apc*^(+/+) ($p = 0.027$). No significant difference between *Tcf7l2*^(+/-);*Apc*^(Min/+) and *Tcf7l2*^(+/+);*Apc*^(Min/+) mice was observed via Shapiro-Wilk test and T-test ($T = 0.454$, $p = 0.663$) (Fig. 5.19).

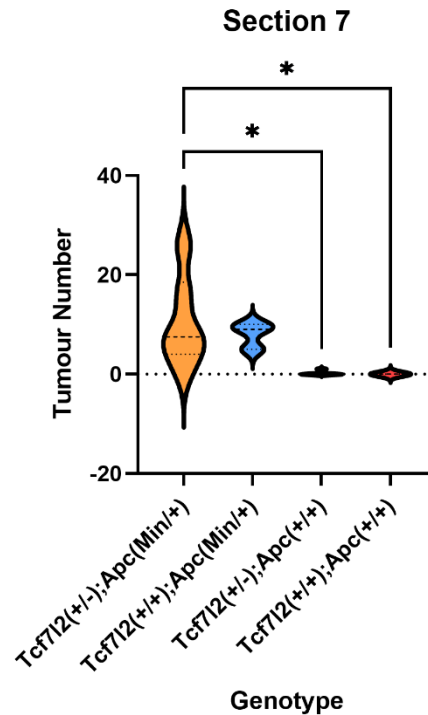


Figure 5.19 Graph showing the number of tumours in section 7 of the small intestine according to genotype. Graph shows the result of the Dunn's multiple comparisons test between all genotypes. * = $p < 0.05$. $n = 20$.

Not enough data points were available to test differences in tumour number for section 8 of the small intestine (**Fig. 5.15**).

These results suggest a significant increase in tumour number in proximal regions of the small intestine with the addition of a heterozygous *Tcf7l2* mutation to an *Apc*^(*Min*/+) murine line.

Tumour volume was also analysed in each section in addition to tumour number for the purpose of incorporating tumour size. In this case, tumour length (mm) was multiplied by both tumour width (mm) and height (1mm) (**Fig. 5.20**).

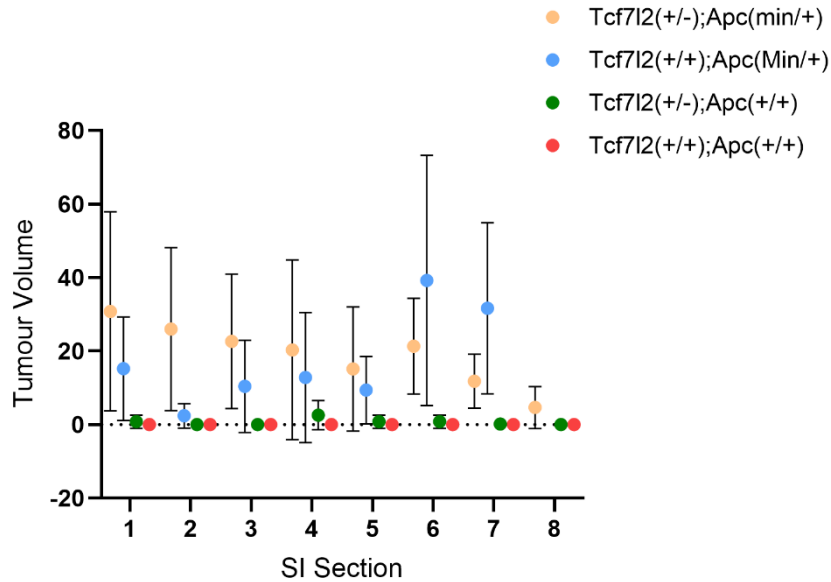
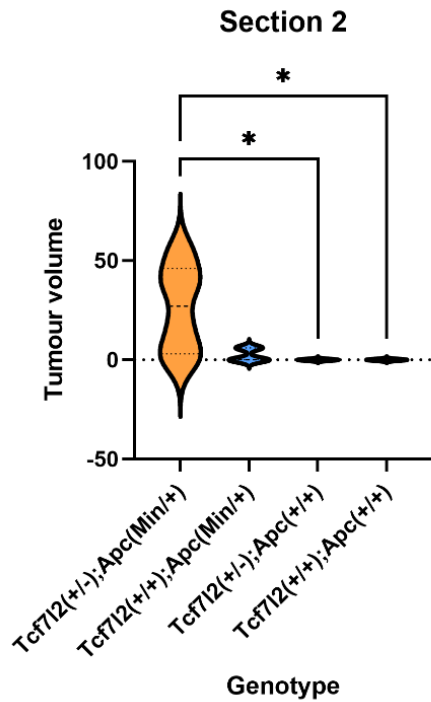


Figure 5.20 Tumour volume by SI section. Graph shows the mean and standard deviation of tumour volume for each 5cm SI section, categorised by murine strain

In section 1, a Kruskal-Wallis test identified a significant difference in tumour volume ($H = 10.46$, $p = 0.015$). However, a subsequent Dunn's multiple comparisons test did not identify any significant differences between the tumour volume of each genotype. A Shapiro-Wilk test followed by a T-test also indicated no significant difference in tumour volume between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice in section 1 of the small intestine.

In section 2, a Kruskal-Wallis test identified a significant difference between tumour volumes in mice of different genotypes in section 2. A Dunn's multiple comparisons test showed that there was a significant difference between tumour volumes in section 2 between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and both $Tcf7l2^{(+/-)};Apc^{(+/+)}$ ($p = 0.023$) and $Tcf7l2^{(+/+)};Apc^{(+/+)}$ ($p = 0.040$). A Shapiro-Wilk test followed by a Mann-Whitney U test indicated that a significantly higher number of tumours were observed in section 2 of the small intestine for $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ compared with $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice ($U = 2.340$, $p = 0.044$) (**Fig. 5.21**).

A



B

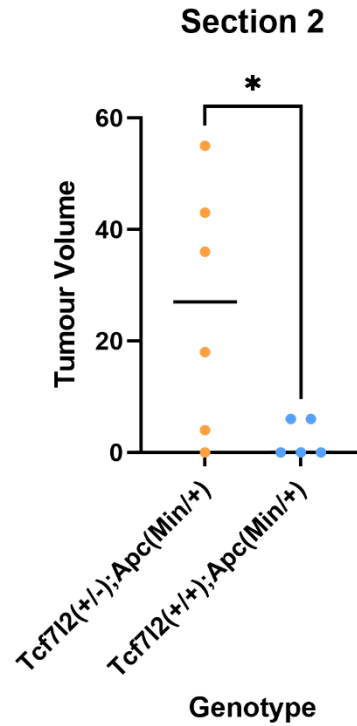


Figure 5.21 An analysis of tumour volume in section 2 of the small intestine according to genotype. A) Graph shows the result of the Dunn's multiple comparisons test between all genotypes. B) Graph shows the result of the Mann-Whitney U test between the main test and control mice, when only incorporating the main test and control mice: $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/-)};Apc^{(Min/+)}$. Horizontal bar = Mean. * = $p < 0.05$. $n = 20$.

Similarly, when analysing tumour volume in section 3 of the small intestine, a Kruskal-Wallis test identified a significant difference in volume depending on strain ($H = 13.45$, $p = 0.0038$). A Dunn's multiple comparisons test identified significant differences in tumour volume between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and both $Tcf7l2^{(+/-)};Apc^{(+/+)}$ ($p = 0.010$) and $Tcf7l2^{(+/-)};Apc^{(+/+)}$ ($p = 0.018$) (**Fig 5.22**). No significant difference in tumour volume was found between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ mice in section 3 of the small intestine ($U = 7$, $p = 0.160$).

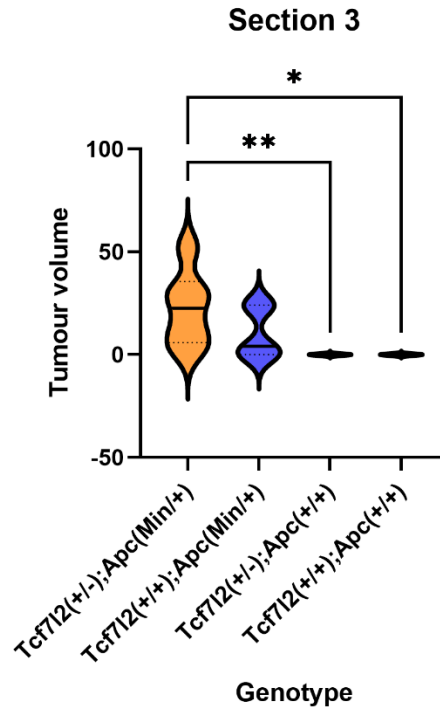


Figure 5.22 Graph showing tumour volume in section 3 of the small intestine according to genotype. Graph shows the result of the Dunn's multiple comparisons test between all genotypes. * = $p < 0.05$, ** = $p < 0.01$. $n = 20$.

No significant difference was found in tumour volume between mice strains in section 4 of the small intestine via Kruskal-Wallis ($H = 7.31$, $p = 0.63$). There was also no significance difference in tumour volume found between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice when testing by Shapiro-Wilk test followed by a T-test ($T = 0.573$, $p = 0.581$).

No significant difference in tumour volume was observed between strains in section 5 of the small intestine when testing by Kruskal-Wallis test ($H = 6.42$, $p = 0.093$). There was also no significant difference in tumour volume found between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice when testing via Shapiro-Wilk followed by a Mann-Whitney U test ($U = 13.50$, $p = 0.805$).

Interestingly, a Kruskal-Wallis test identified a significant difference in tumour volume between strains in section 6 of the SI ($H = 14.76$, $p = 0.002$). A Dunn's multiple comparisons test identified a significant difference in tumour volume of section 6 between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(+/+)}$ ($p = 0.040$) strains. Notably, this test also identified a significant difference in tumour volume between $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ and both $Tcf7l2^{(+/-)};Apc^{(+/+)}$ ($p = 0.040$) and $Tcf7l2^{(+/+)};Apc^{(+/+)}$ ($p = 0.030$) (**Fig. 5.23**). There was no significant difference in tumour volume between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice when testing via Shapiro-Wilk followed by a T-test ($T = 1.194$, $p = 0.267$).

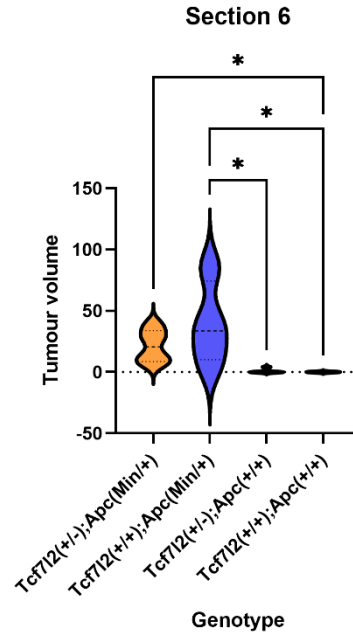


Figure 5.23 Graph showing tumour volume in section 6 of the small intestine according to genotype. Graph shows the result of the Dunn's multiple comparisons test between all genotypes. * = $p < 0.05$. $n = 20$.

A Kruskal-Wallis test also showed a significant difference in tumour volume depending on murine strain in section 7 of the SI. A Dunn's multiple comparisons test identified significant differences between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and both $Tcf7l2^{(+/-)};Apc^{(+/+)}$ ($p = 0.040$) and $Tcf7l2^{(+/+)};Apc^{(+/+)}$ ($p = 0.030$) (Fig. 5.24). No significant difference in tumour volume was identified between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice when using a Shapiro-Wilk, followed by a T-test ($T = 2.020$, $p = 0.083$).

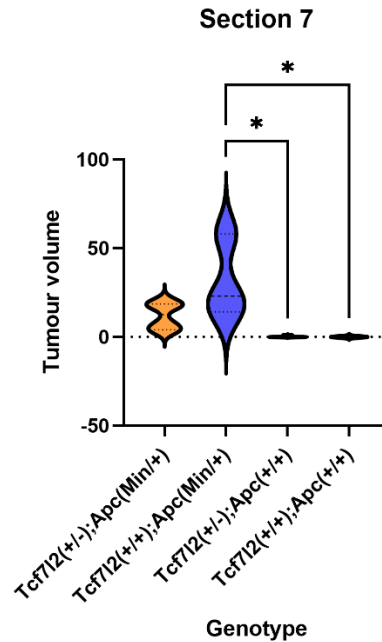
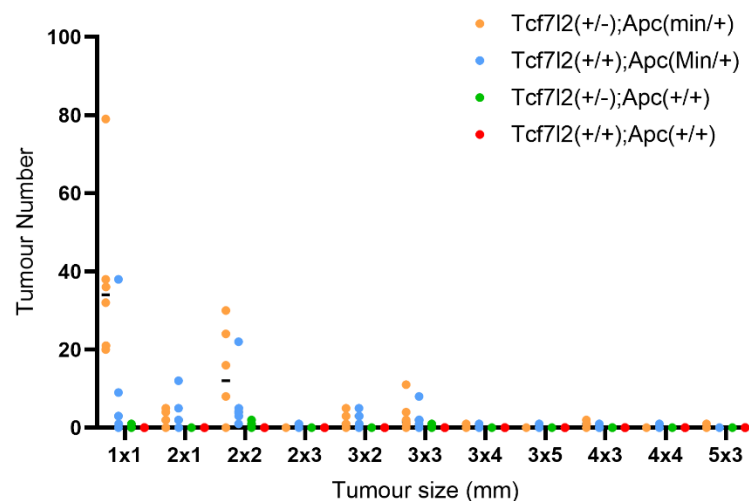


Figure 5.24 Graph showing tumour volume in section 7 of the small intestine according to genotype. Graph shows the result of the Dunn's multiple comparisons test between all genotypes. * = $p < 0.05$. $n = 20$.

