THE UNIVERSITY *of* EDINBURGH

DNA Methylation & its Regulation in Colorectal Tumours

Joseph Christopher Ward

Thesis Presented for the Degree of Doctor of Philosophy

The University of Edinburgh

December 2023

Declaration

I can confirm that this thesis is entirely of my own composition and the work presented within represents the research for which I have contributed significantly as a member of the Tomlinson Group. The contributions of others have been clearly indicated where appropriate.

I also confirm that this work has not been submitted in order to meet the requirements for any other degree or qualification.

Joe Ward.

Abstract

Colorectal cancer (CRC) is the fourth most common form of cancer in the UK and is associated with approximately 17,000 deaths each year. DNA methylation represents a critical epigenetic mechanism by which gene expression is controlled in cells, with aberrant DNA hyper-methylation often reported in CRC. Significant DNA hyper-methylation has been reported in a subset of CRCs and is referred to as the CpG Island Methylator Phenotype (CIMP). The precise mechanisms underlying CIMP and how DNA hyper-methylation contributes to colorectal tumorigenesis remains incompletely understood, therefore this thesis seeks to address both of these questions.

Firstly, Genome-Wide Association Study (GWAS) meta-analysis data revealed an association between the 4q24 locus of chromosome 4 and CRC predisposition in Europeans. Subsequent conditional analysis, fine-mapping and *in silico* functional annotation identified a list of candidate causal polymorphisms with functional relevance in colonic tissues. Mendelian Randomisation and genotype-expression analysis identified down-regulation of Ten-Eleven Translocation 2 (*TET2*) as the candidate causal factor underpinning the GWAS association with CRC risk at this locus. The *TET2* gene is involved in active DNA de-methylation, with pathogenic loss-of-function mutations thought to drive a hyper-methylated phenotype in acute myeloid leukaemia (AML). Similarly, pathogenic mutations in isocitrate dehydrogenase (IDH) 1 and 2 have also been reported in AML and present with a similar hyper-methylated phenotype. It has been hypothesised that 2-hydroxyglutarate produced by mutant IDH drives this hyper-methylation via the competitive inhibition of *TET2.* Mutations in *TET2* and IDH are uncommon in CRC but it is possible that mutations in these genes may drive CIMP in such cancers.

Methylation array data of colorectal adenocarcinomas from The Cancer Genome Atlas identified elevated DNA methylation in *TET2*-mutant and IDH-mutant CRCs compared to their wild-type (WT) counterparts, as well as a significant correlation between IDH mutations and $CIMP^+$ CRC. $CIMP^+$ CRCs presented with an enrichment of significantly hyper-methylated probes in bivalent promoter regions, with *TET2*-mutant and IDH-mutant cancers possibly showing an even greater enrichment at these regions than their WT $CIMP^+$ counterparts. A number of significantly hyper-methylated probes in *TET2*-mutant $CIMP^+$ cancers were also hyper-methylated in IDH-mutant $CIMP^+$ cancers, localising around the transcription start site of a number of candidate tumour suppressor genes – indicating *TET2* and IDH mutations may both drive $CIMP^+$ CRC via the aberrant hyper-methylation and transcriptional silencing of tumour suppressor genes.

5-methylcytosine (5-mC) regularly undergoes spontaneous deamination to thymine. The methyl-CpG binding domain 4 (*MBD4*) gene encodes a protein involved in catalysing the repair of this deamination by excising the thymine from the resultant T:G DNA mismatch. Germline inactivation of *MBD4* has been suggested to predispose affected individuals to intestinal polyposis and some forms of cancer, including uveal melanoma. This predisposition is potentially driven by pathogenic C → T mutations at CpG sites arising within the protein-coding sequence of cancer driver genes as a consequence of spontaneous deaminations of 5-mC going unrepaired. Individuals with germline pathogenic mutations in

MBD4 and subsequent somatic loss of heterozygosity were shown to have increased numbers of C → T mutations at CpG sites, including a number of pathogenic mutations in cancer driver genes that have been previously reported in the literature. It was also found that CpG sites that were highly-methylated and in later-replicating regions of the genome were at the greatest risk of spontaneous deamination, thereby describing *MBD4* as a cancer predisposition gene and how DNA methylation is mechanistically associated with C → T mutagenesis at CpG sites.

In addition to unrepaired spontaneous deaminations, errors made by DNA polymerase epsilon (*POL-ε*) during DNA replication are an alternative mechanism by which C → T mutations at CpG sites may arise. CRCs with *POL-ε* exonuclease domain mutations (EDMs) and microsatellite unstable (MSI^+) cancers presented with more C → T mutations at CpG sites than microsatellite stable (MSS) polymerase wild-type (POL-WT) CRCs. There was an excess of C → T mutations at CpG sites on the leading strand template in MSI^+ CRCs and cancers with *POL-ε* EDMs, indicating that these mutations were likely the result of unrepaired DNA mismatches arising as a consequence of the erroneous incorporation of adenine opposite a template 5-mC by *POL-ε* propagating into C → T mutations in the next round of DNA replication. Similarly to spontaneous deaminations, C → T mutagenesis was more common at highly-methylated CpG sites, indicating that DNA methylation also influences the likelihood of replication errors. Therefore, methylation-induced DNA replication error may represent another mechanism by which DNA methylation may contribute to colorectal tumorigenesis.

In conclusion, this thesis provides an in-depth investigation into the role of DNA methylation in CRC pathogenesis and suggests that DNA hyper-methylation may drive colorectal tumorigenesis via multiple mechanisms, including the direct modification of the expression of CRC driver genes or by indirectly driving C → T mutagenesis at CpG sites.

Lay Summary

Cancer represents one of the most pressing worldwide health concerns, with millions of new cases and disease-related deaths reported every year. Colorectal (bowel) cancer is the fourth most common cancer in the UK, responsible for approximately 17,000 deaths each year. Fundamentally, DNA in every cell of the body is an "instruction manual", containing the "instructions" for building proteins, which are what cells require in order to function correctly. Included in this "instruction manual" are proteins that control when cells divide, some proteins act to promote cell division whereas other proteins act to restrain it. Cancer is caused by the de-regulation of cell division, resulting in cells dividing out of control. This can be caused by alterations to the DNA "instruction manual" (mutations) that result in the production of either hyper-active proteins that promote cell division or non-functional proteins that can no longer restrict cell division.

In addition to this, different cells in different tissues in the body have different functions (e.g. brain cells vs skin cells) and therefore require different proteins to one another in order to perform these functions. Since the DNA "instruction manual" contains the "instructions" for every protein, cells need a way of being told which "pages" of the "instruction manual" are relevant for their function. DNA methylation represents one of these methods, which can instruct cells to "ignore" certain "pages" that are not relevant for their function. However, DNA methylation patterns in cells can also be altered in cancer, which can lead to cells "ignoring" the "instructions" for proteins that act to restrict cell division, potentially leading to cancer development. This DNA "hyper-methylation" has been observed in several cancer types, including bowel cancer.

This thesis seeks to explore the role of DNA methylation in bowel cancer development, starting with the exploration of the importance of a protein that regulates DNA methylation within cells (known as *TET2*) – and whether alterations to this protein may be associated with bowel cancer development. This thesis also explores the role of another protein that acts to maintain DNA methylation in cells (*MBD4*) in the development of intestinal polyps and other types of cancer. The ultimate goal of the exploration of these genes is to identify if inherited variation in either of these genes acts to predispose an individual to bowel cancer, potentially adding to the list of genes known to increase an individual's risk of cancer – thus widening genetic screening programmes. In addition to this, this thesis also seeks to explore how DNA methylation may also influence the likelihood of a mutation occurring and explores the potential mechanisms by which this mutation may occur, including "mistakes" made during DNA replication and the spontaneous degradation of DNA methylation within cells.

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.

All illustrations were made using BioRender (https://app.biorender.com/). In addition to this, there are a great number of people without whom this thesis would not have been possible. These include all the technicians from the Bioresearch & Veterinary Services facility at the Institute of Genetics & Cancer (University of Edinburgh) for their help with the animal work performed in this thesis. Furthermore, I would like to thank Professor Mark Arends (University of Edinburgh) for his technical advice throughout the project. I would also like to thank Dr Ioannis Kafetzopoulos (Babraham Institute) for providing me with replication timing reference data, Dr Claire Palles (University of Birmingham) and Dr Sara Galavotti (University of Birmingham) for performing the sample preparation for the *MBD4*$^{-/-}$ colorectal polyp whole-genome sequencing, Dr Ceres Fernández-Rozadilla (Instituto de Investigación Sanitaria de Santiago de Compostela) for the provision of the GWAS, TWAS and gene expression data used throughout Chapter II of this thesis, Dr Enric Domingo (University of Oxford) for the DNA methylation data used in Chapter III of this thesis and Dr Marketa Tomkova (University of Oxford) for providing the replication strand data used in Chapter V of this thesis.

I would also like to thank Dr Juan Fernández-Tajes (University of Oxford), Dr Steve Thorn (University of Oxford), Dr Albert Nobre de Menezes (formerly of The University of Edinburgh) and Dr Alison Meynert (University of Edinburgh) for their superb bioinformatics support throughout the duration of this project. I would like to acknowledge the role of Dr Kevin Myant (University of Edinburgh), Dr Andrew Sims (University of Edinburgh) and Professor Adele Murrell (University of Bath) for their advice throughout the project as members of my thesis committee. I am also grateful for the additional supervisory input of Dr Chiara Bardella (University of Birmingham) and Dr Lennard Lee (University of Oxford).

I am also extremely grateful to all members of the Tomlinson Group (past and present) for the advice, support and encouragement they provided throughout the duration of the project. These include: Nathalie Feeley, Dr Alina Finch, Harvinder Gala, Jurriaan van Ginkel, Dr Rachel Guest, Güler Gül, Chloe Henry, Tiffany Idle, Dr Emma Jaeger, Rosie Matthews, Dr Amy McCorry, Melissa Morgan, Dr Michael Nicholson, Dr Sophie Roper, Dr Archana Sharma-Oates, Kitty Sherwood, Dr James Wood and Connor Woolley. I would particularly like to thank Dr Ignacio Soriano (University of Oxford) and Dr Nagore De León (University of Oxford) for the day-to-day supervision, advice and support.

This project would not have been possible without my principal supervisory team of Professor Ian Tomlinson (University of Oxford) and Dr Duncan Sproul (University of

Edinburgh). I would therefore like to thank both Ian and Duncan for their guidance, insight, adaptability and, above all, patience throughout our time working together – I will forever be grateful to them both. I would also like to thank the Wellcome Trust, whose funding made this project possible.

Finally, I would like to thank my family – in particular my parents Su & Chris and my sisters Becky & Amy for their support that helped make this thesis possible.

List of Abbreviations

2-HG – 2-Hydroxyglutarate

5-aza – 5-aza-2-deoxycytidine

5-hmC – 5-hydroxymethylcytosine

5-mC – 5-methylcytosine

100KGP – The 100,000 Genomes Project

α-KG – α-Ketoglutarate

AML – Acute Myeloid Leukaemia

APC – Adenomatous Polyposis Coli

BER – Base Excision Repair

BHC – Benjamini-Hochberg Corrected

cCRE – Candidate Cis-Regulatory Element

CIMP – CpG Island Methylator Phenotype

COSMIC – Catalogue of Somatic Mutations in Cancer

CRC – Colorectal Cancer

DFCI – Dana-Farber Cancer Institute

DNMT – DNA Methyltransferase

EAS – East-Asian

EDM – Exonuclease Domain Mutation

ENCODE – The Encyclopaedia of DNA Elements

eQTL – Expression Quantitative Trait Loci

FAP – Familial Adenomatous Polyposis

FET – Fisher's Exact Test

GTEx – Genotype-Tissue Expression

GWAS – Genome-Wide Association Study

HMPS – Hereditary Mixed Polyposis Syndrome

IDH – Isocitrate Dehydrogenase

INDEL – Insertion / Deletion

JPS – Juvenile Polyposis Syndrome

LD – Linkage Disequilibrium

LoH – Loss of Heterozygosity

LS – Lynch Syndrome

MAF – Minor Allele Frequency

MBD4 – Methyl-CpG Binding Domain 4

MLH1 – MutL Homologue 1

MMR – Mismatch Repair

MR – Mendelian Randomisation

MSH2 – MutS Homologue 2

MSH6 – MutS Homologue 6

MSI – Microsatellite Instability

MSS – Microsatellite Stable

PAINTOR – Probabilistic Annotation Integrator

PJS – Peutz-Jeghers Syndrome

POL – DNA Polymerase

POL-δ – DNA Polymerase Delta

POL-ε – DNA Polymerase Epsilon

PPAP – Polymerase Proofreading-Associated Polyposis

SBS – Single-Base Substitution

S:CORT – Stratification in Colorectal Cancer: From Biology to Treatment Prediction

SCREEN – Search Candidate Cis-Regulatory Elements by ENCODE

SNP – Single-Nucleotide Polymorphism

SNV – Single-Nucleotide Variation

TCGA – The Cancer Genome Atlas

TCGA-COAD – The Cancer Genome Atlas (Colorectal Adenocarcinoma Project)

TCGA-READ – The Cancer Genome Atlas (Rectal Adenocarcinoma Project)

TDG – Thymine DNA Glycosylase

TET2 – Ten-Eleven Translocation 2

TWAS – Transcriptome-Wide Association Study

TWMR – Transcriptome-Wide Mendelian Randomisation

VCF – Variant Call File

WT – Wild-Type

Table of Contents

Table of Figures

List of Tables

Chapter I - Introduction

1.1 – Heritable Genetic Traits & Cancer Predisposition

Cancer represents one of the most significant challenges in worldwide healthcare, with an estimated twenty-million newly diagnosed cases and ten-million deaths globally in the year 2020 (1). The most commonly diagnosed cancers include lung, breast and colorectal – with lung cancer proving to be the leading cause of cancer-related deaths (1). Over recent years, owing to a combination of earlier detection, better treatment options and lifestyle alterations (e.g. a reduction in the number of people who smoke), the mortality rates of cancer have improved – saving an estimated three-million lives in the USA alone in the last thirty years(2). However, there are some cancer types whose prognoses have not improved in recent years, an example of this is pancreatic cancer – where the five-year survival rate remains below 10% (3). This is a consequence of the tumour developing asymptomatically before presenting clinically at a late stage or as a metastatic disease (3).