Not enough data points were available to test difference in tumour volume between strains for section 8 of the SI.

These results suggest a significant increase in tumour volume in proximal regions of the small intestine with the addition of a heterozygous *Tcf7l2* mutation to an *Apc*^(Min/+) murine line.



5.25 Distribution of tumour sizes across different genotypes. Graph shows the number of tumours of varying sizes (length x width) across four different genotypes: *Tcf7l2*^(+/-);*Apc*^(Min/+), *Tcf7l2*^(+/+);*Apc*^(Min/+), *Tcf7l2*^(+/-);*Apc*^(+/+) and *Tcf7l2*^(+/+);*Apc*^(+/+).

A plot of the tumour size distribution is suggestive that $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ mice have a greater number of 1x1 mm tumours than any other murine strain (**Fig. 5.25**). A subsequent Shapiro-Wilk test and Mann-Whitney U test determined that there was no significant difference in the number of 1x1 tumours between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice, albeit at borderline significance ($U = 4.500$, $p = 0.0563$).

The distribution of the number of the tumours of the large intestine of all strains was then analysed (**Fig. 5.26**).

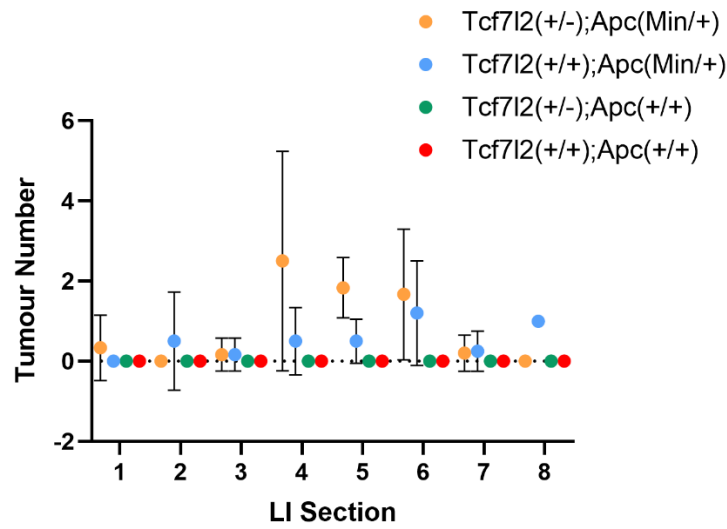


Figure 5.26 The distribution of tumours in the large intestines of multiple murine strains. The large intestines of each mouse were divided and labelled into 1cm sections. The tumours in each section were counted. This graph shows mean number of tumours per section for each genotype, plus standard deviation as whiskers.

Using a Kruskal-Wallis test, no significant differences in tumour numbers between strains were found in sections 1, 2, 3 and 7 of the large intestines. Further to this, no significant difference in tumour number was identified in any of the aforementioned regions when testing the differences between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice alone by multiple Shapiro-Wilk tests and Mann-Whitney U tests ($p > 0.05$). Section 8 lacked sufficient data points for analysis. A significant difference in tumour number based on genotype was identified in section 4 of the large intestine by Kruskal-Wallis test ($H = 7.926$, $p = 0.0476$). A subsequent Dunn's multiple comparisons test showed no significant differences between any of the strains.

A Kruskal-Wallis test showed a significant difference in tumour numbers dependent on murine strain in section 5 of the large intestine ($H = 15.26$, $p = 0.002$). A Dunn's multiple comparison test identified a significant difference in tumour numbers in section 5 between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and both $Tcf7l2^{(+/-)};Apc^{(+/+)}$ ($p = 0.004$) and $Tcf7l2^{(+/+)};Apc^{(+/+)}$ ($p = 0.009$). A subsequent Mann-Whitney U test was utilised to determine if there was a significant difference in tumour number in section 5 between

$Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice; the results indicated a significantly increased number of tumours in $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ mice in section 5 of the large intestine ($U = 3$, $p = 0.033$) (**Fig. 5.27**).

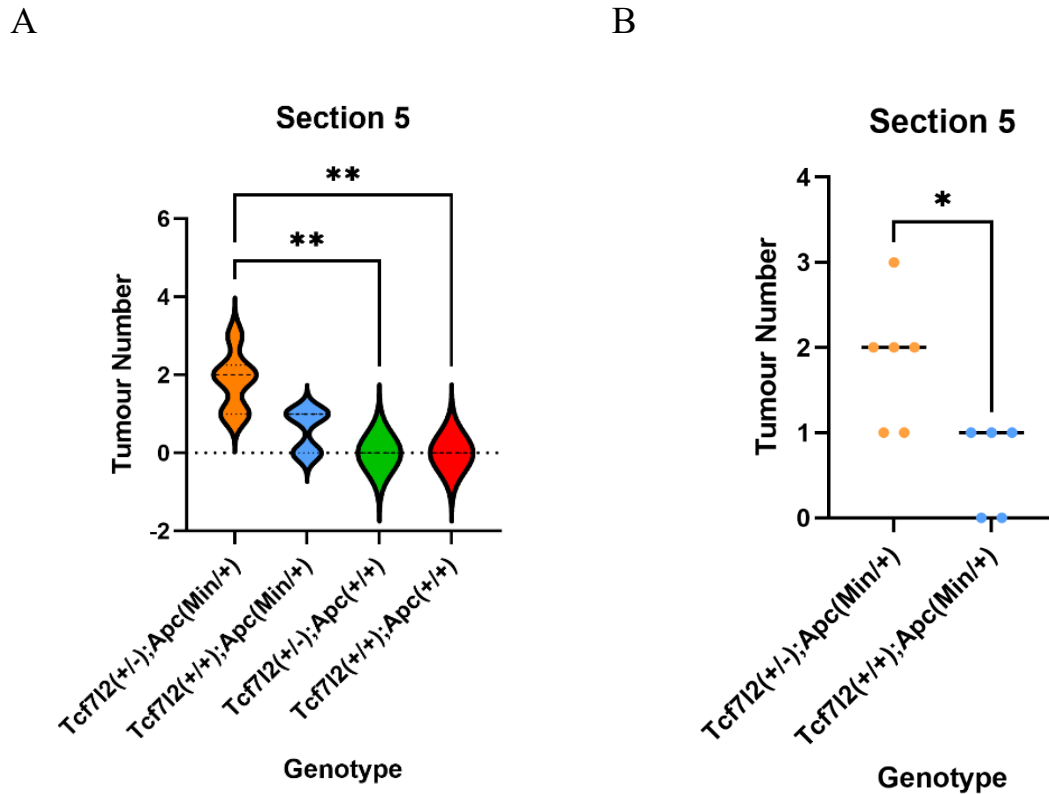


Figure 5.27 Graphs showing the number of tumours in section 5 of the large intestine according to genotype. A) Graph shows the result of the Dunn's multiple comparisons test between all genotypes. B) Graph shows the result of the Mann-Whitney U test between the main test mice and control mice, when only incorporating the main test and control mice: $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$. Horizontal bar = Mean. * = $p < 0.05$. ** = $p < 0.01$. $n = 20$.

In section 6 of the large intestine, the Kruskal-Wallis test also identified a significant difference in tumour numbers based on genotype, but no significant differences between groups were found using a Dunn's multiple comparison test. In addition, no significant difference in tumour number was identified in section 6 of the large intestine between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice when carrying out a Shapiro-Wilk test for normality followed by a T-test ($T = 0.171$, 0.869).

Tumour volume per section of large intestine was then analysed (**Fig. 5.28**).

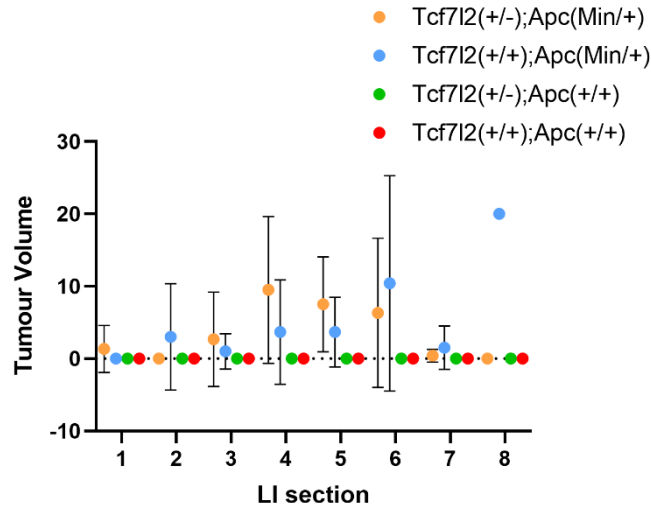


Figure 5.28 Tumour volume distribution across large intestine (LI) sections in different genotypes. This graph shows the mean and standard deviation (whiskers) of the distribution of tumour volumes across different 1cm sections of the large intestine in four genotypes: $Tcf7l2^{(+/-)};Apc^{(Min/+)}$, $Tcf7l2^{(+/+)};Apc^{(Min/+)}$, $Tcf7l2^{(+/-)};Apc^{(+/+)}$ and $Tcf7l2^{(+/+)};Apc^{(+/+)}$.

Using a Kruskal-Wallis test, no significant difference in tumour volume in different strains was observed in sections 1, 2, 3, 4, 7 and 8. In section 6, a significant difference in tumour volume was identified dependent on strain ($H = 9.442$, $p = 0.024$). A subsequent Dunn's multiple comparison test identified no significant difference between groups. Notably, in section 5 a significant difference in tumour volume was identified through Kruskal-Wallis testing ($H = 12.62$, $p = 0.002$). A subsequent Dunn's multiple comparison test identified a significant difference in tumour volume between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and both $Tcf7l2^{(+/-)};Apc^{(+/+)}$ ($p = 0.002$) and $Tcf7l2^{(+/+)};Apc^{(+/+)}$ ($p = 0.029$). Multiple Shapiro-Wilk tests and Mann-Whitney U/T-tests were carried out to determine if there was a significant difference in tumour volume between $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ and $Tcf7l2^{(+/+)};Apc^{(Min/+)}$ mice in any sections in the large intestines, but no significant difference was identified.

5.4.3 *Tcf7l2* RNA expression in $Tcf7l2^{(+/-)}$ mice.

In order to test the expression of *Tcf7l2* in both $Tcf7l2^{(+/-)}$ and $Tcf7l2^{(+/+)};$ mice, quantitative polymerase chain reaction (qPCR) assays were performed on RNA derived from splenic, large intestinal and small intestinal tissue, in triplicate. Tissue was derived from three mice of genotype $Tcf7l2^{(+/+)};Apc^{(Min/+)}$, three mice of genotype $Tcf7l2^{(+/+)};Apc^{(+/+)}$ and three $Tcf7l2^{(+/-)};Apc^{(Min/+)}$ mice. Initially, a standard Taqman *Tcf7l2* probe, spanning the boundary between exons 5 and 6 was obtained from ThermoFisher and utilised in a qPCR assay. A Mann-Whitney U test showed no significant difference in the RNA expression of *Tcf7l2* between $Tcf7l2^{(+/-)}$ and $Tcf7l2^{(+/+)};$ splenic tissue, relative to *Gapdh* ($U = 4$, $p = 0.643$) (Fig. 5.29).

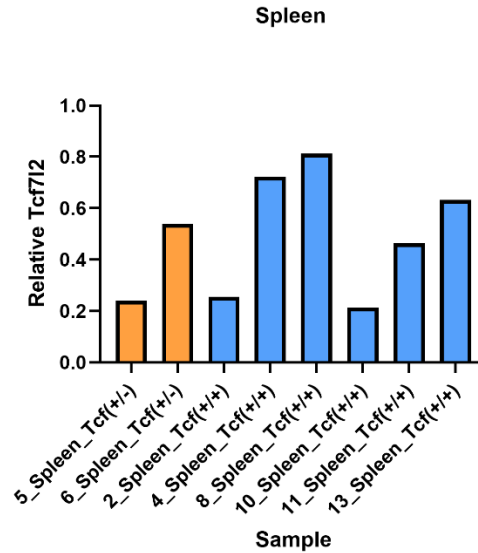


Figure 5.29 Analysis of *Tcf7l2* RNA expression in splenic tissue. Bar chart shows expression levels of *Tcf7l2* relative to *Gapdh* in both $Tcf7l2^{+/-}$ (orange) and $Tcf7l2^{+/+}$ (blue) mice.

The same probe was used to test *Tcf7l2* expression levels in large intestinal tissue of both $Tcf7l2^{+/-}$ and $Tcf7l2^{+/+}$ mice. Again, no significant difference in *Tcf7l2* expression was observed between $Tcf7l2^{+/-}$ and $Tcf7l2^{+/+}$ mice ($U = 5$, $p = 0.381$) (Fig. 5.30).

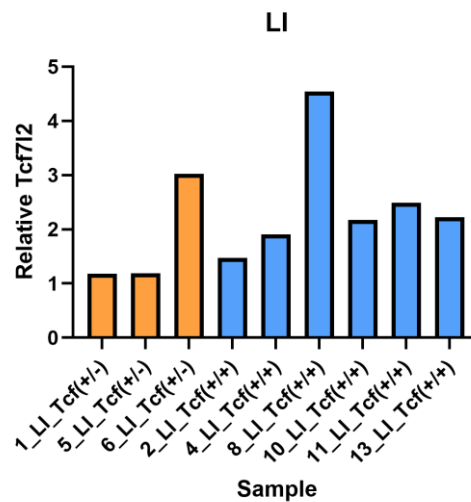


Figure 5.30 Analysis of *Tcf7l2* RNA expression in large intestinal tissue. Bar chart shows expression levels of *Tcf7l2* relative to *Gapdh* in both $Tcf7l2^{+/-}$ (orange) and $Tcf7l2^{+/+}$ (blue) mice.

Finally, the ThermoFisher probe was utilised to test *Tcf7l2* expression in small intestinal tissue. Again, no significant difference in expression was observed between $Tcf7l2^{+/-}$ and $Tcf7l2^{+/+}$ mice ($U = 5$, $p = 0.38$) (Fig. 5.31).

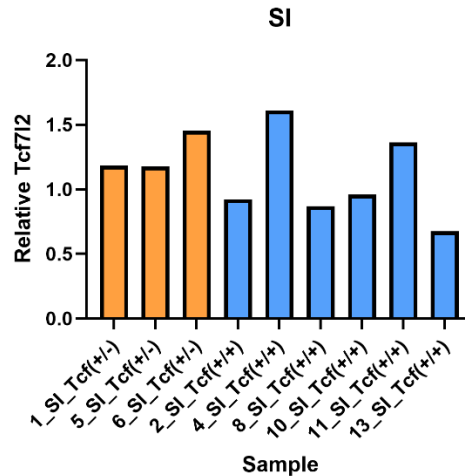


Figure 5.31 Analysis of *Tcf7l2* expression in small intestinal tissue. Bar chart shows expression levels of *Tcf7l2* relative to *Gapdh* in both *Tcf7l2*^(+/-) (orange) and *Tcf7l2*^(+/+) (blue) mice.

These results indicate no apparent change in *Tcf7l2* RNA expression levels in *Tcf7l2*^(+/-) mice as measured using a qPCR assay with an exon 5/6 probe. However, a start codon is present at the beginning of exon 5, potentially allowing translation from this location, stabilising mRNA molecules, and therefore mRNA levels may be recovered from this point leading to the production of mRNA at exon 5-6 around normal levels. For this reason, a second qPCR assay was utilised with primers spanning exon 1. In this case, melt curves for many samples consisted of two peaks (Fig. 5.32).

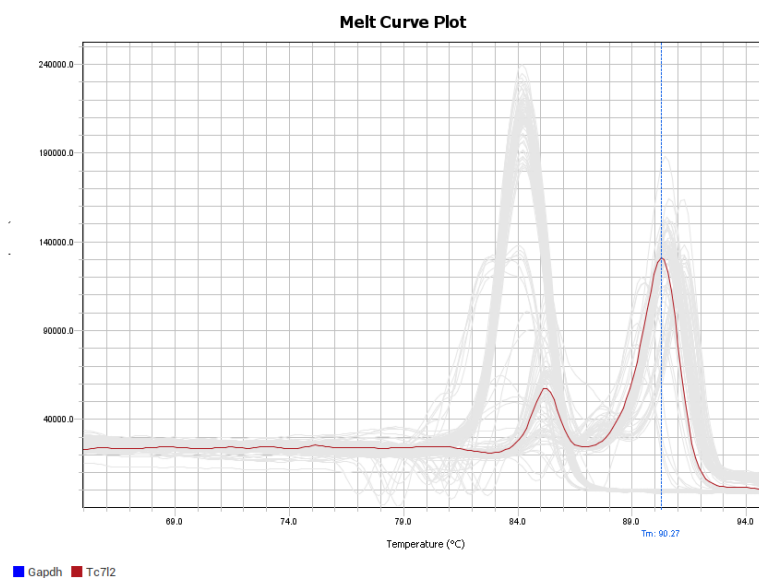


Figure 5.32 Melt curve plot for a heterozygous *Tcf7l2*^(+/-) qPCR liver sample. This melt curve plot shows two melt curves peaks for *Tcf7l2*^(+/-) indicating the presence of two different mRNA products.

The presence of two melt curve peaks is likely to represent the amplification of two different mRNA products of different sizes, owing to the deletion of exon 1 in one allele of heterozygous *Tcf7l2*^(+/-) samples. This made it unclear which allele the resulting expression values corresponded to.

To rule out contamination, fresh SYBER green and other reagents were utilised along with an additional test of a no reverse transcriptase template control. In addition, another qPCR was conducted using one primer located inside the deleted region of exon 1 and another primer outside exon 1, ensuring that expression values would only reflect the WT allele. Theoretically therefore, expression should be halved. However, a fluorescence peak was observed in the no reverse transcriptase controls, indicating contamination.

A western blot was also attempted in order to quantify *Tcf7l2* expression using liver tissue samples from the same mice as used for the qPCR assay. The monoclonal anti-Tcf7l2 antibody bound to the N-terminal section of Tcf7l2 (aa1-49) which is within exon 1. The results of this western blot were inconclusive due to poorly defined bands, which compromised the clarity and interpretability of the data. This requires further optimisation.

5.5 Discussion

This part of the study aimed to investigate the role of *Tcf7l2* as a Wnt pathway mini-driver gene and explore further evidence for the just-right model, through the analysis of a heterozygous knock-out, *Tcf7l2* mouse model, created by Angus-Hill *et al.*, (2011), crossed with the well-established Min mouse model. The results of the study by Angus-Hill *et al.*, (2011) show an increase in tumour number in the colon of *Apc*^(Min/+) mice with the addition of a heterozygous knock-out *Tcf7l2* mutation. This is supportive of the mini-driver and just-right model, suggesting that tumour development is optimised through the combination of a Wnt major-driver mutation in *Apc* and a Wnt mini-driver mutation in *Tcf7l2*. This chapter analysed tumour development in both the small and large intestine of *Tcf7l2*^(+/-); *Apc*^(Min/+) and *Tcf7l2*^(+/-); *Apc*^(Min/+) mice to reproduce the findings of Angus-Hill *et al.*, (2011).

Through sequencing of splenic DNA from *Tcf7l2*^(fl/fl) mice, the *Tcf7l2* construct and allele sequence maps were clarified, revealing that the deletion in the *Tcf7l2* heterozygous mouse model, created by Angus-Hill *et al.*, (2011) is limited to the translated region of exon 1, rather than encompassing the entire exon. Angus-Hill *et al.*, (2011) suggested that this partial deletion would disrupt the splice site and introduce a strong poly(A) signal, reducing the likelihood of producing alternative isoforms of Tcf7l2. However, in the current study, qPCR using primers flanking exons 5-6 showed no reduction in *Tcf7l2* RNA expression across multiple tissue types, suggesting either no change in *Tcf7l2* RNA expression, or the presence of an alternative RNA isoform. A western blot with attempted but produced poor quality results. This experiment should be optimised and repeated to ensure a reduction in the Tcf7l2 protein.

RNA sequencing could be additionally employed to verify whether the knock-out of exon 1 was effective by detecting truncated isoform transcripts containing altered exons. Further RNA and protein analyses would also provide insight into how the Wnt signalling pathway may be affected by reduced or altered Tcf7l2 protein function. Interestingly, Angus-Hill *et al.*, (2011) observed increased expression of Lef1 and Tcf1 in colonic tumours from *Tcf4*^(Het); *Apc*^(Min) mice during their qPCR analysis. As β -catenin typically interacts with Tcf7l2 to activate genes promoting cell survival and proliferation, they hypothesised that other Tcf/Lef family members might compensate for the loss of Tcf7l2 function. This hypothesis is supported by the upregulation of Lef1 and Tcf1 in colonic tumours of *Tcf4*^(Het); *Apc*^(Min) mice (Angus-Hill *et al.*, 2011). Additionally, it would be valuable to explore gene expression changes across the Wnt pathway using the Wnt genes listed in Chapters 2 and 3, comparing expression profiles between *Tcf7l2*^(+/-); *Apc*^(Min/+) mice and *Tcf7l2*^(+/-); *Apc*^(Min/+) mice. This would provide insight into how the just-right model may be affected across the whole pathway when *Tcf7l2* is and is not mutated on a background of a major-driver *Apc* mutation. RNA sequencing may also be useful to identify if any other pathways are affected by hemizygous loss of *Tcf7l2* and therefore influence tumour formation. For example, TCF7L2 protein binds to both a *MYC* enhancer and a *GREM1* enhancer and

therefore hemizygous loss of *Tcf7l2* may influence pathways other than the Wnt pathway (Law *et al.*, 2019).

Due to time constraints, the *Tcf7l2* heterozygous knock-out in combination with an *Apc*^(Min/+) mutation was limited to a *Pgkl-cre* background and therefore, the *Tcf7l2* mutation was constitutive and ubiquitous. This is representative of a germline mutation rather than a somatic mini-driver mutation and may therefore test the role of *Tcf7l2* in tumour formation rather than or in addition to tumour optimisation. As another model, it may be of interest to study the role of *Tcf7l2* as a somatic mini-driver mutation in tumour optimisation, in the intestinal epithelium, by utilising the floxed version of the model with *villin-Cre-ERT2*. *Villin-Cre-ERT2* is a tamoxifen-dependent Cre recombinase specifically expressed in the intestinal epithelium, it would therefore be able to induce a *Tcf7l2* mutation on a background of *Apc*^(Min/+) at a preferred time point (El Marjou *et al.*, 2004). There are however limitations to the use of a *Villin-Cre-ERT2* system. *Villin-Cre-ERT2* targets the intestinal epithelium, thus the continuous renewal of the gut epithelium could affect the persistence of genetic modifications. A study by El Marjou *et al.*, (2004), evidenced that Cre was no longer active after tamoxifen doses were stopped. However, they also showed that recombined alleles as a result of *Villin-Cre-ERT2* persisted for multiple cycles of epithelial renewal. A limitation of the *Villin-Cre-ERT2* system however is the persistence of recombination is varied between the small and large intestine. El Marjou *et al.*, (2004) found that persistence of recombined alleles at 60 days was approximately 80% of small intestinal epithelia compared to 40% of epithelia in the large intestine.

The intestinal analysis revealed no significant difference in tumour number in both the whole small and the large intestines of the main test *Tcf7l2*^(+/-);*Apc*^(Min/+) mice compared with the main control *Tcf7l2*^(+/+);*Apc*^(Min) mice. This contradicts the findings of Angus-Hill *et al.*, (2011) who found that *Tcf4*^(Het) mice harboured an increased number of tumours in the colon compared with *Apc*^(Min) controls. This is perhaps owing to a limited sample size, and so this result is subject to confirmation by further studies. This chapter however improves upon the work by Angus-Hill *et al.*, (2011) through an investigation of tumour distribution throughout the small and large intestine, providing information on the correlation between *Tcf7l2* and *Apc* mutations and specific locations of tumour formation. By dividing both the small intestine and large intestine into multiple sections, this allowed for the study of specific regions of both the proximal and distal small and large intestines.

Firstly, in sections 2 and 3 of the small intestine (proximal) *Tcf7l2*^(+/-);*Apc*^(Min/+) mice had a significantly higher number of tumours than *Tcf7l2*^(+/+);*Apc*^(Min/+) controls. Section 2 of the small intestine also showed a significantly increased tumour volume in mice with both a *Tcf7l2* and *Apc* mutation. This finding contradicts the findings of Angus-Hill *et al.*, who found no significant difference in the number of tumours in the small intestine between *Tcf4*^(Het) mice and *Apc*^(Min/+) controls. Similarly, section 5 of the large intestine of *Tcf7l2*^(+/-);*Apc*^(Min/+) mice, had significantly more tumours than *Tcf7l2*^(+/+);*Apc*^(Min) controls. This is in agreement with the findings of Angus-Hill *et al.*, (2011) who

found that tumour number was increased in the large intestine in mice with a heterozygous *Tcf7l2* mutation. These results indicate that *Tcf7l2* may be able to optimise tumour development in both the small and the large intestines when combined with an *Apc*, major-driver mutation. This is supportive of the just-right model and supportive of *Tcf7l2* as a mini-driver in the Wnt pathway. More supportive evidence of *Tcf7l2* as a mini-driver gene in the Wnt pathway was seen in the *Tcf7l2*^(+/-);*Apc*^(+/+) mice which were aged over one year and did not produce any intestinal tumours suggesting that the *Tcf7l2* heterozygous mutation alone is insufficient in causing intestinal (or other) tumourigenesis.