The fundamental nature of tumorigenesis is the transformation of a previously normal cell into a malignant cancer cell, driven primarily by the accumulation of genetic alterations (mutations) in key cancer-associated genes (4,5). Tumorigenesis almost always begins with the transformation of a single cell which, as a result of its mutations, has a selective growth advantage over neighbouring cells (5–7). In recent years, there has been an increasing emphasis on the paradigm that a tumour is not solely composed of malignant cells, but also contains multiple types of support cell in a system known as the tumour micro-environment (8,9). This tumour micro-environment acts to facilitate tumour growth through either promoting the proliferation of malignant cells, or by inhibiting their death by apoptosis (8,9).

Genetic mutations within cancer cells can broadly be characterised as inherited (germline) defects or acquired sporadically throughout life by random (somatic) mutation (10). The inheritance of a pathogenic mutation in a key cancer driver gene can markedly increase the risk of an individual developing cancer throughout their lifetime (11,12). These inherited elements are present within about 10% of all cancers (11,12). Some examples of germline mutations include breast cancer gene 1 (*BRCA1*) in breast and ovarian cancer, serine-threonine kinase 11 (*STK11*) in pancreatic cancer and adenomatous polyposis coli (*APC*) in colorectal cancer (13–16). Due to their elevated risk of tumorigenesis, the characterisation of new genes which carry a heritable cancer risk has been the subject of extensive research in recent years. This has led to the discovery of several so-called cancer predisposition syndromes, one such example being Lynch Syndrome (LS), which has been identified in colorectal and ovarian cancers (17). As a result of previous work in the field, genetic screening is now at the forefront of modern cancer care (12,18,19). Examples of this include screening for pathogenic mutations in the *BRCA1* or *BRCA2* genes in breast cancer, which are thought to account for a large proportion of hereditary cases (12,18,20). However, the identification of new cancer-predisposing genes remains of critical importance for effective cancer prevention strategies.

1.2 – The Molecular Pathogenesis of Colorectal Cancer

1.2.1 – Modifiable Risk Factors of Colorectal Cancer

Colorectal cancer (CRC) represents a major health burden worldwide, representing the fourth most common cancer in the UK, accounting for approximately 11% of newly-diagnosed cancers (21). In total, CRC is associated with nearly 17,000 deaths every year in the UK, with a five-year survival rate of 58.4% (21). The identification of several modifiable risk factors and the emergence of new therapeutics has led to a reduction in the mortality of CRC in recent years – for example the correlation between red meat consumption and CRC risk (see below) (22,23). CRC affects men and women at approximately equal rates, while geographical analysis has shown the incidences of CRC to be highest in Australia, the USA and many parts of Europe but lowest in India, China and some regions of South America (24,25). There is also a disparity between the incidence of CRC in developed countries in comparison to developing countries, with risk somewhere between three and four times higher in the former than the latter (26). It is possible that this disparity is a consequence of differences in life expectancy between developed and developing countries, given that CRC is most prevalent in people over the age of fifty, rather than differences in lifestyle (27). However, following age-standardisation, the incidence of CRC was still found to be highest in North American and European populations, whilst being lowest in African and South Asian populations (28,29)

As described above, there are several modifiable (lifestyle) factors that affect an individual's likelihood of developing CRC – with the majority of these being linked to diet (30–34). Obesity has been suggested to increase an individual's risk of CRC by as much as 19% (23). There have also been studies that have identified specific food items that carry with them significant associations with CRC development. For example, it has become well known that the consumption of red or processed meat significantly increases the risk of CRC development (35–38). Excessive alcohol consumption has also been suggested to increase CRC risk, while fruit and vegetable consumption have been inversely correlated with CRC risk (39–41). Other lifestyle factors that have been suggested to influence CRC risk include tobacco smoking (42). The study by Tsoi *et al.* investigated the effect of cigarette smoking on CRC risk in a cohort of nearly 1,500,000 participants, where the smoker group demonstrated a higher risk of CRC development than non-smokers (42).

In addition to these modifiable risk factors, there are a number of genetic elements that may influence the development of CRC – these CRC predisposition genes and syndromes will be discussed in detail in section 1.3.

1.2.2 – The Adenoma-Carcinoma Sequence

For many years, the accepted paradigm for the progression of CRC was thought to involve the sequential acquisition of mutations in a number of known CRC driver genes (43). These mutations would together drive the transition of normal intestinal epithelium into small, initially benign, polyps which can then evolve into more malignant, potentially metastatic, carcinomas (44). Over time, this sequential acquisition of mutations became known as the adenoma-carcinoma sequence, as summarised in Figure 1.1 (45,46). This model suggests that

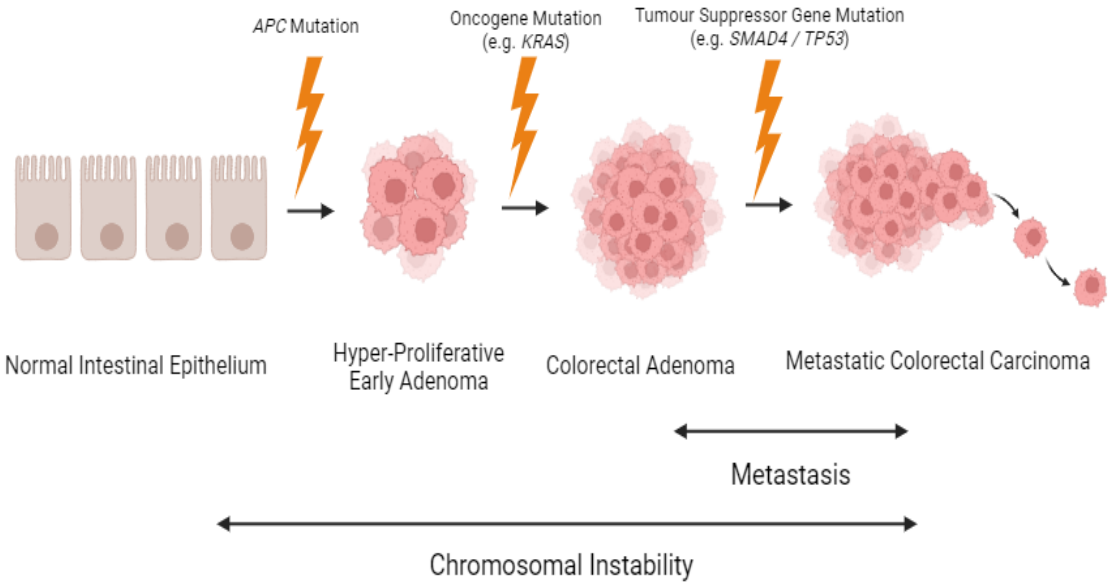


Figure 1.1 – The Adenoma-Carcinoma Sequence: A diagrammatic illustration of the adenoma-carcinoma sequence model of colorectal cancer development. Normal intestinal epithelium transitions into hyper-proliferative early adenomas via the acquisition of a mutation in the adenomatous polyposis coli (*APC*) gene. This early adenoma then becomes a late adenoma following mutations an oncogene such as *KRAS.* Finally, a late adenoma can become a metastatic colorectal cancer via mutations in the key tumour suppressor gene *TP53*. Created with BioRender.com (https://app.biorender.com/).

high-grade, invasive colorectal carcinomas always arise from lower-grade adenomas (47). Typically, high-grade metastatic carcinomas have a poor patient prognosis and only limited treatment options – with an average five-year survival of less than 15% (48).

As described by Fearon & Vogelstein, the adenoma-carcinoma sequence is initiated by a mutation in the *APC* gene, which is sometimes referred to as the "gatekeeper gene" of CRC due to its mutation being the initial step in tumorigenesis (49). Normally, the *APC* gene plays a critical role in the intestinal *Wnt* signalling pathway, which drives the proliferation and self-renewal of intestinal stem cells (50–52). The *Wnt* signalling pathway has been extensively characterised owing to its critical role in intestinal homeostasis and is the subject of several review articles (50,53,54). Briefly, in the absence of the extracellular *Wnt* signalling protein, the *APC* protein forms a "destruction complex" with *axin* and *GSK-3β*, which act together to phosphorylate the intracellular signalling protein *β-catenin,* consequently driving its proteasomal degradation (53,55). However, when the *Wnt* signalling protein binds to its target G-protein coupled receptor *Frizzled* on the surface of cells, this destruction of *β-catenin* does not occur (56). Following receptor binding, via the activity of partner proteins including *Dishevelled*, the destruction complex is disassembled – allowing *β-catenin* to persist within cells. As a result of this, *β-catenin* is then able to migrate into the nucleus and induce the transcription of several genes involved in cellular proliferation and intestinal stem cell renewal (53,57). Therefore, the *APC* protein (alongside its partner proteins) is critical for ensuring there is tight regulation of the *Wnt* signalling pathway by preventing *β-catenin* dependent downstream signalling in the absence of the appropriate extracellular stimulation (58). In the event the *APC* protein is inactivated, potentially as a consequence of a pathogenic

mutation in the *APC* gene, *β-catenin* is not marked for degradation in the proteasome in the absence of the *Wnt* signalling protein, allowing it to exert its effects on cellular transcription without the appropriate upstream stimulus (57). This constitutive *Wnt* signalling is therefore able to drive inappropriate cellular proliferation within the intestinal compartment. Furthermore, inactivation of the *APC* protein has been suggested to disrupt normal intestinal tissue architecture via the weakening of inter-cellular adherens junctions – of which *β-catenin* represents an essential component (59,60). A summary of the *Wnt* signalling pathway is presented in Figure 1.2.