None of the Kruskal-Wallis tests followed by a Dunn's test indicated a statistically relevant difference between the number of tumours between test *Tcf7l2*^(+/-);*Apc*^(Min/+) mice and *Tcf7l2*^(+/+);*Apc*^(Min/+) controls. The only differences were found between the aforementioned genotypes and *Tcf7l2*^(+/-);*Apc*^(+/+) and WT controls. However, when utilising either a Mann-Whitney U or T-test in a more targeted way between these the two groups of interest, statistical significance was observed in two sections of the small intestine and one section of the large intestine. The difference between tests may be attributed to multiple corrections testing producing more conservative results. A larger sample size may provide more power to the analysis and therefore alter the statistical relevance of some of these results when combined with additional controls.

A wide variation in tumour number and volume can be seen in both the experimental *Tcf7l2*^(+/-);*Apc*^(Min/+) mice and *Tcf7l2*^(+/+);*Apc*^(Min/+). This highlights the need for further investigation with a larger sample size, as well as the exploration of other confounding variables such as sex. According to Wang et al. (2023), male *Apc*^(Min/+) mice have significantly larger and more tumours than female mice.

A valuable addition to this study, providing further insight into the just-right model, would be modulation of the specific mutations in both *Apc* and *Tcf7l2*. Evidence suggests that different *TCF7L2* mutations have different effects on its function. Tang et al. (2008) introduced WT *TCF7L2*, along with four different *TCF7L2* mutations into DLD-1 cells via expression plasmids to test effects on growth. These mutations included both mutations in the N terminal (G>T Glu24-stop, Exon1-delT) and in the C terminal (C>T Arg465-Cys, C>T Arg397-Trp) of *TCF7L2*. Interestingly, both overexpression of the WT sequence and Arg465-Cys mutation caused suppressed growth in DLD-1 cells. The other three mutations on the other hand, abolished the ability of *TCF7L2* to suppress cell growth. It would therefore be of interest to determine the effect of modulated *Tcf7l2* mutations *in-vivo*, and in turn, whether this modulates tumour formation in combination with varying mutations in *Apc*. An alternative knock-out of *Tcf7l2* with exons flanking exon 5 was utilised in previous literature studies of diabetic susceptibility (Yang et al., 2012). This would remove the start codon and other sequences within exon 5 and theoretically cause deletion of the high mobility group (HMG) domain and nuclear localisation sequence. In addition to modulation of the *Tcf7l2* mutation the *Apc* mutation could also be modulated. The *Apc*^(Min/+) mutation is positioned at aa850 (p.Arg850Ter). Multiple alternative *Apc* models could be utilised for this study including *Apc*^(1638/+), *Apc*^(716/+) and *Apc*^(242/+). Each of these models produce a

variable number of tumours and would provide insight into the benefit of the addition of a *Tcf7l2* mutation in the context of differing *Apc* major-driver mutants.

Interestingly, in Chapter 3, several co-occurrence patterns were identified in MSS and MSI⁺ CRCs using the 100kGP dataset. Notably, *AMER1* mutations strongly co-occurred with *APC* mutations in MSS CRCs and were associated with an increase in 20AARs in *APC*, making *AMER1* a strong candidate Wnt pathway mini-driver gene. Modelling the co-occurrence of *AMER1* and *APC* mutations in a mouse model could therefore provide valuable insights into tumourigenesis and the just-right model for this gene. Additionally, *TCF7L2* mutations were found to robustly co-occur with *CTNNB1* mutations in both MSS and MSI⁺ CRCs. It would be of particular interest to explore tumour development by varying the major driver mutation, for instance, introducing *Tcf7l2* mutations alongside *RNF43* and *CTNNB1* mutations, modelling the co-occurrence patterns identified and absent in human datasets. Such an approach could yield further insights into the just-right model of CRC evolution.

In conclusion, this study aimed to explore the role of *Tcf7l2* as a potential Wnt pathway mini-driver and provide evidence supporting the just-right model of CRC evolution. By analysing the heterozygous *Tcf7l2*^(+/-) knock-out model, these results showed that no tumours were formed via *Tcf7l2* heterozygous knock-out alone. This supports the categorisation of *Tcf7l2* as a Wnt pathway mini-driver gene as it cannot cause tumourigenesis in isolation. An increased number of tumours in sections of both the small and the large intestines in *Tcf7l2*^(+/-);*Apc*^(Min/+) mice also supports *Tcf7l2* as a Wnt pathway mini-driver gene and supports the just-right model as it suggests that *Tcf7l2* has the ability to optimise tumourigenesis. Notably, this model is a model of a germline mutation and therefore shows that *Tcf7l2* may play a role in tumour formation rather than purely in optimisation in sub-clonal regions as suggested by Castro-Giner, Ratcliffe and Tomlinson (2015). A *TCF7L2* mutation was also found on the trunk of a phylogenetic tree in Chapter 2 which could also be an example of *TCF7L2* playing a role in tumour formation rather than optimisation. In addition, a genome-wide significant SNP at the *TCF7L2* locus was associated with increased TCF7L2 expression in the distal colon. Altogether this evidence suggests that *TCF72* could be categorised as a mini-driver however the definition of a mini-driver as a sub-clonal mutation solely involved in later-stage optimisation, needs further investigation.

Although a difference in tumour number was found in sections of the intestine, providing support for the just-right model, for improvement of this study, a larger number of samples would be beneficial. Increasing the sample size may provide the data required to further support the ability of *Tcf7l2* mutations to significantly increase the number of tumours in the whole intestine when crossed with an *Apc*-mutant strain, providing robust support for the just-right model. Future work may benefit from the incorporation of a conditional knock-out model of *Tcf7l2* using *Villin-cre-ERT2* to analyse the ability of *Tcf7l2* to optimise tumourigenesis. Varying the major-driver mutations in *Apc* and mini-driver mutations in *Tcf7l2* could further clarify the intricacies of the just-right model and its role in CRC evolution.

Chapter 6

Final discussion and conclusions

6. Final discussion and conclusions

The refinement of existing models with improved understanding of selective constraints upon CRC evolution is crucial for gaining a comprehensive insight into the mechanisms driving tumour development and progression. A thorough understanding of CRC progression can assist in improving diagnosis, enhancing prognostic accuracy and developing effective treatment strategies.

This thesis firstly aimed to further explore and refine current models of CRC evolution through the multiregional sequencing of large adenomatous colorectal polyps, both with and without a focus of cancer. This showed that most polyps harboured mutations in driver genes described by the Vogelstein model e.g. *APC*, *KRAS*, *TP53* and 18q loss (Fearon and Vogelstein, 1990). However, multiple findings contradicted the Vogelstein/linear model. Firstly, two polyps with a focus of cancer did not harbour the entire mutation complement as suggested necessary for tumour progression by the Vogelstein model. In addition, each individual polyp was shown to exhibit ITH, and driver ITH. This contradicts both the Vogelstein/linear model through the clear lack of selective sweeps, as well as the big bang model which suggests that all mutations capable of driving tumourigenesis should occur early in development, and thus exhibit monoclonality. Altogether, these results were more consistent with branched and parallel evolution, in which clones with varied driver mutations can exist and evolve in parallel, within the same neoplasm. This model suggests that each branch represents a different route to cancer formation (Gerlinger *et al.*, 2014).

Another key result presented by this chapter was developed through further exploration of ITH. Phylogenetic analysis indicated that despite having no difference in overall mutation burden, polyps with a focus of cancer exhibited enhanced mutation accumulation and genetic diversity at the branch level (non-clonal mutations, exhibited in at least two samples) as opposed to leaf (mutations unique to individual samples). This indicates that critical genetic diversifications, potentially involving key driver mutations, occur early in tumour development, though this happens after the most recent common ancestor and so some driver mutations are not clonal. Cross *et al.*, (2018) propose that evolution following this, pattern in CRCs, is neutral.

This study could provide further insights by comparing the ratio of non-synonymous to synonymous mutations on both the branches and leaves of phylogenetic trees to establish whether neutral evolution occurs following this critical period of advanced genetic diversity in polyps with a focus of cancer, or whether neutral evolution is limited to advanced CRCs. This study would also benefit from more advanced sampling techniques involving sampling at greater depths of the polyp to capture potentially genetically different clones, together with a greater number of samples per polyp so more comparisons could be made between samples, including better sampling of cancerous and non-cancerous regions through histopathological analyses of the samples taken, and the use of MutationTimer to provide a more accurate timeline for genetic events.

The just-right model of CRC was developed through the study of 20AARs in *APC*, in FAP. This model proposed that a just-right level of Wnt pathway activation is achieved through the selection of particular mutations in *APC*, resulting in a just-right number of 20AARs and therefore Wnt activation. Segditsas and Tomlinson (2006) suggested that the just-right model could extend across the Wnt pathway, incorporating mutations in genes other than *APC*. In 2015, Castro-Giner, Ratcliffe and Tomlinson (2015) then proposed that the just-right level of pathway activation can be achieved through a combination of both major-driver and mini-driver mutations. Evidence of the just-right model/mini-driver model was investigated utilising the multiregional polyp samples. This provided broad evidence of major-driver mutual exclusivity and multiple examples of potential mini-drivers occurring in combination with an *APC* major-driver mutation. This study provided broad evidence of the just-right model acting in adenomatous colorectal polyps.

The relationship between major-driver and mini-driver mutations in the Wnt pathway was studied in greater depth using the 100KGP dataset. Wnt pathway mutations in approximately 2,000 CRCs were investigated for patterns of mutual exclusivity and co-occurrence. A key finding of this study was firstly validation of the mutually exclusive relationship between major-driver mutations in Wnt pathway components: *RNF43*, *CTNNB1*, *RSPO* fusions and *APC*. In addition, mini-driver mutations were found to co-occur with major-driver mutations in MSS CRCs e.g. *AMER1* and *APC*, *TCF7L2* and *APC*, and *TCF7L2* and *CTNNB1*.

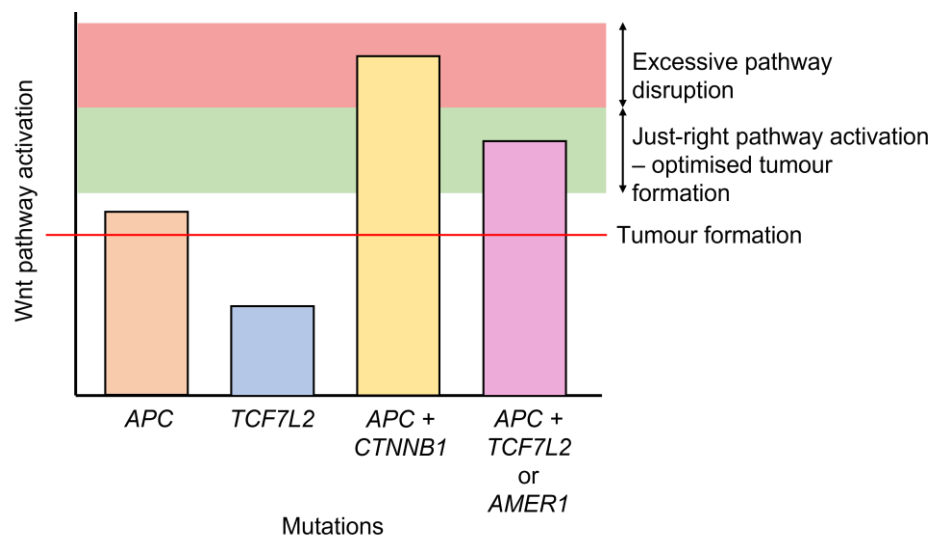


Figure 6.1 A diagrammatic representation of just-right Wnt pathway activation as described by the just-right and mini-driver models of CRC evolution. A just-right level of pathway activation is achieved through the addition of a mini-driver mutation to a major-driver mutation. A mini-driver mutation cannot cause tumourigenesis in isolation. Both *APC* and *CTNNB1* mutations are mutually exclusive in CRC suggesting negative selection perhaps due to excessive pathway disruption leading to apoptosis.

This chapter also revealed that mutational relationships between genes were affected by tumour context (MSS/MSI status) and position in the colon (proximal/distal). MSI is often found in serrated

lesions associated with *BRAF* mutation. *APC* mutations are rare in these tumours, and do not play a significant role in this pathway of carcinogenesis. Thus, the positive association between *SOX9*, *FBXW7* and *APC* were not observed in MSI⁺ CRCs, rather a very strong co-occurrence was seen between *RNF43* and *ZNRF3* mutations. In addition, an increased number of 20AARs in *APC* were found in the proximal colon than the distal showing that the just-right number of 20AARs may vary depending on location.