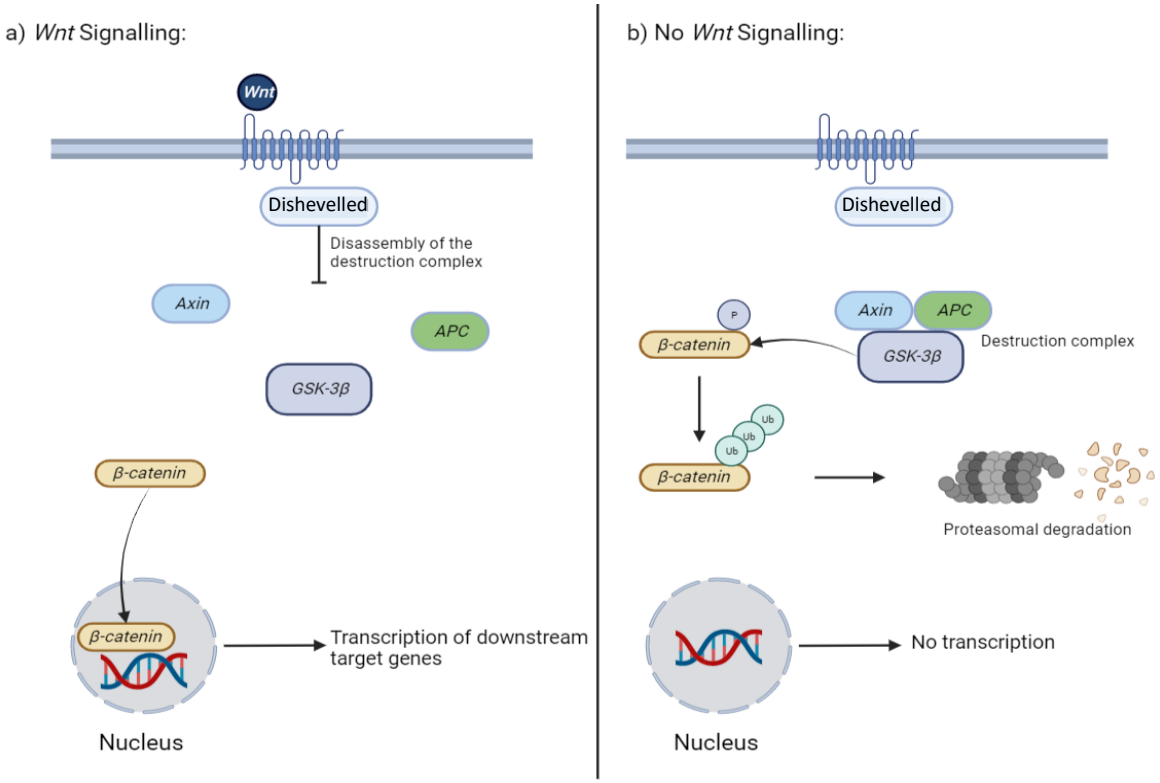


Figure 1.2 – An Overview of the *Wnt* Signalling Pathway: A schematic illustration of the control of intracellular *β-catenin* signalling via the *Wnt* signalling pathway. a) In the presence of extracellular *Wnt* signalling the *β-catenin* destruction complex is disassembled, allowing *β-catenin* migration into the nucleus and subsequent transcription of target genes. b) In the absence of *Wnt* signalling, *β-catenin* is marked for proteasomal degradation via the destruction complex, preventing downstream *β-catenin* mediated gene transcription. Created with BioRender.com (https://app.biorender.com/).

The next stage of the adenoma-carcinoma sequence requires the mutation of an oncogene, thus acting to further drive inappropriate cellular proliferation (61). Common oncogenes mutated in CRC include *KRAS* and *BRAF* – both components of the mitogen-activated protein (MAP) kinase signalling pathway. Briefly, activation of extracellular growth factor receptors induces the activation of *Ras* signalling proteins in a GTP-dependent manner. Activated *Ras* is then able to in turn activate *Raf* kinase family proteins which subsequently

activate the downstream effector *Mek* via phosphorylation which then itself phosphorylates *Erk* via its own intrinsic kinase activity (62,63). Activated *Erk* is then translocated into the nucleus where it is able to induce the transcription of several downstream target genes involved in cellular growth, proliferation and survival (63). A summary of the MAP kinase pathway is provided in Figure 1.3.

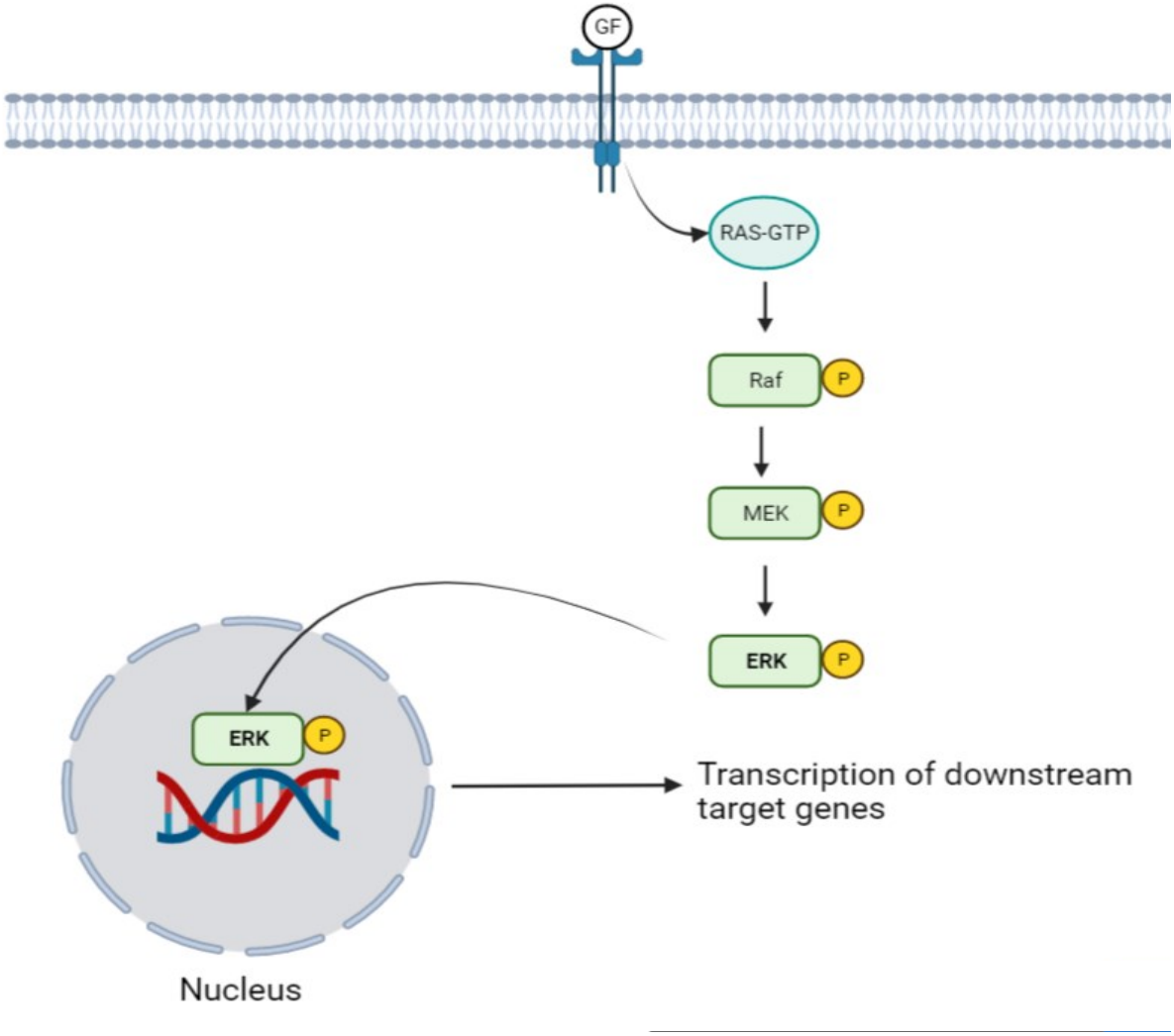


Figure 1.3 – The MAP Kinase Pathway: A summary of the mitogen-activated protein (MAP) kinase intracellular signalling pathway. The binding of an extracellular growth factor (GF) induces the GTP-dependent activation of *Ras* proteins, which in turn phosphorylate *Raf* proteins. Activated *Raf* subsequently phosphorylates *Mek* which in turn phosphorylates *Erk.* Activated *Erk* is then able to translocate to the nucleus and induce the transcription of downstream target genes. Created with BioRender.com (https://app.biorender.com/).

Mutations in the *KRAS* gene, commonly at codons twelve and thirteen, are found in around one third of all CRC cases (61). Mutations in a member of the *Ras* family have previously been studied and identified in 57% of carcinomas and nearly half of large adenomas, but only in 9% of small adenomas (47). This suggests that mutations in oncogenes are required to drive the adenoma-carcinoma sequence following the initial mutation in the *APC* gene (47).

Another commonly mutated oncogene is *BRAF,* which when mutated has been shown to drive the abnormal proliferation of intestinal cells (64). Mutations in *BRAF* have been identified in approximately 10% of all CRCs and also in a larger proportion of melanomas (65,66). The most common mutation in *BRAF* identified in these cancers is $BRAF^{V600E}$, which is a target for therapeutic intervention in melanoma through the *BRAF* inhibitor vemurafenib. The importance of BRAF mutations in CRC pathogenesis will be discussed in section 1.2.3. Mutations in one of these oncogenes is thought to drive the progression of cells through the next stage of the adenoma-carcinoma sequence, driving the formation of a colorectal adenoma.

The later stages of the adenoma-carcinoma sequence are thought to drive the progression of a colorectal adenoma into an invasive, potentially metastatic, colorectal carcinoma. This is achieved through the mutation of one of a number of key tumour suppressor genes – the most common examples being *SMAD4* and *TP53* (43,47,67,68). *SMAD4* mutations have been identified in approximately 20% of CRCs and in 30% of pancreatic adenocarcinomas – and are associated with a poor patient prognosis (69,70). The *SMAD4* protein is involved in regulating the transforming growth factor beta (*TGF-β*) and bone morphogenetic protein (BMP) signalling pathways via heterodimerisation with other members of the SMAD family (70). The *TGF-β* pathway controls several key intracellular processes, including cellular growth, apoptosis and migration (71). Upon binding of the extracellular *TGF-β* or BMP signalling proteins to extracellular receptors, receptor-associated members of the SMAD protein family (*R-SMAD*) are phosphorylated, undergo a conformational change and form a heterodimer with the common mediator *SMAD4* (71). This complex then migrates to the nucleus and induces the transcription of downstream target genes (70,71). *SMAD2* and *SMAD3* represent the *R-SMAD* proteins responsible for transducing *TGF-β* pathway signalling and *SMAD1, SMAD5* and *SMAD8* are BMP-associated (70,71). Canonical *SMAD4*-associated signalling functions at the G_1/S-phase checkpoint of the cell cycle, driving cell cycle arrest or apoptosis via inducing the transcription of cell cycle inhibitors (e.g. *p21*) (71,72). The role of SMAD proteins in the *TGF-β* and BMP signalling pathways is summarised in Figure 1.4.

The functions of the *TP53* protein have been the subject of extensive study. Following the discovery of its essential role within cells in the suppression of malignancy, the *TP53* protein is now commonly referred to as the "Guardian of the Genome" (73,74). The role of *TP53* as a tumour suppressor gene is extensive, from inducing cell cycle arrest to initiating DNA repair or triggering apoptosis (75–77). The expression of *TP53* is induced by a number of cellular stress factors, including DNA damage or aberrant oncogene activation (78). Following cellular stress, *TP53* goes on to induce the transcription of the cell cycle arrest protein *p21* – which is able to halt cell cycle progression via the inhibition of key cyclin-dependent kinase (CDK) proteins *CDK2* and *CDK4* (78). *p21* is also able to halt the process of DNA replication via interactions with proliferating cell nuclear antigen (*PCNA*) (79). In addition to this role in inducing cell cycle arrest, *TP53* also has roles in more permanent cellular