This analysis identified a positive association between *APC* and *AMER1* mutations in CRCs. To further explore this, the presence of *AMER1* mutations were found to be associated with an increased number of 20AARs in *APC*. This provides supporting evidence of *AMER1* acting as a Wnt pathway mini-driver as it suggests that *AMER1* fine-tunes Wnt signalling in the presence of a sub-optimal number of 20AARs in *APC*.

This study provides support for the just-right model in the Wnt pathway and highlights its complexity and context dependence in CRC. It also highlights the importance of understanding the interactive relationship between mutations in different genes, and the importance of mini-driver genes in tumour evolution.

A limitation of this study is that it is based on binary data, capturing only the presence or absence of a pathogenic mutations in each Wnt pathway gene. As discussed in Chapter 3, *AMER1* has dual functionality, and like other genes in the Wnt pathway, different mutations can lead to distinct effects on the resulting protein. Therefore, it would be hugely valuable to investigate the co-occurrence and mutual exclusivity of specific mutations across different genes in the Wnt pathway. Categorising gene-specific mutations as mini-drivers, when appropriate, may provide a more precise and meaningful classification. In addition, the analysis and incorporation of alternative mechanisms of Wnt dysregulation would be a great benefit to this analysis, such as the transcriptional silencing of *DKK1*, *SFRP1* and *WIF1*. Expanding this analysis to investigate if there is a just-right level of pathway activation across multiple different signalling pathways would also be of interest, as well as an investigation to see if the model can be applied to different cancer types e.g. endometrial cancer. Further studies may also look at whether the addition of a mini-driver mutation leads to a change in prognosis and survival. In a recent study, Segura et al. (2023) identified genes mutated at a frequency of less than 5% in a large CRC dataset with the aim of finding candidate mini-driver mutations. They identified *DOCK3*, *FNI*, *PAPPA2*, *DNAH11* and *FBN2*. They found that the presence of these mutations was associated with poor survival, indicating that mini-driver mutations may be useful as prognostic biomarkers. By elucidating and having a clear understanding of just-right pathway activation, this could indicate whether a polyp is likely to progress to a cancer. In addition, through defining the occurrence of parallel evolution and exploring different routes of carcinogenesis, this could indicate different mechanisms of treatment resistance. As stated in Chapter 3, a further study could look at the different Wnt pathway mutations present in tumours that have received treatment and exhibited resistance.

In chapter 3 a pattern of co-occurrence between *TCF7L2* and *APC* mutations was identified. Due to the occurrence of *TCF7L2* mutations in 10% of CRCs, the ability of tumours to evolve without *TCF7L2* mutations, its status as a transcription factor in the Wnt pathway, and its relationship with *APC* mutations, Chapter 4 aimed to clarify the effect of *TCF7L2* SNPs on CRC risk, and Chapter 5 aimed to further assess the evidence for *TCF7L2* as a mini-driver using a heterozygous knock-out mouse model.

Chapter 4 firstly established the presence of two separate genome-wide significant signals at the 10q25.2 region through the conditional analysis of GWAS meta-analysis data, one at the *TCF7L2* locus and another at the *VTI1A* locus. An association was then observed between risk alleles at multiple *TCF7L2* SNP loci and increased *TCF7L2* expression in the rectum, through an eQTL analysis. This eQTL analysis also identified an association between risk alleles at multiple *TCF7L2* SNP loci and decreased *VTI1A* expression in the distal colon, as well as an association between risk alleles at multiple SNP loci at the *VTI1A* locus and increased *VTI1A* expression in the distal colon. These results suggest that both *TCF7L2* and *VTI1A* may have a contributory role in CRC risk subject to further statistical analysis and functional studies. Notably, expression of both *TCF7L2* and *VTI1A* were found to be positively associated with one another in both the proximal colon and the rectum, potentially representing a mini-driver co-occurrence, subject to confirmation by functional studies.

To supplement the GWAS analysis and further investigate the role of *TCF7L2* in tumour development, a heterozygous knock-out murine model of *Tcf7l2* was analysed. The aim of this work was to validate the work of Angus-Hill *et al.*, (2011) and establish if *Tcf7l2* mutations are capable of optimising Wnt signalling in the presence of an *Apc*^(Min/+) major-driver mutation. The results showed that the addition of a heterozygous knock-out *Tcf7l2* mutation (*Tcf7l2*^(+/-)) to an *Apc*^(Min/+) mutation resulted in a significant increase in tumour number in certain sections of both the small and the large intestines compared to *Tcf7l2*^(+/+);*Apc*^(Min/+) control mice. This both supported the findings of Angus-Hill *et al.*, (2011), as they also found an increased number of tumours in the large intestine in *Tcf4*^(Het);*Apc*^(Min) mice and contradicted their findings as they found no significant difference in tumour number between these test and control mice in the small intestine. This result supports the hypothesis that *Tcf7l2* mutations have a role in enhancing tumour formation but are not sufficient to induce tumorigenesis in isolation. This supports its categorisation as a mini-driver and provides supportive evidence for the just-right model.

The definition of a mini-driver gene is complex. Castro-Giner, Ratcliffe and Tomlinson (2015) suggest that mini-drivers are of low frequency and small effect. They also specified that mini-drivers should be sub-clonal due to their role in tumour optimisation rather than formation and be associated with a lack of selective sweeps as a result of their overall weak effect. Mini-drivers should also be incapable of driving tumorigenesis in isolation resulting in a co-occurrence with major-driver mutations. Transcription factors are good candidates for mini-drivers as they allow for the modulation of expression of other genes. *TCF7L2* is therefore a good candidate Wnt pathway mini-driver. Firstly,

TCF7L2 is a transcription factor. Secondly, Chapter 3 showed a co-occurrence between mutations in *TCF7L2* and major-driver *APC* mutations in a large CRC dataset. In addition, the conditional *Tcf7l2* knock-out mouse model analysed in Chapter 5 showed that heterozygous knock-out of *Tcf7l2* is not capable of driving tumourigenesis on its own and that tumour number is increased in *Tcf7l2*^(+/-);*Apc*^(Min/+) mice in certain regions of the small and large intestine compared with *Tcf7l2*^(+/+);*Apc*^(Min/+) controls. However, it is important to note that none of these points are enough to categorise *TCF7L2* as a mini-driver on their own. For example, *KRAS* mutations, despite being major-drivers of CRC, cannot induce tumourigenesis. However, their frequency in CRC is substantially higher than *TCF7L2* (Najumudeen *et al.*, 2024). *KRAS* mutations are present in approximately 40% of cancers, while *TCF7L2* mutations are present in approximately 10% (Fröhlich, Rose and Hecht, 2023; Najumudeen *et al.*, 2024). Notably, *CTNNB1* mutations are present in approximately 6% of CRCs, however they have a large effect on the Wnt pathway. *CTNNB1* mutations have been shown to induce transformation in the colon when homozygously mutated or when membranous β -catenin binding is inhibited. They are also highly selected in MSI⁺ CRCs and are mutually exclusive with major-driver *APC* mutations in the Wnt pathway suggesting that the just-right level of Wnt pathway activation is hypothetically surpassed when both major-driver genes are mutated.

All of the thesis results when taken together indicate that *TCF7L2* can be considered a candidate mini-driver gene, however, the results of many analyses in this thesis show that *TCF7L2* may play a role in tumour formation as well as optimisation in sub-clonal regions of CRAs and CRCs. Firstly, a *TCF7L2* mutation was identified on the phylogenetic tree of a single large polyp in Chapter 2. In addition, the conditional analysis and eQTL analysis suggested that *TCF7L2* may be associated with an increased risk of tumour formation subject to further statistical analysis and functional studies. Finally, a constitutive heterozygous knock-out *Tcf7l2* mutation in combination with a constitutive *Apc*^(Min/+) mutation induced the formation of an increased number of tumours in different sections of the small and large intestines. These results suggest that mini-drivers may also be involved in tumour formation as well as optimisation, further clarifying the definition of a mini-driver mutation.

In summary, the findings presented in this thesis provide substantial support for both the just-right and mini-driver models of CRC evolution. They enhance our understanding of the selective constraints involved in CRC progression and further clarify the role that mini-driver genes play in tumour development. Furthermore, these results pave the way for future research to investigate the just-right model more comprehensively, allowing for deeper exploration of its mechanisms and broader implications for CRC.

Further statements and acknowledgements

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure.

I am a contributing author of the Cornish *et al.*, (2024) paper discussed in this thesis. My work on the correlations between mutations in Wnt pathway genes was included in the supplementary information of this paper. I am also a contributing author of the Brunet *et al.*, (*) paper discussed in this thesis to which I contributed my advice, biological knowledge and understanding of mutations in Wnt pathway genes.

References

- Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013) *Predicting functional effect of human missense mutations using PolyPhen-2*, *Current Protocols in Human Genetics*. Available at: <https://doi.org/10.1002/0471142905.hg0720s76>.
- Aguilera, O. *et al.* (2006) 'Epigenetic inactivation of the Wnt antagonist DICKKOPF-1 (DKK-1) gene in human colorectal cancer', *Oncogene*, 25(29), pp. 4116–4121. Available at: <https://doi.org/10.1038/sj.onc.1209439>.
- Albuquerque, C. *et al.* (2002) 'The "just-right" signaling model: APC somatic mutations are selected based on a specific level of activation of the β -catenin signaling cascade', *Human Molecular Genetics* [Preprint]. Available at: <https://doi.org/10.1093/hmg/11.13.1549>.
- Angus-Hill, M.L. *et al.* (2011) 'T-cell factor 4 functions as a tumor suppressor whose disruption modulates colon cell proliferation and tumorigenesis', *Proceedings of the National Academy of Sciences of the United States of America*, 108(12), pp. 4914–4919. Available at: <https://doi.org/10.1073/pnas.1102300108>.
- Aoki, K. and Taketo, M.M. (2007) 'Adenomatous polyposis coli (APC): A multi-functional tumor suppressor gene', *Journal of Cell Science*, pp. 3327–3335. Available at: <https://doi.org/10.1242/jcs.03485>.
- Ascierto, P.A. *et al.* (2012) 'The role of BRAF V600 mutation in melanoma', *Journal of Translational Medicine* [Preprint]. Available at: <https://doi.org/10.1186/1479-5876-10-85>.
- Bachet, J. *et al.* (2006) 'Priority Report KRAS Mutation Status Is Predictive of Response to Cetuximab Therapy in Colorectal Cancer Astrid Lie', *Cancer Research*, 66(8), pp. 3992–3995. Available at: <https://doi.org/10.1158/0008-5472.CAN-06-0191>.
- Bakker, E.R.M. *et al.* (2013) ' β -Catenin signaling dosage dictates tissue-specific tumor predisposition in Apc-driven cancer', *Oncogene*, 32(38), pp. 4579–4585. Available at: <https://doi.org/10.1038/onc.2012.449>.
- Beggs, A.D. *et al.* (2010) 'Peutz e Jeghers syndrome : a systematic review and recommendations for management', *Gut*, 59(7), pp. 975–986. Available at: <https://doi.org/10.1136/gut.2009.198499>.
- Boland, C.R. and Goel, A. (2010) 'Microsatellite Instability in Colorectal Cancer', *Gastroenterology*, 138(6). Available at: <https://doi.org/10.1053/j.gastro.2009.12.064>.
- Briggs, S. and Tomlinson, I. (2013) 'Germline and somatic polymerase ϵ and δ mutations define a new class of hypermutated colorectal and endometrial cancers', *Journal of Pathology*, 230(2), pp. 148–153. Available at: <https://doi.org/10.1002/path.4185>.
- Brunet Guasch, M. *et al.* (no date) 'Quantifying "just-right" APC inactivation for colorectal cancer initiation'. Available at: <https://doi.org/10.1101/2024.07.10.602868>.
- Bürtin, F., Mullins, C.S. and Linnebacher, M. (2020) 'Mouse models of colorectal cancer: Past, present and future perspectives', *World Journal of Gastroenterology*. Baishideng Publishing Group Co, pp. 1394–1426. Available at: <https://doi.org/10.3748/WJG.V26.I13.1394>.
- Castro-Giner, F., Ratcliffe, P. and Tomlinson, I. (2015) 'The mini-driver model of polygenic cancer evolution', *Nature Reviews Cancer* [Preprint]. Available at: <https://doi.org/10.1038/nrc3999>.