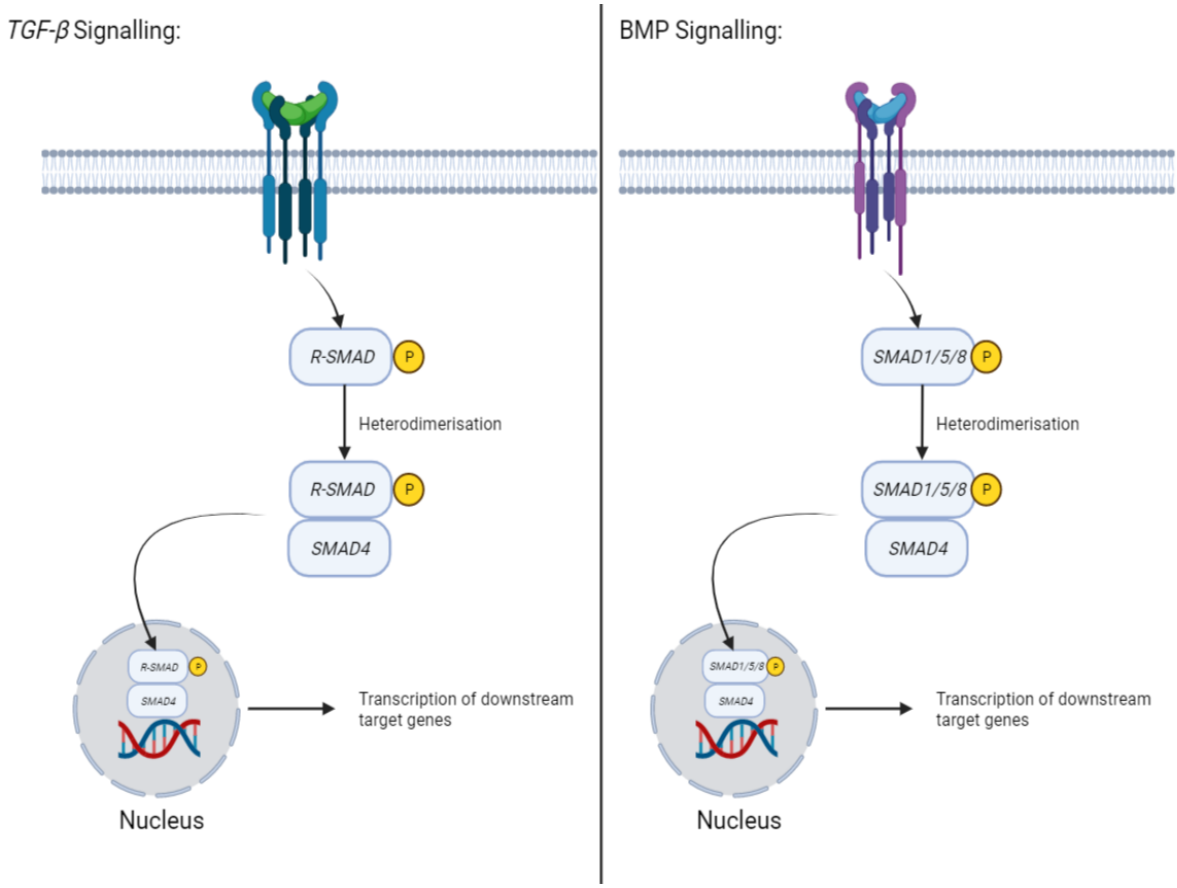


Figure 1.4 – The Role of SMAD Proteins in the *TGF-β* & BMP Signalling Pathways: A summary of the role of the SMAD proteins in the transforming growth factor β (*TGF-β*) and bone morphogenetic protein (BMP) signalling pathways. In the *TGF-β* pathway (left), binding of the *TGF-β* signalling protein to its receptor drives the phosphorylation of receptor SMAD (*R-SMAD*) proteins, which heterodimerise with *SMAD4* and translocate to the nucleus in order to induce the transcription of downstream target genes. In the BMP pathway (right), the binding of extracellular signalling proteins to their receptor protein triggers a similar cascade, using *SMAD1, SMAD5* or *SMAD8* in place of *R-SMAD* proteins. Created with BioRender.com (https://app.biorender.com/).

responses to stressors – including permanent cell senescence and apoptosis (78). Apoptosis can be triggered in mammalian cells via either the intrinsic mitochondrial pathway or extrinsic death-receptor pathway (80). In the intrinsic pathway, pro-apoptotic proteins are released by the mitochondria in response to pro-apoptotic signalling from members of the *BCL2* family, which consequently results in caspase protein activation and cell death (80). Alternatively, the extrinsic pathway functions via the *Fas* ligand (*Fas-L*) interacting with the *Fas* receptor on the surface of target cells. Following this interaction, initiator caspases are activated which in turn activate executor caspases – thus driving cellular apoptosis (81). The *TP53* protein has been suggested to influence both pathways by inducing the transcription of pro-apoptotic proteins to activate the intrinsic apoptosis pathway and also up-regulate the expression of the *Fas* receptor on the surface of the cell – increasing the likelihood of apoptosis via the extrinsic pathway (80,81). Given the critical roles outlined above, it is perhaps not surprising that *TP53* mutations are present in up to 50% of all CRC cases (68). Following the loss of these key tumour suppressor genes, cells lose control of cell cycle arrest, apoptosis and, in the case of the *TGF-β* pathway, cellular migration (82,83). This can

lead to the progression from an adenoma to a carcinoma, which may either exist *in situ* or have the potential to develop into metastatic disease. Potential sites of metastatic disease include the liver, lung or nervous system (84). Between 15-25% of CRC patients present with metastatic disease, with variable clinical implications depending on the site of metastasis (85).

Alongside direct mutation of the key driver genes listed above, there are a number of other factors which also drive cells through the adenoma-carcinoma sequence. As illustrated in Figure 1.1, increasing genomic instability (either at the microsatellite or chromosomal level – see section 1.2.3) and angiogenesis are examples of some of the key cellular processes that are disrupted throughout the adenoma-carcinoma sequence (86,87). The critical nature of these processes in driving tumorigenesis is exemplified by the fact that both are referred to as "Hallmarks of Cancer" (9). Alterations of cellular DNA methylation patterns also represents a key mechanism of tumorigenesis in CRC, potentially by aberrantly modulating the transcription of key tumour suppressor genes. The role of DNA methylation in the regulation of gene expression will be discussed in detail in section 1.4. Briefly, aberrant promoter hyper-methylation of cancer-associated tumour suppressor genes has been suggested to play a key role in CRC pathogenesis (88,89). For example, nearly 20% of CRCs lacking a pathogenic *APC* mutation may instead present with *APC* promoter hyper-methylation, representing an alternative mechanism for *APC* inactivation in CRC (88).

1.2.3 – The Serrated Pathway of Colorectal Cancer

The adenoma-carcinoma sequence represents the predominant mechanism for CRC pathogenesis, accounting for approximately 70-80% of cases (90). These adenoma-carcinoma sequence driven tumours are characterised by genomic instability at the chromosomal level but present with few mutations at microsatellite regions of the genome (microsatellite stable – MSS) (90). However, while these tumours represent the most common form of CRC, there are other pathways that can drive CRC tumorigenesis (90,91). In the early 1990s, Jass & Smith reported mucinous hyperplastic colorectal polyps with a serrated (saw-like) appearance (92). Serrated adenomas represent multiple distinct sub-types of polyps, including traditional serrated adenomas, sessile serrated adenomas, mixed polyps and hyperplastic polyps (93,94). For many years, these serrated polyps were thought to be benign, with little tumour-initiating potential (91). However, following more in-depth analysis of these polyps revealed an alternative mechanism of CRC pathogenesis to the widely-accepted adenoma-carcinoma sequence (90,91,93).

These serrated adenomas are characterised by microsatellite instability (MSI), which describes hyper-mutation of microsatellite regions of the genome (93,95). MSI is a feature of approximately 15% of all CRCs and is the consequence of pathogenic mutations in components of the DNA mismatch repair (MMR) pathway (see sections 1.3.2 and 1.4.3 for a detailed description of MSI in CRC) (95,96). Interestingly, serrated adenomas have been shown to harbour a different profile of driver mutations to adenomas driven by the adenoma-carcinoma sequence, with these different drivers disrupting the same intracellular signalling pathways (91,93,97). For example, a number of serrated adenomas also harbour pathogenic *BRAF* mutations, as opposed to pathogenic *KRAS* mutations which characterise CRCs arising

from the adenoma-carcinoma sequence (91,93). Despite affecting different genes, both of these mutations drive tumorigenesis via disrupting the MAP kinase pathway (see Figure 1.3) (91,93,98). Furthermore, while mutations in SMAD proteins are common in the adenoma-carcinoma sequence, serrated adenomas often present with activating mutations in the *TGF-β* receptor *TGFBR2,* with both mutations acting to activate the *TGF-β* pathway (see Figure 1.4) (97). Finally, serrated adenomas often harbour pathogenic mutations in the pro-apoptotic gene *BAX,* resulting in the same resistance to apoptosis displayed by adenoma-carcinoma sequence cancers with pathogenic *TP53* mutations (see section 1.2.2) (91). Perhaps the most unique feature of serrated adenomas compared to those arising as a part of the adenoma-carcinoma sequence is the presence of abnormal patterns of DNA hyper-methylation (90,93). The role of DNA methylation in regulating gene expression will be described in detail in section 1.4.2. Briefly, abnormal hyper-methylation of genomic CpG islands is a key characteristic of the CpG island methylator phenotype (CIMP), the precise causes and importance of which are still yet to be fully understood (see section 1.4.4) (99–101). Overall, serrated adenomas are estimated to account for somewhere between 15-30% of all CRCs, providing an alternative model of tumorigenesis to the adenoma-carcinoma sequence (90). A summary of the serrated adenoma sequence of colorectal cancer is provided in Figure 1.5.

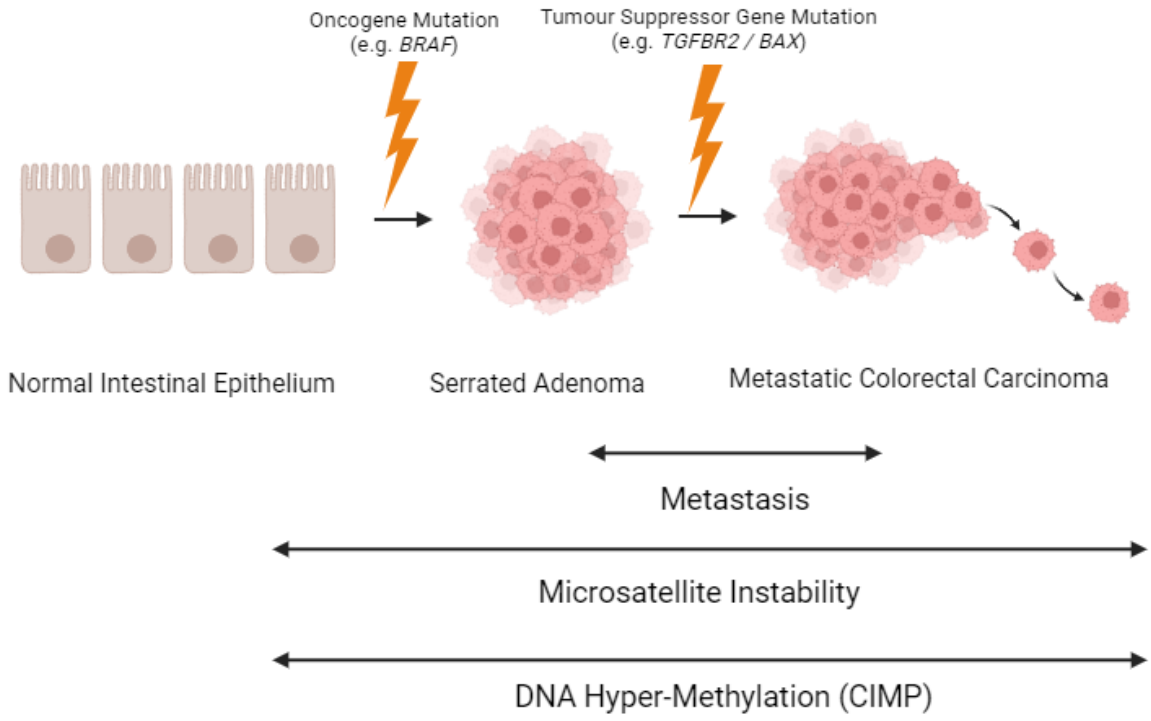


Figure 1.5 – The Serrated Pathway of Colorectal Cancer: A summary of the serrated pathway of colorectal cancer development. Previously normal intestinal epithelia can transition into a serrated adenoma following pathogenic mutations in the oncogene *BRAF.* These adenomas can subsequently develop into potentially metastatic carcinomas via mutations in *BAX* and *TGFBR2.* Accompanying these changes are DNA hyper-methylation, a feature of the CpG island methylator phenotype (CIMP) and hyper-mutation of microsatellite regions of the genome, resulting in microsatellite instability. Created with BioRender.com (https://app.biorender.com/).

1.2.4 – Intra-Tumour Heterogeneity in Colorectal Cancer

As previously noted, as cells progress towards malignancy, there is a marked increase in genomic instability that accompanies this progression (102). As described above, this genomic instability can occur at the chromosomal (adenoma-carcinoma sequence) or microsatellite (serrated pathway) level (90). As a consequence of this genomic instability, there is often a large degree of intra-tumour heterogeneity (ITH) within CRCs – resulting in the development of several distinct sub-clones with their own unique profile of mutations (103,104). ITH has been extensively investigated in recent years due to its importance predicting a patient's response to anti-cancer therapies (105–107). The study by Losi *et al.* suggested that ITH could be found in as many as 90% of CRCs, suggesting that the landscape of mutations within tumour cells is not homogenous in CRC (108). The relatively linear nature of the adenoma-carcinoma sequence, where the sequential acquisition of mutations in characterised driver genes is thought to drive tumour progression, does little to explain the apparent ITH present within CRCs (109,110). As a result of this, a number of alternative models of CRC tumorigenesis have aimed to explain ITH and expand what is known about CRC pathogenesis from the adenoma-carcinoma sequence.