- Cerami, E. *et al.* (2012) 'The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data', *Cancer Discovery*, 2(5), pp. 401–404. Available at: <https://doi.org/10.1158/2159-8290.CD-12-0095>.
- Chakravarty, D. *et al.* (2017) 'OncoKB : A Precision Oncology Knowledge Base', *JCO Precision Oncology*, (1), pp. 1–16.
- Chalmers, Z.R. *et al.* (2017) 'Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden', *Genome Medicine*, 9(1). Available at: <https://doi.org/10.1186/s13073-017-0424-2>.
- Chang, J.J. *et al.* (2020) 'Long term outcomes of colon polyps with high grade dysplasia following endoscopic resection', *BMC Gastroenterology*, 20(1). Available at: <https://doi.org/10.1186/s12876-020-01499-2>.
- Chang, L. *et al.* (2023) 'Systematic profiling of conditional pathway activation identifies context-dependent synthetic lethalties', *Nature Genetics*, 55(10), pp. 1709–1720. Available at: <https://doi.org/10.1038/s41588-023-01515-7>.
- Chatila, W.K. *et al.* (2023) 'Integrated clinical and genomic analysis identifies driver events and molecular evolution of colitis-associated cancers', *Nature Communications*, 14(1). Available at: <https://doi.org/10.1038/s41467-022-35592-9>.
- Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, Anthony J, *et al.* (2016) 'Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications.', *Bioinformatics (Oxford, England)*, 32(8), pp. 1220–1222. Available at: <https://doi.org/10.1093/bioinformatics/btv710>.
- Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, Anthony J., *et al.* (2016) 'Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications', *Bioinformatics*, 32(8), pp. 1220–1222. Available at: <https://doi.org/10.1093/bioinformatics/btv710>.
- Choi, J. *et al.* (2023) 'Influence of location-dependent sex difference on PD-L1, MMR/MSI, and EGFR in colorectal carcinogenesis', *PLoS ONE*, 18(2 February). Available at: <https://doi.org/10.1371/journal.pone.0282017>.
- Christie, M. *et al.* (2013) 'Different APC genotypes in proximal and distal sporadic colorectal cancers suggest distinct WNT/ β -catenin signalling thresholds for tumourigenesis', *Oncogene*, 32(39), pp. 4675–4682. Available at: <https://doi.org/10.1038/onc.2012.486>.
- Cornish, A.J. *et al.* (2024a) 'The genomic landscape of 2,023 colorectal cancers', *Nature*, 633(8028), pp. 127–136. Available at: <https://doi.org/10.1038/s41586-024-07747-9>.
- Cornish, A.J. *et al.* (2024b) 'The genomic landscape of 2,023 colorectal cancers', *Nature* [Preprint]. Available at: <https://doi.org/10.1038/s41586-024-07747-9>.
- Cross, W. *et al.* (2018) 'The evolutionary landscape of colorectal tumorigenesis', *Nature Ecology and Evolution* [Preprint]. Available at: <https://doi.org/10.1038/s41559-018-0642-z>.
- Danecek, P. *et al.* (2021) 'Twelve years of SAMtools and BCFtools', *GigaScience*, 10(2). Available at: <https://doi.org/10.1093/gigascience/giab008>.
- D.L. Swofford (2003) 'PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.0b10', *Sinauer Associates* [Preprint], (4).

- Doll, R. and Hill, A.B. (1956) 'Lung Cancer And Other Causes Of Death In Relation To Smoking A Second Report On The Mortality Of British Doctors', *British Medical Journal*, 2(5001), p. 1071. Available at: <https://doi.org/10.1136/bmj.2.5001.1071>.
- Dunlop, M.G. (2002) 'Guidance on gastrointestinal surveillance for hereditary non-polyposis colorectal cancer, familial adenomatous polyposis, juvenile polyposis, and Peutz-Jeghers syndrome', *Gut*, 51(SUPPL. 5), pp. 21–27. Available at: https://doi.org/10.1136/gut.51.suppl_5.v21.
- Fang, D.-C. *et al.* (2002) 'Mutation analysis of APC gene in gastric cancer with microsatellite instability', *World Journal of Gastroenterology*, 8(5). Available at: www.wjgnet.com.
- Farnhammer, F., Colozza, G. and Kim, J. (2023) 'RNF43 and ZNRF3 in Wnt Signaling - A Master Regulator at the Membrane', *International Journal of Stem Cells*, (7), pp. 1–9.
- Favoriti, P. *et al.* (2016) 'Worldwide burden of colorectal cancer: a review', *Updates in Surgery*, 68(1), pp. 7–11. Available at: <https://doi.org/10.1007/s13304-016-0359-y>.
- Fearon, E.R. and Vogelstein, B. (1990a) 'A genetic model for colorectal tumorigenesis', *Cell* [Preprint]. Available at: [https://doi.org/10.1016/0092-8674\(90\)90186-I](https://doi.org/10.1016/0092-8674(90)90186-I).
- Fearon, E.R. and Vogelstein, B. (1990b) 'A genetic model for colorectal tumorigenesis', *Cell* [Preprint]. Available at: [https://doi.org/10.1016/0092-8674\(90\)90186-I](https://doi.org/10.1016/0092-8674(90)90186-I).
- Fernandez-Rozadilla, C. *et al.* (2023a) '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. Available at: <https://doi.org/10.1038/s41588-022-01222-9>.
- Fernandez-Rozadilla, C. *et al.* (2023b) '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. Available at: <https://doi.org/10.1038/s41588-022-01222-9>.
- Fernandez-Rozadilla, C. *et al.* (2023c) '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. Available at: <https://doi.org/10.1038/s41588-022-01222-9>.
- Fisher, R., Pusztai, L. and Swanton, C. (2013) 'Cancer heterogeneity: Implications for targeted therapeutics', *British Journal of Cancer*, pp. 479–485. Available at: <https://doi.org/10.1038/bjc.2012.581>.
- Fontana, E. and Valeri, N. (2019) 'Class(Y) dissection of BRAF heterogeneity: Beyond non-V600', *Clinical Cancer Research* [Preprint]. Available at: <https://doi.org/10.1158/1078-0432.CCR-19-2732>.
- Fox, E.J. and Loeb, L.A. (2010) 'Lethal Mutagenesis: Targeting the Mutator Phenotype in Cancer', *Seminars in Cancer Biology*, pp. 353–359. Available at: <https://doi.org/10.1016/j.semcancer.2010.10.005>.
- Fröhlich, J., Rose, K. and Hecht, A. (2023) 'Transcriptional activity mediated by β -CATENIN and TCF/LEF family members is completely dispensable for survival and propagation of multiple human colorectal cancer cell lines', *Scientific Reports*, 13(1). Available at: <https://doi.org/10.1038/s41598-022-27261-0>.
- Gallagher, M.D. and Chen-Plotkin, A.S. (2018) 'The Post-GWAS Era: From Association to Function', *American Journal of Human Genetics*, 102(5), pp. 717–730. Available at: <https://doi.org/10.1016/j.ajhg.2018.04.002>.

- Gay, D.M. *et al.* (2019) 'Loss of BCL9/9l suppresses Wnt driven tumourigenesis in models that recapitulate human cancer', *Nature Communications*, 10(1). Available at: <https://doi.org/10.1038/s41467-019-08586-3>.
- Geoghegan, G. *et al.* (2019) 'Targeted deletion of Tcf7l2 in adipocytes promotes adipocyte hypertrophy and impaired glucose metabolism', *Molecular Metabolism*, 24, pp. 44–63. Available at: <https://doi.org/10.1016/j.molmet.2019.03.003>.
- Gerlinger, M. *et al.* (2014) 'Cancer: Evolution Within a Lifetime', *Annual Review of Genetics* [Preprint]. Available at: <https://doi.org/10.1146/annurev-genet-120213-092314>.
- Gerstung, M. *et al.* (2020) 'The evolutionary history of 2,658 cancers', *Nature*, 578(7793), pp. 122–128. Available at: <https://doi.org/10.1038/s41586-019-1907-7>.
- Giannakis, M. *et al.* (2014a) 'RNF43 is frequently mutated in colorectal and endometrial cancers', *Nature Genetics*, 46(12), pp. 1264–1266. Available at: <https://doi.org/10.1038/ng.3127>.
- Giannakis, M. *et al.* (2014b) 'RNF43 is frequently mutated in colorectal and endometrial cancers', *Nature Genetics*, 46(12), pp. 1264–1266. Available at: <https://doi.org/10.1038/ng.3127>.
- Gloyn, A.L., Braun, M. and Rorsman, P. (2009) 'Type 2 diabetes susceptibility gene TCF7L2 and its role in β -cell function', *Diabetes*, pp. 800–802. Available at: <https://doi.org/10.2337/db09-0099>.
- Guastadisegni, C. *et al.* (2010) 'Microsatellite instability as a marker of prognosis and response to therapy: A meta-analysis of colorectal cancer survival data', *European Journal of Cancer*, 46(15), pp. 2788–2798. Available at: <https://doi.org/10.1016/j.ejca.2010.05.009>.
- Guo, Q. *et al.* (2021) 'A β -catenin-driven switch in tcf/lef transcription factor binding to dna target sites promotes commitment of mammalian nephron progenitor cells', *eLife*, 10, pp. 1–47. Available at: <https://doi.org/10.7554/eLife.64444>.
- Half, E., Bercovich, D. and Rozen, P. (2009) 'Familial adenomatous polyposis', *Orphanet Journal of Rare Diseases*, 4(1). Available at: <https://doi.org/10.1186/1750-1172-4-22>.
- Hanahan, D. and Weinberg, R.A. (2011) 'Hallmarks of cancer: The next generation', *Cell*, 144(5), pp. 646–674. Available at: <https://doi.org/10.1016/j.cell.2011.02.013>.
- Hao, H.X. *et al.* (2012) 'ZNRF3 promotes Wnt receptor turnover in an R-spondin-sensitive manner', *Nature* [Preprint]. Available at: <https://doi.org/10.1038/nature11019>.
- Harrison, P.W. *et al.* (2024) 'Ensembl 2024', *Nucleic Acids Research*, 52(D1), pp. D891–D899. Available at: <https://doi.org/10.1093/nar/gkad1049>.
- Hirschhorn, J.N. and Daly, M.J. (2005) 'Genome-wide association studies for common diseases and complex traits', *Nature Reviews Genetics*, 6(2), pp. 95–108. Available at: <https://doi.org/10.1038/nrg1521>.
- Houlston, R.S. *et al.* (2016) 'Europe PMC Funders Group Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41 , 3q26 . 2 ', 42(11), pp. 973–977. Available at: <https://doi.org/10.1038/ng.670>. Meta-analysis.
- Huge, N. *et al.* (2020) 'Wnt status-dependent oncogenic role of BCL9 and BCL9L in hepatocellular carcinoma', *Hepatology International*, 14(3), pp. 373–384. Available at: <https://doi.org/10.1007/s12072-019-09977-w>.

- Hutchinson, A., Asimit, J. and Wallace, C. (2020) 'Fine-mapping genetic associations', 29(August), pp. 82–89. Available at: <https://doi.org/10.1093/hmg/ddaa148>.
- Huyghe, J.R. *et al.* (2019a) 'Discovery of common and rare genetic risk variants for colorectal cancer', *Nature Genetics*, 51(1), pp. 76–87. Available at: <https://doi.org/10.1038/s41588-018-0286-6>.
- Huyghe, J.R. *et al.* (2019b) 'Discovery of common and rare genetic risk variants for colorectal cancer', *Nature Genetics*, 51(1), pp. 76–87. Available at: <https://doi.org/10.1038/s41588-018-0286-6>.
- International HapMap Consortium (2003) 'International HapMap Consortium. The International HapMap Project', *Nature*, 426(6968), pp. 789–796.
- Ioannidis, N.M. *et al.* (2016) 'REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants', *American Journal of Human Genetics*, 99(4), pp. 877–885. Available at: <https://doi.org/10.1016/j.ajhg.2016.08.016>.
- Johnson, C.M. *et al.* (2013) 'Meta-Analyses of colorectal cancer risk factors', *Cancer Causes and Control*, 24(6), pp. 1207–1222. Available at: <https://doi.org/10.1007/s10552-013-0201-5>.
- Keum, N.N. and Giovannucci, E. (2019) 'Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies', *Nature Reviews Gastroenterology and Hepatology*, 16(12), pp. 713–732. Available at: <https://doi.org/10.1038/s41575-019-0189-8>.
- Kichaev, G. *et al.* (2014) 'Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies', *PLoS Genetics*, 10(10). Available at: <https://doi.org/10.1371/journal.pgen.1004722>.
- Kim, S. *et al.* (2018) 'Strelka2: fast and accurate calling of germline and somatic variants', *Nature Methods*, 15(8), pp. 591–594. Available at: <https://doi.org/10.1038/s41592-018-0051-x>.
- Kim, S. and Jeong, S. (2019) 'Mutation hotspots in the β -catenin gene: Lessons from the human cancer genome databases', *Molecules and Cells*. Korean Society for Molecular and Cellular Biology, pp. 8–16. Available at: <https://doi.org/10.14348/molcells.2018.0436>.
- Knudson, A.G. (1971) 'Mutation and cancer: statistical study of retinoblastoma.', *Proceedings of the National Academy of Sciences of the United States of America*, 68(4), pp. 820–823. Available at: <https://doi.org/10.1073/pnas.68.4.820>.
- Kohler, E.M. *et al.* (2009) ' β -Catenin degradation mediated by the CID domain of APC provides a model for the selection of APC mutations in colorectal, desmoid and duodenal tumours', *Human Molecular Genetics*, 18(2), pp. 213–226. Available at: <https://doi.org/10.1093/hmg/ddn338>.
- Komiya, Y. and Habas, R. (2008) 'Wnt signal transduction pathways', *Organogenesis* [Preprint]. Available at: <https://doi.org/10.4161/org.4.2.5851>.
- Koo, B.K. *et al.* (2012) 'Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors', *Nature* [Preprint]. Available at: <https://doi.org/10.1038/nature11308>.
- Lamlum, H. *et al.* (1999a) 'The type of somatic mutation at APC in familial adenomatous polyposis is determined by the site of the germline mutation: A new facet to Knudson's "two-hit" hypothesis', *Nature Medicine* [Preprint]. Available at: <https://doi.org/10.1038/12511>.
- Lamlum, H. *et al.* (1999b) 'The type of somatic mutation at APC in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's "two-hit" hypothesis', *Nature*, 5(9), pp. 1071–1075.