An example of these alternative non-linear models of CRC progression is the "Big Bang" model, developed by Sottoriva and colleagues, which postulates that in CRCs develop as a single entity populated by a series of genetically distinct sub-clones (111). The model suggests that mutations can be divided into "public" and "private" classes, where "public" mutations are common to all sub-clones of a tumour and can be traced back to the tumour-initiating cell, the common ancestor of all sub-clones within the tumour (111). On the other hand, "private" mutations are unique mutations found within individual sub-clones as a result of their own distinct mutational processes and offer no selective advantages over other sub-clones due to the rapidly-expanding nature of the tumour (111). Therefore, it is assumed that these "private" mutations drive the massive ITH seen within CRCs and that "public" mutations in driver genes occur early in tumour development (109,111). For example, in the study by Gerlinger *et al.,* whole-exome sequencing of multiple distinct regions of a renal tumour demonstrated that up to 69% of all somatic mutations were not conserved across all sub-clones (112). The implication of this is that the prevalence of a specific sub-clone within a tumour is governed by the time at which the sub-clone first appeared and not by any selective advantages conferred to the sub-clone by its mutational profile (113). Consequently, according to the "Big Bang" model, the most prevalent sub-clones within a tumour are not the fittest, but the oldest, this concept of tumour evolution not being driven by selective sweeps has become known as neutral growth (111). This model may partially explain why some CRCs do not reach the final stage of the adenoma-carcinoma sequence and develop into metastatic disease (109). The suggestion from the "Big Bang" model is that all driver gene mutations are present at the initiation of the cancer and are common to all sub-clones of the tumour, therefore if mutations in drivers that promote metastasis are not present at tumour initiation, a tumour will never be able to metastasise (109,114). Alternatively, if at tumour initiation there are mutations in driver genes that promote metastasis, it is likely the resulting cancer will develop into metastatic disease as a result of the "*born to be bad"* model

(109,111). Overall, this model suggests that the massive ITH within CRCs is primarily driven by "private" mutations within each sub-clone which confer no selective advantage over other sub-clones – promoting tumour evolution by neutral growth.

Other non-linear models of CRC progression suggest that not all driver gene mutations are "public" – and therefore may not be present within all sub-clones of a tumour (115). van Ginkel *et al.* describe this concept as "parallel evolution", where sub-clones with distinct driver gene profiles can spatially co-exist within the same tumour (115). It is not clear what may cause this divergence in sub-clonal driver mutation profiles, but it is assumed that the different driver mutation profiles of an individual sub-clone confers no selective advantage over other sub-clones within the same tumour – meaning no sub-clones are eradicated by Darwinian selection (115). It is thought that parallel evolution is more common in early colorectal adenomas, which develop into CRC via the emergence of a dominant sub-clone – however it is again unclear how this sub-clone emerges given the aforementioned lack of selective advantage over other sub-clones (115).

1.3 – Colorectal Cancer Predisposition Syndromes

As previously discussed in section 1.2.1, there are a number of modifiable factors that affect the likelihood of an individual developing CRC – with many of these linked to diet and physical activity. While it is generally accepted that many CRCs are driven by somatic mutations in key driver genes, it is also apparent that underlying genetic predisposition is the driving force in a number of CRCs (116,117). An estimated 5% of CRCs are the result of germline mutations in known autosomal-dominant cancer predisposition syndromes, while it has been suggested that 20-25% of CRCs have an underlying genetic association (116,117). This means that our current understanding of the mechanisms underlying germline predisposition to CRC remains incomplete and more emphasis should be placed on identifying new genes associated with CRC predisposition.

1.3.1 – Familial Adenomatous Polyposis

Perhaps the most well-described of all the CRC predisposition syndromes is familial adenomatous polyposis (FAP), an autosomal-dominant syndrome first described as early as the 1950s by Gardner and colleagues (16). FAP is the result of a germline pathogenic mutation in the *APC* gene and is the most common polyposis syndrome of the gastrointestinal tract – affecting approximately 1 in 10,000 people but accounting for less than 1% of CRCs (118).

FAP has been characterised by the development of several, initially benign, adenomatous polyps of the colon which can increase in number with age – potentially resulting in hundreds of intestinal polyps (119). These polyps have the potential to develop into malignant colonic tumours if not properly managed or monitored, with CRC development common in FAP patients before the age of forty (120). As discussed in section 1.2.2, the *APC* gene plays a crucial role in the *Wnt* signalling pathway and its mutation is associated with the first step of

tumour development according to the adenoma-carcinoma sequence. Alongside the development of these intestinal polyps, FAP patients may also present with abnormalities of the bone, thyroid and soft tissue (121). The severity of FAP is known to vary depending on the location of the inherited *APC* mutation. Mutations which affect a region of the protein between codons 1,250 and 1,464 – especially at codon 1,309 – are associated with a more severe form of FAP, whereas mutations at either terminus of the gene are associated with more mild disease (122).

Since the discovery of FAP and the associated genetic alterations which underpin the syndrome, screening of at risk individuals has reduced the number of FAP patients presenting with CRC at first clinical evaluation by 55% (121). While this has improved the overall survival of FAP patients, the risk of developing CRC remains high (121). Alongside enhanced colonic surveillance of FAP patients, there are more extreme alternative treatments – including prophylactic surgical removal of the colon – that may be of clinical benefit. The two major procedures involved here are colectomy with ileorectal anastomosis and proctocolectomy with ileal pouch anal anastomosis (122). The decision on which procedure is used is influenced by a number of factors, including the patient's age and the severity of the polyposis (122).

1.3.2 – Hereditary Non-Polyposis Colorectal Cancer

Another CRC predisposition syndrome that has been the subject of extensive research is hereditary non-polyposis colorectal cancer (HNPCC), otherwise referred to as Lynch Syndrome (LS). LS is an autosomal-dominant syndrome which predisposes individuals to the development of CRC and was first described in the 1960s (123). In the study by Hampel *et al.,* it was shown that LS could be identified in 2.2% of patients in a CRC cohort of over 1,000 patients (124). Similarly, the study by Haraldsdottir *et al.* concluded that 1.8% of CRCs were a result of LS, again from a cohort of over 1,000 patients (125). Therefore, a number of studies have concluded that LS is the most common form of inherited CRC (125,126).

LS is the result of germline pathogenic mutations in genes involved in the DNA MMR pathway, including MutL homologue 1 (*MLH1*), MutS homologue 2 (*MSH2*) or 6 (*MSH6*) and *PMS2*. Briefly, *MSH2* and *MSH6* form a heterodimer involved in the initial recognition of a DNA mismatch, which allows the recruitment of the *MLH1-PMS2* heterodimer to the mismatch and subsequent nicking of the DNA around the mismatch (127). Following this, *EXO1*-mediated removal of mispaired nucleotides and re-synthesis and ligation of the DNA completes the MMR process (127). Inactivation of MMR pathway genes, via pathogenic mutation or epigenetic silencing (see section 1.4.4) are associated with the development of MSI, a characteristic identified in approximately 15% of all CRCs – whilst also being identified in endometrial and gastric cancers (128). As seen in Figure 1.1, MSI is a feature of the adenoma-carcinoma sequence and increases genomic instability, a critical component of colorectal tumorigenesis. The role of the MMR pathway in CRC pathogenesis will be investigated in more detail in Chapter V of this thesis.

In order to diagnose LS, a patient must meet what is referred to as the Amsterdam Criteria. The second generation of these Amsterdam Criteria have an estimated sensitivity of detecting

LS of 78% ad an estimated specificity of 61% (129,130). The Amsterdam Criteria are summarised in Table 1.1. In addition to this, the Bethesda Guidelines were published in the 1990s to act as a guide on which CRC patients should be sent for MSI testing. Patients who then tested positive for MSI could then be sent for specific MMR testing, in theory leading to the effective diagnosis of LS (131). When first implemented, the Bethesda Guidelines demonstrated a sensitivity of 94% but a relatively poor specificity of only 25% (130). A refined set of Bethesda Guidelines were subsequently issued and are summarised in Table 1.2.

Amsterdam Criteria II:	
Criterion:	Description:
Family History	At least three relatives with CRC or LS-associated cancer
Relatedness	At least one affected relative should be first-degree
Generations	At least two successive generations should be affected
Age	At least one relative should have been diagnosed before the age of fifty
FAP	FAP-associated CRCs should be excluded from the analysis
Tumour Verification	Cancers should be histologically verified

Table 1.1 – The Amsterdam Criteria of Lynch Syndrome Diagnosis: The second generation of Amsterdam Criteria followed for diagnosing Lynch Syndrome (LS). According to the criteria, at least three relatives across at least two successive generations, one of whom is a first-degree relative, must have been diagnosed with colorectal cancer (CRC) or a LS-associated cancer. One of these cancers should be in a patient under fifty years old. Familial adenomatous polyposis (FAP) must be excluded from the analysis and all tumours should have received histological verification. If all of these criteria have been met, a CRC patient can be diagnosed with LS (132).

Revised Bethesda Guidelines:	
Criterion:	Description:
Age	CRC must be diagnosed before the age of fifty
Tumour Type	Patient presents with multiple LS-associated tumours, either at the same time or at separate time points (regardless of age)
MSI Phenotype	Patient presents with LS-associated tumour with MSI characteristics before the age of sixty
Family History	At least one first-degree relative has LS-associated tumour diagnosed before the age of fifty
Family History	CRC diagnosed in at least two first or second-degree relatives with LS-associated tumours (regardless of age)

Table 1.2 – The Revised Bethesda Guidelines: The revised guidelines on which newly-diagnosed colorectal cancer (CRC) patients should be sent for microsatellite instability (MSI) testing. According to the guidelines, any CRC patient who is diagnosed before the age of fifty should be sent for screening. If a patient presents with multiple Lynch Syndrome (LS) associated tumours throughout their life, they should also be sent for screening. If the newly-diagnosed CRC presents with MSI characteristics, the age limit to be sent for testing is raised to sixty. Other criteria for sending patients for MSI screening include having a first–degree relative diagnosed with an LS-associated tumour before the age of fifty or having two or more first or second-degree relatives with LS-associated tumours also diagnosed with CRC. Following MSI screening, patients can also be screened for LS (131).

On average, patients with LS present clinically at a much younger age than sporadic CRC, which most commonly presents in patients over the age of fifty and is most common in those over the age of seventy-five (26). In LS, the age of disease presentation can vary depending on which MMR gene is mutated. In addition to this, LS patients have different prognoses depending on the MMR gene mutated (126). LS patients with mutations in *MLH1* or *MSH2* have a mean CRC diagnosis age of forty-four, whereas LS-associated CRCs with *MSH6* mutations do not typically develop until between forty-two and sixty-nine years of age (133). LS-associated CRC with *PMS2* mutations have an average diagnosis age of between sixty-one and sixty-six (133). Similarly, it is estimated that the lifetime risk of developing CRC in LS patients also varies depending on the mutated gene. Patients with *MLH1* or *MSH2* mutations have a lifetime risk of developing CRC of 30-74%, whereas this risk drops to between 10-22% in *MSH6*-mutant patients and 15-20% in patients with *PMS2* mutations (126). LS patients are thought to progress through the adenoma-carcinoma sequence of CRC tumorigenesis at a more rapid rate than sporadic CRCs, on average taking just five years compared to potentially decades in sporadic CRCs (134). It is because of this that management of LS is biannual colonic surveillance, often beginning between twenty and twenty-five years of age in order to identify potential tumours at an earlier stage, therefore improving prognosis (126,135).

From a therapeutic perspective, LS patients may benefit from alternative treatments to conventional chemotherapeutics used to treat sporadic CRCs. The study by Smyrk *et al.* found that MSI^{High} CRCs have a significantly higher number of tumour-infiltrating lymphocytes (TILs) than MSS tumours (136). In addition to this, the study by Llosa *et al.* identified infiltrating T-Helper cells and subsequent production of interferon-gamma (*IFN-γ*) within MSI^{High} CRC (137). To counteract the pro-immune response characteristics of the tumour micro-environment in MSI^{High} CRCs, there is also the expression of a number of immunosuppressive signalling proteins, including programmed cell death protein 1 (*PD-1*), its ligand (*PD-L1*) and cytotoxic T-lymphocyte associated protein 4 (*CTLA4*) – which act to suppress the anti-tumour immune response (137). The presence of TILs and an immunosuppressive environment in MSI^{High} CRCs presents a new avenue for therapeutic intervention using immune-checkpoint blockade immunotherapy. In the study by Le *et al.,* patients with MSI^{High} CRC, MSI^{High} non-CRC or MSS CRC were treated with the anti *PD-1* monoclonal antibody pembrolizumab. Both the immune response rate and progression-free survival of the MSI^{High} cohorts was significantly higher than that of the MSS cohort (138). The study by Overman *et al.* investigated the effect of the anti *PD-1* monoclonal antibody nivolumab in combination with the anti-*CTLA4* monoclonal antibody ipilimumab in MSI^{High} metastatic CRCs and demonstrated a twelve-month overall survival of 85% (139). It is perhaps unsurprising given the apparent efficacy of immune-checkpoint blockade immunotherapeutics in MSI^{High} CRCs that pembrolizumab, nivolumab and ipilimumab have all been approved by the US food and drugs administration (FDA) for use in MSI^{High} metastatic CRC.

1.3.3 – Peutz-Jeghers Syndrome

Another, somewhat rarer, autosomal-dominant condition which predisposes to colorectal polyps is Peutz-Jeghers Syndrome (PJS) – which is driven by germline mutations in the *STK11* gene (see section 1.1) (140). An early case of PJS was described in the 1920s by Peutz, who identified rectal polyps and unusual pigmentation of the skin and mucous membranes within several members of the same family (141). Despite these observations, it was not until the 1950s that PJS was coined by Bruwer *et al.,* who combined the observations of Peutz with similar observations made in another patient cohort by Jegher *et al.* (142,143). The incidence of PJS is less common than that of FAP or LS, affecting 1 in 50,000 to 200,000 live births (144).

The unusual dark brown or blue-brown pigmentations associated with approximately 95% of PJS cases often appear early in life, commonly in infancy (140,144). These pigmentations are commonly identified at the vermilion border of the lips, buccal mucosa, hands and feet of affected individuals in 94%, 66%, 74% and 62% of cases respectively (144). Alongside this pigmentation, harmatomatous polyps of the gastrointestinal tract are the other major feature of PJS – which are present in up to 88% of affected individuals but are far fewer in number to the number of polyps observed in FAP (145). Polyps are commonly located in the small intestine and colon in 60-90% and 50-64% of affected individuals respectively, however polyps can also present in the bladder and gallbladder (140). Characteristic features of PJS-associated polyps include their frond-like structure, composed of epithelial cells of the gastrointestinal region of origin and smooth muscle (144). These polyps often grow early in an affected individual's life and clinically present themselves between the ages of ten and thirty. Consequently the average age of diagnosis of PJS in both males and females is in the early to mid-twenties (144). Diagnosis of PJS can be made if the affected individual meets at least one of the criteria outlined in Table 1.3 (140).

Diagnosing Peutz-Jeghers Syndrome:	
Criterion:	Description:
Polyps	Presence of two or more polyps with PJS characteristics
Family History	Any number of PJS-associated polyps in a patient with at least one close relative whom has been diagnosed with PJS
Pigmentation	Characteristic PJS pigmentation in an individual with at least one close relative whom has been diagnosed with PJS
Polyps + Pigmentation	Any number of PJS-associated polyps in an individual who also presents with characteristic PJS pigmentation

Table 1.3 – Diagnosis Criteria for Peutz-Jeghers Syndrome: The criteria used to diagnose Peutz-Jeghers syndrome (PJS) in an individual. These include the presence of characteristic PJS-associated polyps, family history of the disease, characteristic pigmentation of the skin and/or mucous membranes or a combination of polyps and pigmentation. An affected individual need only meet one of the above criteria to be diagnosed with PJS (140).

PJS is the result of germline pathogenic mutations in the *STK11* gene located on chromosome 19, which encodes the serine-threonine kinase *Lkb1* (146). *Lkb1* is involved in a number of intracellular processes but is most noted for its role in cellular responses to stress – including low-nutrient or low-energy states (146). When a cell enters one of these states, adenosine

monophosphate (AMP) binds to AMP-activated protein kinase (*AMPK*), which is then phosphorylated and activated by *Lkb1* (147). This drives a move towards intracellular catabolism to generate ATP to maintain intracellular ATP:AMP ratios – however in the absence of *Lkb1,* cells move away from oxidative phosphorylation to aerobic glycolysis-based metabolism (146,148). Changes to cellular metabolism have been described as a hallmark of cancer development in the work of Hanahan & Weinberg (9). Another role of *Lkb1* involves the regulation of the mammalian target of rapamycin (*mTOR*) signalling pathway, a regulator of cellular growth and proliferation (149). *mTOR* is known to promote cellular growth and proliferation via the phosphorylation of downstream targets *p20* ribosomal S6 kinase (*S6K*) and eukaryotic translation initiation factor 4E binding protein 1 (*eIF4E-BP1*), which promote the initiation of translation within cells (149). This pro-growth signalling can be inhibited by the activity of *Lkb1.* Following the *Lkb1*-mediated phosphorylation and activation of *AMPK,* the tuberous sclerosis complex 2 (*TSC2*) is activated following *AMPK*-mediated phosphorylation (150). Activated *TSC2* is then able to inhibit *mTOR* signalling, therefore preventing cellular growth in low-energy or low-nutrient states (150). Therefore, a consequence of *Lkb1* inactivation in PJS allows uncontrolled and unregulated pro-growth and pro-proliferative intracellular signalling via *mTOR* (149).

Alongside the enhanced risk of developing CRC, PJS patients are also predisposed to other forms of cancer – including cancers of the lung, breast and pancreas (151–153). It is estimated that PJS patients have a greater than 90% chance of developing cancer at some point in their lives, with breast cancer carrying the highest cumulative risk of development between the ages of fifteen and sixty-four at 54%, followed by CRC in second at 39% (144,149,154). The study of Giardiello *et al.* identified nine cases of CRC in PJS patients with an average diagnosis age of approximately forty-six, which is significantly younger than the average age of diagnosis of sporadic CRCs – a trend also seen in FAP and LS patients (154). Similarly to FAP and LS, the recommended management of PJS often involves enhanced colonic surveillance – with several health authorities recommending an investigative colonoscopy every two-three years beginning at the age of eighteen in affected individuals (144). In addition to this, it is recommended that colonic polyps greater than 1cm in size identified during surveillance are removed via a polypectomy – while rapidly-growing polyps or those larger than 1.5cm in size are recommended for surgical removal (144).

1.3.4 – Other Autosomal-Dominant Colorectal Cancer Predisposition Syndromes

Another example of an autosomal-dominant syndrome associated with gastrointestinal polyp development is polymerase proofreading-associated polyposis (PPAP), which can predispose affected individuals to CRC (155). PPAP was first reported in 2013 by Palles *et al.,* who described patients with multiple colorectal adenomas and/or CRCs without an MSI phenotype (156). This phenotype was characterised by germline pathogenic mutations in the exonuclease domain of either DNA polymerase delta (*POL-δ*) or DNA polymerase epsilon (*POL-ε*), which was suggested to impair the intrinsic error-correcting capabilities of each protein (156). Germline exonuclease domain mutations (EDMs) in either enzyme show high penetrance and enhanced predisposition to CRC, with germline EDMs in *POL-δ* also predisposing to endometrial cancer (156,157). In addition to this, a number of somatic *POL-ε*

EDMs have been identified in CRC and endometrial cancer, recapitulating the phenotype of PPAP of hyper-mutated tumours on an MSS background (157). Palles *et al.* identified *POL-δ^{S478N}* and *POL-ε^{L424V}* as the changes associated with PPAP pathogenesis, with *POL-δ^{P327L}* also identified by subsequent analysis (156,157). Data from The Cancer Genome Atlas (TCGA) suggests that approximately 3% of CRCs harboured somatic EDMs in *POL-ε,* with *POL-ε^{P286R}, POL-ε^{V411L}* and *POL-ε^{S459F}* being common alterations (69,157). Other studies have also suggested that *POL-ε^{P286R}* is one of the most common *POL-ε* EDMs (158). Interestingly, somatic *POL-δ* EDMs are much rarer, only found in an estimated 0.5% of CRC samples (157). The mechanism of DNA polymerase exonuclease domain-mediated error correction will be described in Chapter V of this thesis, but the failure to repair DNA mismatches arising from DNA replication errors results in hyper-mutated tumours with thousands of mutations in exonic DNA alone (69). The study by Palles *et al.* investigated 132 patients with pathogenic EDMs in either *POL-δ* or *POL-ε* and found that 85% of *POL-δ* and 95% of *POL-ε* mutants had colorectal adenomas or CRC, with a median age of cancer diagnosis being in the forties for both genes (159). In addition to this, the median number of polyps present within *POL-δ* and *POL-ε*-mutants was thirteen and twelve respectively (159). The study, similarly to other CRC predisposition syndromes discussed, recommends enhanced colonic surveillance of affected individuals starting at the age of fourteen (159).

Alongside PJS, there are other CRC predisposition syndromes associated with the development of several harmatomatous polyps, including Juvenile Polyposis Syndrome (JPS). JPS was first reported in 1964 by McColl *et al.* and affects approximately 1 in 100,000 people (160,161). As opposed to the age of polyp presentation, the term juvenile is used as a histological descriptor of the polyps associated with JPS, which commonly present in the colon and small intestine of affected individuals (162). JPS is the result of inherited pathogenic mutations in either *BMPR1A* or *SMAD4* which, as discussed in section 1.2.2, are important regulators of cell cycle arrest (162). Polyps often appear in the first decade of life and are thought to increase the risk of CRC development by approximately 20% (163). Suggestions for management of JPS includes routine colonic surveillance every three years from the age of fifteen with surgical intervention or polypectomy where appropriate (162).

Hereditary mixed polyposis syndrome (HMPS) was first described in an Ashkenazi Jewish family who presented with multiple gastrointestinal polyps of multiple distinct histological sub-types or individual polyps that display characteristics of more than one histological sub-type (164). The genetic alteration that underpins HMPS are duplications of regulatory regions of the *GREM1* gene on chromosome 15, resulting in over-expression of the *GREM1* protein (164). The study by Lieberman *et al.* reported four families that harboured the aforementioned *GREM1* regulatory region duplication, encompassing a forty-kilobase (kb) region of chromosome 15 (164). Across the four families, a total of sixteen individuals were identified with a *GREM1* region duplication, with three families presenting with rapidly-growing intestinal polyps and the fourth fulfilling the Amsterdam Criteria for LS diagnosis – potentially complicating the distinction between the two conditions (164). Owing to the rapid growth and early presentation of intestinal polyps in HMPS, which can often be in the early twenties of affected individuals, Lieberman *et al.* consequently recommended the initiation of colonic screening in early adolescence for effective management of the syndrome (164,165).

Taken together, there are a great number of autosomal-dominant CRC predisposition syndromes which markedly increase an affected individual's lifetime risk of CRC development. There are also examples of autosomal-recessive syndromes that are known to predispose to CRC development, including *MUTYH*-associated polyposis, which often presents clinically as similar to attenuated forms of FAP (166). However, while there has been extensive research into the molecular mechanisms underpinning each of the CRC predisposition syndromes listed above, they only account for approximately 5% of all CRC cases (162). As previously stated, it is estimated that some 20-25% of CRCs have an inherited component underlying tumorigenesis whose mechanisms are as yet not completely understood (116,162). In light of this, it remains of critical importance that new genes whose pathogenic mutation predisposes to CRC development are identified and characterised via downstream functional validation experiments.

1.4 – DNA Methylation & Roles in Colorectal Tumorigenesis

1.4.1 – A History of Epigenetics

It has long been hypothesised that changes to the environment could drive heritable changes in the phenotype of an organism (167,168). As early as the 1890s, Weismann amputated the tails of successive generations of mice in an attempt to determine if this change could be inherited (168). However, it wasn't until the 1940s when Waddington was able to prove this interplay from his work with *Drosophila* (168,169). Waddington observed the development of unique thoracic and wing structures in *Drosophila* living in different temperatures, observing that after enough generations these unique structures could be stably inherited despite the absence of the initial environmental stimulus (168).

Waddington hypothesised that these interactions between the environment and the phenotype was the cause for "canalisation", the process by which cells with an identical genotype could vary in their phenotype (167). This led to the suggestion of the existence of a programme of "epigenetics" (literally processes "over genetics") that allowed canalisation to occur (167,169). Over the subsequent years, following the discovery of processes that regulate the expression of genes within cells, the definition of epigenetics shifted (170). The idea that epigenetics was the major mechanism used by cells to control their unique gene expression signature compared to other cell types, despite having the same genetic material, led to epigenetics being defined as the change in the expression of a gene in the absence of a change to the underlying DNA sequence (170). The epigenetic profile of cells remains a heritable feature that can be transmitted to subsequent generations, as well as a profile that can be modified by the environment (171).

Despite not being associated with a change to the DNA base sequence, epigenetics allows for the control of how DNA is organised within cells, which can either encourage or discourage the transcription of a gene (172). Modification of histones, the proteins around which DNA is wound, can result in such conformational changes (172). Alternatively, modification of the DNA bases themselves in a process known as DNA methylation can also influence the

transcription and downstream expression of genes, representing a critical epigenetic mechanism by which gene expression is regulated (172).

1.4.2 – The Importance of DNA Methylation in Mammalian Genomes

Effective intracellular control of gene expression is essential for proper function. The de-regulation of gene expression, either via protein over-expression or absence due to pathogenic mutation, has consequences in tumorigenesis – with examples of both being provided in section 1.3. In light of this, tight regulation of the transcription and subsequent translation of protein-coding genes is the major mechanism employed by cells to control gene expression (173). DNA methylation is the best-described mechanism for the epigenetic regulation of gene expression, which is characterised by the addition of a methyl-group to the fifth carbon of cytosine to produce 5-methylcytosine (5-mC) – most commonly in the genetic context CpG (174,175).