- Latchford, A.R. *et al.* (2011) 'Peutz-Jeghers Syndrome : Intriguing Suggestion From Surveillance', *Diseases of the Colon & Rectum*, 54(12), pp. 1547–1551. Available at: <https://doi.org/10.1097/DCR.0b013e318233a11f>.
- Law, P.J. *et al.* (2019) 'Association analyses identify 31 new risk loci for colorectal cancer susceptibility', *Nature Communications*, 10(1). Available at: <https://doi.org/10.1038/s41467-019-09775-w>.
- Layer, R.M. *et al.* (2014) 'LUMPY: a probabilistic framework for structural variant discovery.', *Genome biology*, 15(6), p. R84. Available at: <https://doi.org/10.1186/gb-2014-15-6-r84>.
- Lei, S. *et al.* (2023) 'AMER1 deficiency promotes the distant metastasis of colorectal cancer by inhibiting SLC7A11- and FTL-mediated ferroptosis', *Cell Reports*, 42(9). Available at: <https://doi.org/10.1016/j.celrep.2023.113110>.
- Li, H. (2011) 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics*, 27(21), pp. 2987–2993. Available at: <https://doi.org/10.1093/bioinformatics/btr509>.
- Liao, G. *et al.* (2006) *Jun NH 2-terminal kinase (JNK) prevents nuclear-catenin accumulation and regulates axis formation in Xenopus embryos*. Available at: www.pnas.org/cgi/doi/10.1073/pnas.0602557103.
- Lien, W.H. and Fuchs, E. (2014) 'Wnt some lose some: Transcriptional governance of stem cells by Wnt/ β -catenin signaling', *Genes and Development*. Cold Spring Harbor Laboratory Press, pp. 1517–1532. Available at: <https://doi.org/10.1101/gad.244772.114>.
- Lièvre, A. *et al.* (2006) 'KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer', *Cancer Research*, 66(8). Available at: <https://doi.org/10.1158/0008-5472.CAN-06-0191>.
- Liu, J. *et al.* (2022) 'Wnt/ β -catenin signalling: function, biological mechanisms, and therapeutic opportunities', *Signal Transduction and Targeted Therapy*, 7(1). Available at: <https://doi.org/10.1038/s41392-021-00762-6>.
- Lonai, P. (1998) *Maternally expressed PGK-Cre transgene as a tool for early and uniform activation of the Cre site-specific recombinase*.
- Longley, D.B. and Johnston, P.G. (2005) 'Molecular mechanisms of drug resistance', *Journal of Pathology*, 205(2), pp. 275–292. Available at: <https://doi.org/10.1002/path.1706>.
- Lynch, H.T. *et al.* (2015) 'Milestones of Lynch syndrome 1895-2015', *Nat Rev Cancer*, 15, pp. 181–194.
- Lynch, H.T. and Smyrk, T. (1996) 'Hereditary nonpolyposis colorectal cancer (Lynch syndrome): An updated review', *Cancer*, pp. 1149–1167. Available at: [https://doi.org/10.1002/\(SICI\)1097-0142\(19960915\)78:6<1149::AID-CNCR1>3.0.CO;2-5](https://doi.org/10.1002/(SICI)1097-0142(19960915)78:6<1149::AID-CNCR1>3.0.CO;2-5).
- Mai, J. *et al.* (2023) 'Transcriptome-wide association studies: recent advances in methods, applications and available databases', *Communications Biology*. Nature Research. Available at: <https://doi.org/10.1038/s42003-023-05279-y>.
- Mamlouk, S. *et al.* (2020) 'Malignant transformation and genetic alterations are uncoupled in early colorectal cancer progression', *BMC Biology*, 18(1). Available at: <https://doi.org/10.1186/s12915-020-00844-x>.

- Manolio, T.A. *et al.* (2009) 'Finding the missing heritability of complex diseases', *Nature*, 461(7265), pp. 747–753.
- Marees, A.T. *et al.* (2018) 'A tutorial on conducting genome-wide association studies: Quality control and statistical analysis', *International Journal of Methods in Psychiatric Research*, 27(2), pp. 1–10. Available at: <https://doi.org/10.1002/mpr.1608>.
- El Marjou, F. *et al.* (2004a) 'Tissue-specific and inducible Cre-mediated recombination in the gut epithelium', *Genesis (United States)*, 39(3), pp. 186–193. Available at: <https://doi.org/10.1002/gene.20042>.
- El Marjou, F. *et al.* (2004b) 'Tissue-specific and inducible Cre-mediated recombination in the gut epithelium', *Genesis (United States)*, 39(3), pp. 186–193. Available at: <https://doi.org/10.1002/gene.20042>.
- Martínez-Jiménez, F. *et al.* (2020) 'A compendium of mutational cancer driver genes', *Nature Reviews Cancer*. Nature Research, pp. 555–572. Available at: <https://doi.org/10.1038/s41568-020-0290-x>.
- Masica, D.L. *et al.* (2017) 'CRAVAT 4: Cancer-related analysis of variants toolkit', *Cancer Research*, 77(21), pp. e35–e38. Available at: <https://doi.org/10.1158/0008-5472.CAN-17-0338>.
- Matsui, C. *et al.* (2008) 'Identification of a link between the SAMP repeats of adenomatous polyposis coli tumor suppressor and the Src homology 3 domain of DDEF', *Journal of Biological Chemistry*, 283(47), pp. 33006–33020. Available at: <https://doi.org/10.1074/jbc.M800420200>.
- McLaren, W. *et al.* (2016) 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17(1). Available at: <https://doi.org/10.1186/s13059-016-0974-4>.
- Meyerson, W. *et al.* (2020) 'Origins and characterization of variants shared between databases of somatic and germline human mutations', *BMC Bioinformatics*, 21(1), pp. 1–22. Available at: <https://doi.org/10.1186/s12859-020-3508-8>.
- Mitroi, A.F. *et al.* (2022) 'Association of TCF7L2, CASC8 and GREM1 Polymorphisms in Patients with Colorectal Cancer and Type II Diabetes Mellitus', *Genes*, 13(8). Available at: <https://doi.org/10.3390/genes13081297>.
- Molinari, F. and Frattini, M. (2014) 'Functions and regulation of the PTEN gene in colorectal cancer', *Frontiers in Oncology*, 4 JAN. Available at: <https://doi.org/10.3389/fonc.2013.00326>.
- Morgan, E. *et al.* (2023) 'Global burden of colorectal cancer in 2020 and 2040: Incidence and mortality estimates from GLOBOCAN', *Gut*, 72(2), pp. 338–344. Available at: <https://doi.org/10.1136/gutjnl-2022-327736>.
- Moser, A.R., Pitot, H.C. and Dove, W.F. (1990) *A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse*.
- Najumudeen, A.K. *et al.* (2024) 'KRAS allelic imbalance drives tumour initiation yet suppresses metastasis in colorectal cancer in vivo', *Nature Communications*, 15(1). Available at: <https://doi.org/10.1038/s41467-023-44342-4>.
- Nelson, C.M. (2021) 'The mechanics of crypt morphogenesis', *Nature Cell Biology*, 23(7), pp. 678–679. Available at: <https://doi.org/10.1038/s41556-021-00703-z>.
- Ng, P.C. and Henikoff, S. (2003) 'SIFT: Predicting amino acid changes that affect protein function', *Nucleic Acids Research*, 31(13), pp. 3812–3814. Available at: <https://doi.org/10.1093/nar/gkg509>.

- Nguyen, H.T. and Duong, H.Q. (2018) 'The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy (review)', *Oncology Letters*. Spandidos Publications, pp. 9–18. Available at: <https://doi.org/10.3892/ol.2018.8679>.
- Nicoletti, C. (2000) 'Unsolved mysteries of intestinal M cells', *Gut*, 47(5), pp. 735–739. Available at: <https://doi.org/10.1136/gut.47.5.735>.
- Niida, A. *et al.* (2021) 'Modeling colorectal cancer evolution', *Journal of Human Genetics*. Springer Nature, pp. 869–878. Available at: <https://doi.org/10.1038/s10038-021-00930-0>.
- Novellasmunt, L., Antas, P. and Li, V.S.W. (2015) 'Targeting Wnt signaling in colorectal cancer. A Review in the Theme: Cell Signaling: Proteins, Pathways and Mechanisms', *American Journal of Physiology Cell Physiology*, 309(8), pp. 511–521.
- Nowell, P.C. (1976) 'The clonal evolution of tumor cell populations', *Science* [Preprint]. Available at: <https://doi.org/10.1126/science.959840>.
- Palles, C., Latchford, A. and Valle, L. (2018) *Adenomatous polyposis syndromes: Polymerase proofreading-associated polyposis, Hereditary Colorectal Cancer: Genetic Basis and Clinical Implications*. Available at: https://doi.org/10.1007/978-3-319-74259-5_8.
- De Palma, F.D.E. *et al.* (2019) 'The molecular hallmarks of the serrated pathway in colorectal cancer', *Cancers*. MDPI AG. Available at: <https://doi.org/10.3390/cancers11071017>.
- Pedregosa, F. Varoquaux, G. Gramfort, A. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830.
- Pruim, R.J. *et al.* (2011) 'LocusZoom: Regional visualization of genome-wide association scan results', in *Bioinformatics*. Oxford University Press, pp. 2336–2337. Available at: <https://doi.org/10.1093/bioinformatics/btq419>.
- Purcell, S. *et al.* (2007) 'PLINK: A tool set for whole-genome association and population-based linkage analyses', *American Journal of Human Genetics*, 81(3), pp. 559–575. Available at: <https://doi.org/10.1086/519795>.
- Ramakrishnan, A.B. *et al.* (2023) 'SOX9 and TCF transcription factors associate to mediate Wnt/ β -catenin target gene activation in colorectal cancer', *Journal of Biological Chemistry*, 299(1). Available at: <https://doi.org/10.1016/j.jbc.2022.102735>.
- Rausch, T. *et al.* (2012) 'DELLY: structural variant discovery by integrated paired-end and split-read analysis.', *Bioinformatics (Oxford, England)*, 28(18), pp. i333–i339. Available at: <https://doi.org/10.1093/bioinformatics/bts378>.
- Rembacken, B.J. *et al.* (2000) 'Flat and depressed colonic neoplasms: A prospective study of 1000 colonoscopies in the UK', *Lancet*, 355(9211), pp. 1211–1214. Available at: [https://doi.org/10.1016/S0140-6736\(00\)02086-9](https://doi.org/10.1016/S0140-6736(00)02086-9).
- Rim, E.Y., Clevers, H. and Nusse, R. (2022) 'The Wnt Pathway: From Signaling Mechanisms to Synthetic Modulators', *Annual Review of Biochemistry*, (91), pp. 571–598. Available at: <https://doi.org/10.1146/annurev-biochem-040320>.
- Russo, A. *et al.* (2005) 'The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: Influence of tumor site, type of mutation, and adjuvant treatment', *Journal of Clinical Oncology*, 23(30), pp. 7518–7528. Available at: <https://doi.org/10.1200/JCO.2005.00.471>.

- Rutlin, M. *et al.* (2020) 'The Villin1 Gene Promoter Drives Cre Recombinase Expression in Extraintestinal Tissues', *CMGH*, 10(4), pp. 864-867.e5. Available at: <https://doi.org/10.1016/j.jcmgh.2020.05.009>.
- Sai Ge *et al.* (2022) 'Using genotype to assist clinical surveillance: a retrospective study of Chinese familial adenomatous polyposis patients', *Am J Cancer Res*, 12(9), pp. 4254–4266.
- de Sainte Agathe, J.M. *et al.* (2023) 'SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation', *Human Genomics*, 17(1). Available at: <https://doi.org/10.1186/s40246-023-00451-1>.
- Saito, T. *et al.* (2018) 'A temporal shift of the evolutionary principle shaping intratumor heterogeneity in colorectal cancer', *Nature Communications*, 9(1). Available at: <https://doi.org/10.1038/s41467-018-05226-0>.
- Salahshor, S. *et al.* (1999) 'Colorectal cancer with and without microsatellite instability involves different genes', *Genes Chromosomes and Cancer*, 26(3), pp. 247–252. Available at: [https://doi.org/10.1002/\(SICI\)1098-2264\(199911\)26:3<247::AID-GCC9>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1098-2264(199911)26:3<247::AID-GCC9>3.0.CO;2-H).
- Sanz-Pamplona, R. *et al.* (2015a) 'Exome sequencing reveals AMER1 as a frequently mutated gene in colorectal cancer', *Clinical Cancer Research*, 21(20), pp. 4709–4718. Available at: <https://doi.org/10.1158/1078-0432.CCR-15-0159>.
- Sanz-Pamplona, R. *et al.* (2015b) 'Exome sequencing reveals AMER1 as a frequently mutated gene in colorectal cancer', *Clinical Cancer Research*, 21(20), pp. 4709–4718. Available at: <https://doi.org/10.1158/1078-0432.CCR-15-0159>.
- Schneikert, J. *et al.* (2014) 'Different roles for the Axin interactions with the SAMP versus the second twenty amino acid repeat of adenomatous polyposis coli', *PLoS ONE*, 9(4). Available at: <https://doi.org/10.1371/journal.pone.0094413>.
- Seabold, S. and Perktold, J. (2010) 'statsmodels: Econometric and statistical modeling with python', *9th Python in Science Conference* [Preprint].
- Segditsas, S. and Tomlinson, I. (2006) 'Colorectal cancer and genetic alterations in the Wnt pathway', *Oncogene* [Preprint]. Available at: <https://doi.org/10.1038/sj.onc.1210059>.
- Segura, A.V.C. *et al.* (2023) 'Impact of mini-driver genes in the prognosis and tumor features of colorectal cancer samples: a novel perspective to support current biomarkers', *PeerJ*, 11. Available at: <https://doi.org/10.7717/PEERJ.15410>.
- Seshagiri, S. *et al.* (2012) 'Recurrent R-spondin fusions in colon cancer', *Nature*, 488(7413), pp. 660–664. Available at: <https://doi.org/10.1038/nature11282>.
- Seshan, E.V. and Karissa, W. (2023) *clinfun: Clinical Trial Design and Data Analysis Functions R Package version 1.1.1*.
- Shy, B.R. *et al.* (2013) 'Regulation of Tcf7l1 DNA Binding and Protein Stability as Principal Mechanisms of Wnt/ β -Catenin Signaling', *Cell Reports*, 4(1), pp. 1–9. Available at: <https://doi.org/10.1016/j.celrep.2013.06.001>.
- Smith, G. *et al.* (2002) 'Mutations in APC, Kirsten-ras, and p53 - Alternative genetic pathways to colorectal cancer', *Proceedings of the National Academy of Sciences of the United States of America* [Preprint]. Available at: <https://doi.org/10.1073/pnas.122612899>.

- Smits, R. *et al.* (1999) 'Apc1638T: A mouse model delineating critical domains of the adenomatous polyposis coli protein involved in tumorigenesis and development', *Genes and Development* [Preprint]. Available at: <https://doi.org/10.1101/gad.13.10.1309>.
- Soravia, C. *et al.* (1998) 'Genotype-phenotype correlations in attenuated adenomatous Polyposis coli', *American Journal of Human Genetics*, 62(6), pp. 1290–1301. Available at: <https://doi.org/10.1086/301883>.
- Sottoriva, A. *et al.* (2015) 'A big bang model of human colorectal tumor growth', *Nature Genetics* [Preprint]. Available at: <https://doi.org/10.1038/ng.3214>.
- Spain, S.L. and Barrett, J.C. (2015a) 'Strategies for fine-mapping complex traits', *Human Molecular Genetics*, 24(R1), pp. R111–R119. Available at: <https://doi.org/10.1093/hmg/ddv260>.
- Spain, S.L. and Barrett, J.C. (2015b) 'Strategies for fine-mapping complex traits', *Human Molecular Genetics*, 24(R1), pp. R111–R119. Available at: <https://doi.org/10.1093/hmg/ddv260>.
- Suto, R. *et al.* (2004) 'Dominant-negative mutant of c-Jun gene transfer: A novel therapeutic strategy for colorectal cancer', *Gene Therapy*, 11(2), pp. 187–193. Available at: <https://doi.org/10.1038/sj.gt.3302158>.
- Szewczyk, L.M. *et al.* (2024) 'Astrocytic β -catenin signaling via TCF7L2 regulates synapse development and social behavior', *Molecular Psychiatry*, 29(1), pp. 57–73. Available at: <https://doi.org/10.1038/s41380-023-02281-y>.
- Talseth-palmer, B.A. (2017) 'The genetic basis of colonic adenomatous polyposis syndromes', *Hereditary Cancer in Clinical Practice*, 15(5), pp. 1–7. Available at: <https://doi.org/10.1186/s13053-017-0065-x>.
- Tanaka, T. *et al.* (2006) 'Chromosome 18q deletion and Smad4 protein inactivation correlate with liver metastasis: A study matched for T- and N- classification', *British Journal of Cancer*, 95(11), pp. 1562–1567. Available at: <https://doi.org/10.1038/sj.bjc.6603460>.
- Tang, F. *et al.* (2022) 'The Role of Vti1a in Biological Functions and Its Possible Role in Nervous System Disorders', *Frontiers in Molecular Neuroscience*. Frontiers Media S.A. Available at: <https://doi.org/10.3389/fnmol.2022.918664>.
- Tang, W. *et al.* (2008) *A genome-wide RNAi screen for Wnt/-catenin pathway components identifies unexpected roles for TCF transcription factors in cancer*. Available at: www.pnas.org/cgi/content/full/.
- Tanneberger, K., Pfister, A.S., Brauburger, K., *et al.* (2011) 'Amer1/WTX couples Wnt-induced formation of PtdIns(4,5)P2 to LRP6 phosphorylation', *EMBO Journal*, 30(8), pp. 1433–1443. Available at: <https://doi.org/10.1038/emboj.2011.28>.
- Tanneberger, K., Pfister, A.S., Kriz, V., *et al.* (2011) 'Structural and functional characterization of the Wnt inhibitor APC membrane recruitment 1 (Amer1)', *Journal of Biological Chemistry* [Preprint]. Available at: <https://doi.org/10.1074/jbc.M111.224881>.
- Tanton, H. *et al.* (2022) *A novel β -catenin/BCL9 complex inhibitor blocks oncogenic Wnt signaling and disrupts cholesterol homeostasis in colorectal cancer*, *Sci. Adv.* Available at: <https://www.science.org>.
- Tenesa, A. *et al.* (2009) 'Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21', 40(5), pp. 631–637. Available at: <https://doi.org/10.1038/ng.133>. Genome-wide.

- Tomlinson, I. *et al.* (2007) 'A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21', *Nature Genetics*, 39(8), pp. 984–988. Available at: <https://doi.org/10.1038/ng2085>.
- Topol, L. *et al.* (2009a) 'Sox9 inhibits Wnt signaling by promoting β -catenin phosphorylation in the nucleus', *Journal of Biological Chemistry*, 284(5), pp. 3323–3333. Available at: <https://doi.org/10.1074/jbc.M808048200>.
- Topol, L. *et al.* (2009b) 'Sox9 inhibits Wnt signaling by promoting β -catenin phosphorylation in the nucleus', *Journal of Biological Chemistry* [Preprint]. Available at: <https://doi.org/10.1074/jbc.M808048200>.
- Tu, J. *et al.* (2019) 'The most common RNF43 mutant G659Vfs*41 is fully functional in inhibiting Wnt signaling and unlikely to play a role in tumorigenesis', *Scientific Reports*, 9(1). Available at: <https://doi.org/10.1038/s41598-019-54931-3>.
- Vaughan-Shaw, P.G. *et al.* (2021) 'Differential genetic influences over colorectal cancer risk and gene expression in large bowel mucosa', *International Journal of Cancer*, 149(5), pp. 1100–1108. Available at: <https://doi.org/10.1002/ijc.33616>.
- Virtanen, P. *et al.* (2020) 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nature Methods*, 17(3), pp. 261–272. Available at: <https://doi.org/10.1038/s41592-019-0686-2>.
- Vösa, U. *et al.* (2021) 'Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression', *Nature Genetics*, 53(9), pp. 1300–1310. Available at: <https://doi.org/10.1038/s41588-021-00913-z>.
- Wall, J.D. and Pritchard, J.K. (2003) 'Haplotype blocks and linkage disequilibrium in the human genome', *Nature Reviews Genetics*, 4(8), pp. 587–597. Available at: <https://doi.org/10.1038/nrg1123>.
- Wang, L. *et al.* (2023) 'Male-Biased Gut Microbiome and Metabolites Aggravate Colorectal Cancer Development', *Advanced Science*, 10(25). Available at: <https://doi.org/10.1002/advs.202206238>.
- Waskom, M. *et al.* (2017) 'mwaskom/seaborn: v0.8.1 (September 2017) (v0.8.1). Zenodo.'
- Weber, U., Paricio, N. and Mlodzik, M. (2000) 'Jun mediates Frizzled-induced R3/R4 cell fate distinction and planar polarity determination in the Drosophila eye', *Development*, 127(16), pp. 3619–3629.
- 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. Available at: <https://doi.org/10.1038/s41388-020-1259-7>.
- Williams, M.J. *et al.* (2016) 'Identification of neutral tumor evolution across cancer types', *Nature Genetics*, 48(3), pp. 238–244. Available at: <https://doi.org/10.1038/ng.3489>.
- Wu, M. *et al.* (2023) 'The Wnt-dependent and Wnt-independent functions of BCL9 in development, tumorigenesis, and immunity: Implications in therapeutic opportunities', *Genes and Diseases*. KeAi Communications Co., pp. 701–710. Available at: <https://doi.org/10.1016/j.gendis.2023.03.012>.
- Yang, J. *et al.* (2012) 'Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits', *Nature Genetics*, 44(4), pp. 369–375. Available at: <https://doi.org/10.1038/ng.2213>.

- Yang, J. *et al.* (2013a) 'Genome-Wide Complex Trait Analysis (GCTA): Methods, Data Analyses, and Interpretations. In Gondro, C., van der Werf J., Hayes, B. (eds) *Genome-Wide Association Studies and Genomic Prediction.*', *Methods in Molecular Biology*, 1019.
- Yang, J. *et al.* (2013b) 'Genome-Wide Complex Trait Analysis (GCTA): Methods, Data Analyses, and Interpretations. In Gondro, C., van der Werf J., Hayes, B. (eds) *Genome-Wide Association Studies and Genomic Prediction.*', *Methods in Molecular Biology*, 1019.
- Yao, Z. *et al.* (2017) 'Tumours with class 3 BRAF mutants are sensitive to the inhibition of activated RAS', *Nature* [Preprint]. Available at: <https://doi.org/10.1038/nature23291>.
- Yeh, C.H., Bellon, M. and Nicot, C. (2018) 'FBXW7: A critical tumor suppressor of human cancers', *Molecular Cancer*. BioMed Central Ltd. Available at: <https://doi.org/10.1186/s12943-018-0857-2>.
- Yiu, R. *et al.* (2005) 'Mechanisms of microsatellite instability in colorectal cancer patients in different age groups', *Diseases of the Colon and Rectum*, 48(11), pp. 2061–2069. Available at: <https://doi.org/10.1007/s10350-005-0171-0>.
- Zaidi, S.H. *et al.* (2020) 'Landscape of somatic single nucleotide variants and indels in colorectal cancer and impact on survival', *Nature Communications*, 11(1). Available at: <https://doi.org/10.1038/s41467-020-17386-z>.
- Zhang, B. *et al.* (2014) 'Large-scale genetic study in east Asians identifies six new loci associated with colorectal cancer risk', *Nature Genetics*, 46(6), pp. 533–542. Available at: <https://doi.org/10.1038/ng.2985>.
- Zhang, L. and Shay, J.W. (2017) 'Multiple Roles of APC and its Therapeutic Implications in Colorectal Cancer', *Journal of the National Cancer Institute*. Oxford University Press. Available at: <https://doi.org/10.1093/jnci/djw332>.
- Zhang, P. *et al.* (2018) 'Targeting CDK1 and MEK/ERK overcomes apoptotic resistance in BRAF-mutant human colorectal cancer', *Molecular Cancer Research* [Preprint]. Available at: <https://doi.org/10.1158/1541-7786.MCR-17-0404>.
- Zhao, H. *et al.* (2022) 'Wnt signaling in colorectal cancer: pathogenic role and therapeutic target', *Molecular Cancer*. BioMed Central Ltd. Available at: <https://doi.org/10.1186/s12943-022-01616-7>.
- Zhu, L. *et al.* (2021) 'APC Promoter Methylation in Gastrointestinal Cancer', *Frontiers in Oncology*. Frontiers Media S.A. Available at: <https://doi.org/10.3389/fonc.2021.653222>.