The methylation of DNA is regulated by the family of DNA methyltransferase (DNMT) proteins, which are responsible for both the deposition of methyl-groups onto target cytosine residues and the maintenance of 5-mC patterns throughout the genome following DNA replication (174,175). The deposition of a methyl-group onto a target cytosine is referred to as *de novo* methylation and is primarily accomplished by *DNMT3A* and *DNMT3B*, in addition to the accessory protein *DNMT3L* (174–176). Due to their roles in *de novo* methylation, these DNMTs are most highly expressed in undifferentiated cells (175,176). In addition to these proteins, *DNMT1* is responsible for the faithful propagation of DNA methylation patterns to daughter cells following replication, allowing the efficient transmission of gene expression profiles between generations of cells (174).

Alongside its role in the regulation of gene expression, DNA methylation also plays an important role in other biological processes – including embryogenesis, genomic imprinting and X-chromosome inactivation (177). The importance of DNA methylation during embryogenesis is highlighted by the embryonic lethality of *DNMT1*-deficient and *DNMT3B*-deficient mouse models (174,176,178). Mouse models of *DNMT3A* deficiency are initially viable but suffer from impaired post-natal development and die before the age of one month (176). During embryonic development, there are two distinct waves of DNA de-methylation followed by DNA re-methylation. The first of these waves occurs rapidly post-fertilisation, with re-methylation beginning at the implantation stage of development – while the second wave is initiated at the primordial germ cell stage (174).

DNA methylation of gene promoters acts as the foremost method of transcriptional repression (179). In the human genome approximately 80% of CpG dinucleotides are methylated, with unmethylated sites tending to cluster together in regions known as CpG islands (179). These CpG islands are common at gene promoters, allowing tight control of gene expression by the co-ordinated methylation of the promoters of genes to be silenced (179). DNA methylation can repress gene expression via both direct and indirect mechanisms. Following the deposition of 5-mC at gene promoters, transcription-initiating proteins are blocked from accessing the DNA and therefore are not able to bind and drive gene transcription (174,179).

In addition to this direct mechanism, the methylation of DNA can recruit methyl-CpG binding proteins to the site of DNA methylation, which act to facilitate interactions between the methylated DNA and modified histones – resulting a shift in local chromatin structure towards a more repressive conformation, further limiting the access of transcription-associated proteins to the DNA (174,179). The importance of DNA methylation in regulating gene expression is best exemplified in the study of ESCs (180). The *OCT4* gene is associated with the maintenance of pluripotency in stem cells (181). Therefore, following differentiation, the promoter of this gene undergoes *de novo* methylation, resulting in its transcriptional silencing in differentiated cells (180,182). In addition to *OCT4,* there are a number of other pluripotency-associated genes, including *NANOG* that undergo *de novo* DNA methylation and transcriptional silencing during cell differentiation (183). The study by Hackett *et al.* performed an epigenetic disruption and recovery analysis of somatic cells in order to identify genes whose expression is directly controlled by DNA methylation (184). DNA hypo-methylation via 5-aza-2-deoxycytidine (5-aza, see section 1.4.3) treatment revealed a significant up-regulation of a number of genes involved in distinct biological pathways, including immune response and gamete genesis (184). Interestingly, of these up-regulated genes, many were rapidly down-regulated again following withdrawal of 5-aza, indicating that DNA methylation may not be directly responsible for the control of the expression of these genes due to the lack of extensive *de novo* methylation in somatic cells (184). However, genes with distinct roles in the germline, including *Asz1* and *Ant4,* remained up-regulated following withdrawal of 5-aza, indicating that DNA methylation may be directly responsible for regulating the expression of these genes (184). The direct and indirect mechanisms of DNA methylation mediated regulation of gene expression described above are summarised in Figure 1.6. While the direct methylation of gene promoters appears to play a role in cell differentiation and embryonic development, the role of the indirect role of DNA methylation in transcriptional silencing and development is less clear (183,185). While a number of studies have demonstrated that recruitment of histone de-acetylases to gene promoters is mediated by methyl-CpG binding proteins *in vitro* (186,187), there are few instances of this form of regulation being demonstrated *in vivo.* While the study by Schmolka *et al.* suggested that methyl-CpG binding domain 3 (*MBD3*) was crucial for the differentiation of neuronal cells, this was independent of its methyl-CpG binding domain (188). Similarly, Caballero *et al.* demonstrated that *in vivo* several methyl-CpG binding proteins were not essential for neuronal differentiation (185). Overall, while this indirect mechanism by which DNA methylation may regulate gene expression appears to be present *in vitro,* the utility in differentiation and development *in vivo* is yet to be fully understood.

Conversely, while DNA methylation at promoters is associated with transcriptional repression, there is evidence to suggest that DNA methylation within the coding regions of genes may promote gene activation (174). Hellman & Chess demonstrated that there are differences in the DNA methylation profiles between the active and inactive X-chromosome following inactivation – with the active X-chromosome surprisingly showing a higher degree of DNA methylation than the inactive X-chromosome (189). The inactive X-chromosome was demonstrated to have DNA methylation localised to gene promoters, whereas the active X-chromosome showed intra-genic DNA methylation – the level of which was correlated with gene expression (189). Subsequently, this phenomenon has been identified across the genome, implicating DNA methylation in more processes than just transcriptional repression

(190). To further support the role of DNA methylation in controlling gene expression, there is a clear correlation between DNA methylation and DNA replication timing (190). DNA which replicates in early S-phase of the cell cycle is known to be more euchromatic, gene-rich and actively transcribed when compared to DNA replicated in late S-phase – which has less genes and is associated with regions of heterochromatin (191,192). Consequently, these early-replicating regions of the genome also show higher degrees of DNA methylation than late-replicating regions (190). It is because of these crucial roles in epigenetic regulation of gene expression that de-regulation of cellular DNA methylation patterns have been implicated in the development of several types of cancer – including CRC.

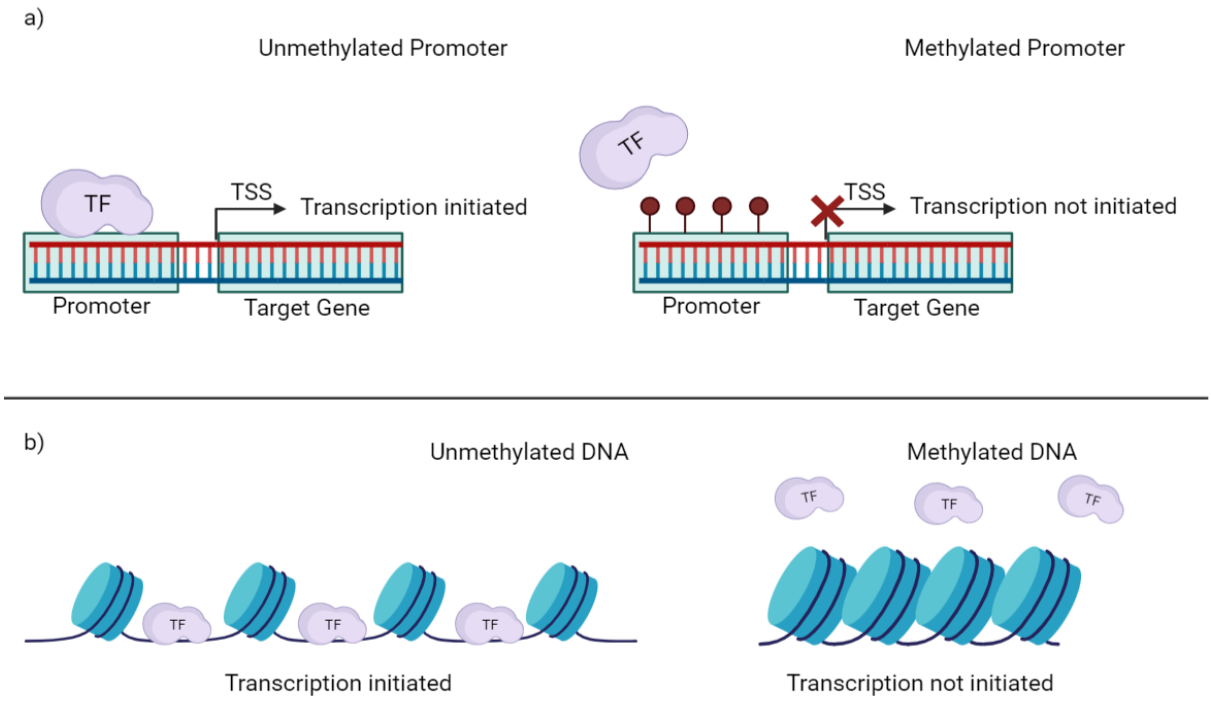


Figure 1.6 – Mechanisms of DNA Methylation-Mediated Control of Gene Expression: A schematic illustration of the direct and indirect mechanisms by which DNA methylation influences gene expression. a) Methylation can directly control gene expression by preventing a transcription factors (TF) from accessing gene promoters and driving gene transcription from the transcription start site (TSS). b) DNA methylation can also indirectly control gene expression by inducing changes in chromatin conformation to a more closed and repressive structure, again preventing TFs from accessing gene promoters and subsequently initiating transcription. Created with BioRender.com (https://app.biorender.com/).

1.4.3 – DNA Methylation Alterations in Colorectal Tumorigenesis

The adenoma-carcinoma sequence represents a general model for the development of CRC, driven by the sequential acquisition of pathogenic mutations in specific CRC driver genes – for a full description of the adenoma-carcinoma sequence see section 1.2. In addition to these mutations, more emphasis has recently been placed on over-arching changes that take place within the intestinal compartment during tumorigenesis, for example alterations to inter-cellular adhesion or angiogenesis (43). Alterations to cellular DNA methylation patterns are

thought to be a key characteristic of CRC tumorigenesis, occurring early in the adenoma-carcinoma sequence (see Figure 1.1) (43). Consequently, these alterations in DNA methylation are hypothesised to play a part in driving the progression of previously normal (or hyper-proliferative) intestinal epithelia into colorectal adenomas (43).

Alterations in DNA methylation are a common feature of several tumour types, usually characterised by genome-wide hypo-methylation interspersed with localised regions of DNA hyper-methylation at CpG islands of gene promoters (193). A summary of these methylation changes in cancer is provided in Figure 1.7. It has been suggested that global DNA hypo-methylation may drive increased genomic instability, another key feature of CRC tumorigenesis (194). In addition to this, alterations to cellular DNA methylation profiles may also have an impact on the expression of several oncogenes or tumour suppressor genes owing to the role of DNA methylation in regulating gene expression (see section 1.4.2) (195). Promoter hyper-methylation has been demonstrated to be a mechanism of tumour suppressor gene silencing in several forms of cancer, examples including the *BRCA1* gene in breast cancer or Glutathione S-Transferase P (*GSTP1*) in prostate cancer (see Figure 1.6 for an illustration of DNA methylation induced control of gene expression) (193,196). In line with this phenomenon, promoter hyper-methylation of several CRC driver genes have been reported. Examples of these driver genes include *APC* and *MLH1* (89,197). The study by Liang *et al.* reported that the hyper-methylation of the *APC* promoter was significantly correlated with the incidence of CRC and the frequency of *APC* promoter hyper-methylation was more common in colorectal adenomas and stage I CRC than normal colorectal tissue – with the frequency of *APC* promoter hyper-methylation being nearly ten-fold higher in CRC than normal tissues (197).

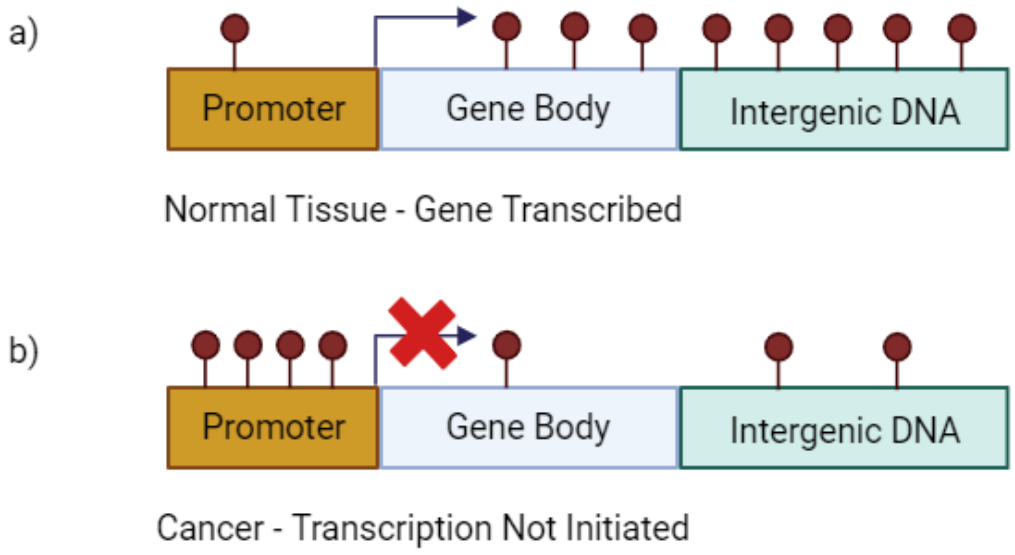


Figure 1.7 – The Methylation Landscape of Cancer: An illustration of the DNA methylation changes that occur in cancer. In normal tissues (a), CpG islands located within gene promoters are generally unmethylated, while DNA methylation (red lollipops) is common within protein-coding and intergenic DNA. However, in cancer (b), there is global DNA hypo-methylation (especially in protein-coding and intergenic DNA), coupled with DNA hyper-methylation at CpG islands within gene promoters, which can prevent transcription factor binding and subsequent transcription of a gene. Created with BioRender.com (https://app.biorender.com/).

Perhaps a more well-known example of aberrant promoter hyper-methylation in CRC is at the *MLH1* gene locus. The hyper-methylation of the *MLH1* promoter has been shown to reduce the expression of the *MLH1* protein, thereby driving the development of MSI^{High} CRC (89). Following a meta-analysis by Li *et al., MLH1* promoter hyper-methylation was identified in 18.7% of sporadic CRCs across twenty-nine separate studies and 16.4% of LS-associated CRCs across eight independent studies (89). When this analysis was refined to only investigate MSI^{High} CRCs, a total of 73.6% of sporadic CRCs and 15.3% of LS-associated CRC presented with *MLH1* promoter hyper-methylation (89). In addition to CRC, *MLH1* promoter hyper-methylation has also been identified in approximately 40% of MMR-deficient endometrial cancers (198). In the context of CRC, Li *et al.* also reported significant associations between *MLH1* promoter hyper-methylation and tumour location, patient sex and the presence of pathogenic *BRAF* mutations (89). There are also examples of promoter hyper-methylation of genes that can drive CRC tumorigenesis via a more indirect mechanism than previous examples. The study by Neri *et al.* demonstrated that the expression of ten-eleven translocation 1 (*TET1*) is down-regulated early in CRC development following analysis of 887 colorectal adenocarcinomas (199). The study also provided evidence that *in vitro* repression of *TET1* expression drives an increase in the proliferation rate of CRC cell lines (199). The *TET1* protein, in addition to the other members of the family (*TET2* and *TET3*) is involved in the active de-methylation of DNA via catalysing the conversion of 5-mC to 5-hydroxymethylcytosine (5-hmC) (200). Therefore, mutations in *TET1* resulted in a lack of promoter de-methylation and subsequent reduced expression of *DKK3* and *DKK4,* which are both inhibitors of the *Wnt* signalling pathway (199). The potential role of members of the TET family in CRC tumorigenesis will be explored in detail in Chapter II and Chapter III of this thesis.

In light of the aberrant DNA methylation patterns driving inappropriate expression of key CRC driver genes, there has been some investigation into the utility of DNA methylation profiles as a biomarker of CRC tumorigenesis (201). The study by Rasmussen *et al.* evaluated seventy-four previous investigations into the utility of DNA methylation analysis of blood or stool samples in the early detection or prognosis of CRC (201). Multiple studies investigated *APC* promoter methylation as a CRC biomarker, including Pack *et al.,* who used methylation-specific polymerase chain reaction (MS-PCR) and identified hyper-methylation of the *APC* promoter in the blood with a sensitivity and specificity of 57% and 86% respectively (202). Conversely, Leung *et al.* did not find a significant difference in the DNA methylation of the *APC* promoter in the blood of CRC cases vs controls, but interestingly, did find a significant association for the *MLH1* gene (203). Furthermore, altered DNA methylation profiles in CRC patients may, in addition to potentially providing a novel biomarker for early detection, be targeted therapeutically. One such example of this was the study by Yang *et al.,* who treated the CRC cell line HCT116 with low doses of the DNA hypo-methylating agent 5-aza for twenty-four hours (204). The results of the study included impaired colony-formation of the 5-aza treated cells, as well as an initial reduction in cellular growth rates – with the population doubling time of the cells decreasing from twenty-six hours to thirty-eight hours before returning to pre-treatment levels after several weeks (204). Global DNA de-methylation was identified in the 5-aza treated cells five days post-treatment, potentially suggesting a link between DNA methylation profiles and cellular proliferation (204). The therapeutic potential of 5-aza may also extend to the re-activation of genes previously

silenced by aberrant promoter hyper-methylation. The study by Mossman *et al.* treated five CRC cell lines for seventy-two hours with 5-aza and demonstrated global de-methylation and re-activation of a number of previously silenced driver genes – including *MLH1* in HCT116 cells, which showed no DNA methylation at promoter CpG islands (205). This effect could not be recapitulated in the cell line SW48, which showed evidence of *MLH1* expression whilst retaining promoter hyper-methylation (205). However, despite the therapeutic promise of 5-aza, its toxicity has limited its clinical utility (206,207). As an analogue of deoxycytidine, 5-aza is incorporated into DNA and results in DNMT inhibition (206,208,209). The incorporation of 5-aza into the DNA sequence covalently traps *DNMT1,* the maintenance DNA methyltransferase, to DNA – thus preventing propagation of DNA methylation profiles in cells following DNA replication (206). This block in DNA methylation and covalent linkage of *DNMT1* to the DNA explains the toxicity of 5-aza treatment, given the critical role of *DNMT1* in non-cancerous cells (207).

1.4.4 – The CpG Island Methylator Phenotype

Localised promoter hyper-methylation in tumours forms a key characteristic of the $CIMP^{+}$ tumours (100). CIMP was first described in the late 1990s, where abnormal DNA hyper-methylation was identified in primary CRC (100). Toyota *et al.* observed seven regions that were hyper-methylated specifically in cancers, which were coined as methylated in tumours (MINTs) (100). Cancers with hyper-methylation at three or more of these seven MINTs often presented with promoter hyper-methylation of tumour suppressor genes (e.g. p16) and *MLH1,* suggesting that their epigenetic silencing could drive tumorigenesis (100,210). CIMP is characterised by distinct patterns of promoter hyper-methylation and is often associated with an MSI^{High} phenotype (99,211). CIMP is present in distinct subsets of CRCs, but has also been identified in other cancer types, including pancreatic and gastric cancers (99). The study by Samowitz *et al.* investigated a cohort of 864 CRCs using MS-PCR at the loci of *MLH1, CDKN2A, MINT1*, *MINT2* and *MINT31* (212). The study concluded that a CRC could be classified as $CIMP^{High}$ if there was evidence of promoter hyper-methylation in two or more of these panel genes and a cancer could be defined as $CIMP^{Low}$ if promoter hyper-methylation was detected in less than two of the panel genes (212). This characterisation was also used in the comprehensive characterisation of CRC in a TCGA study, who identified distinct $CIMP^{High}$ and $CIMP^{Low}$ CRC clusters (69). The study by Shen *et al.* analysed both the genetic and epigenetic traits of ninety-seven primary CRCs by characterising mutations in previously identified CRC driver genes and the DNA methylation status of twenty-seven promoter-associated CpG islands (213). The results of this study identified three separate CIMP clusters in CRC, two for $CIMP^{+}$ cancers and one for $CIMP^{-}$. Of the $CIMP^{+}$ clusters, the first (termed "CIMP1" in the study) was characterised by MSI (80% of cancers) and *BRAF* mutation (53% of cancers), whereas the other cluster ("CIMP2") was associated with mutations in *KRAS* (92% of cancers) (213). Interestingly, the $CIMP^{-}$ cluster presented with a greater prevalence of *TP53* mutations than either of the $CIMP^{+}$ clusters (213). The use of MS-PCR in the early studies of CIMP on a limited number of panel genes suffered from a number of limitations (214). There was little consistency in the gene panels used to define CIMP in many early studies, while MS-PCR technology was limited by PCR bias, false-

positive results and only allowing a qualitative, not quantitative, assessment of methylation (101,214).

The clinical utility of CIMP has been of great scientific interest in recent years – either from a prognostic perspective or in predicting response to anti-cancer therapies. The study by Juo *et al.* performed a meta-analysis of thirty-three previous CIMP studies in CRC (215). Of these studies, eleven were used to assess the association between CIMP and disease-free survival (DFS). Of these eleven studies, eight showed no association between CIMP and DFS, two identified $CIMP^{+}$ CRCs having a worse DFS than $CIMP^{-}$ CRCs and one found CIMP to be associated with a longer DFS (215). In addition to this, one study, despite having no overall significant association between CIMP and DFS, found evidence of a worse DFS of $CIMP^{+}$ tumours when compared to $CIMP^{-}$ tumours of the proximal colon (215). Juo *et al.* also used these previous studies to assess the utility of CIMP in predicting patient response to 5-fluorouracil (5-FU) chemotherapy. In total, seven studies assessed the association between CIMP and DFS in CRCs that received adjuvant 5-FU treatment following surgical resection. Of these, three studies concluded that $CIMP^{+}$ CRCs have a better response to 5-FU treatment than $CIMP^{-}$ (215). However, the overall utility of the meta-analysis is limited by the differences between studies in CIMP classification. Another meta-analysis by Wang *et al.* investigated twenty-six studies, comprising 2,142 $CIMP^{High}$ CRCs from a total of 12,930 CRCs (216). Of these studies, twenty-one assessed $CIMP^{High}$ cancers against $CIMP^{Low}$ and $CIMP^{-}$ cancers whereas six studies analysed $CIMP^{High}$, $CIMP^{Low}$ and $CIMP^{-}$ cancers separately. In both of these groups $CIMP^{High}$ CRCs were significantly associated with a poorer overall survival than $CIMP^{-}$ CRCs (216). Overall survival was also significantly poorer in $CIMP^{High}$ vs $CIMP^{-}$ CRCs of stage III-IV, but not in stage I-II CRCs (216). Interestingly, Wang *et al.* identified no significant differences in overall survival of $CIMP^{High}$ vs $CIMP^{-}$ MSI^{+} cancers, but these differences were present in CRCs with proficient DNA MMR (183). In contrast to the meta analysis by Juo *et al.,* post-surgical 5-FU treatment had no significant effect on DFS in either $CIMP^{High}$ or $CIMP^{Low}$/$CIMP^{-}$ CRCs (216). Overall, the utility of CIMP in patient prognosis in CRC is limited due to the varying results of previous studies – including meta-analyses, inconsistencies in the gene panels and sequencing technologies used to classify CIMP and difficulties controlling for tumour stage or prior chemotherapies (215,217). For example, the study by Advani *et al.* performed a meta-analysis into the prevalence of CIMP across global populations, where there were approximately sixty different methods for determining CIMP status across these studies, further highlighting the difficulties of clearly defining a $CIMP^{+}$cancer (218).

$CIMP^{High}$ CRCs appear to present with distinct clinical features, including localisation to the proximal colon and an older age of onset than $CIMP^{-}$ CRCs (217). However, despite the identification of these features, the driving force that underpins the development of $CIMP^{+}$ disease remains unclear. There has been suggestion that mutations in DNMT or TET genes may drive CIMP via aberrant hyper-methylation or loss of DNA de-methylation respectively (217). However, mutations in either of these families of proteins occur in less than 10% of CRCs and may not provide a full explanation of the mechanisms which underpin CIMP pathogenesis (217). As previously discussed, mutations in *BRAF* often co-occur with CIMP, leading to speculation that the two may be linked (212,217). Fang *et al.,* hypothesised that mutations in *BRAF* may drive aberrant promoter hyper-methylation via the increased

phosphorylation and expression of the transcriptional repressor MAF BZIP transcription factor G (*MAFG*) – which was thought to be a consequence of increased MAP kinase pathway activity in *BRAF* mutant CRCs (219). Fang *et al.* found *MAFG* was able to recruit a series of transcriptional co-repressors and influence the hyper-methylation of the *MLH1* promoter in CRC cell lines and the knockdown of *DNMT3B* was able to alleviate this repression (219). The suggestion that activation of the MAP kinase pathway may driver DNA hyper-methylation led to the hypothesis that *KRAS* may also represent a CIMP driver gene (220). The study by Serra *et al.* suggested that activating mutations in *KRAS* drove the up-regulation of the *ZNF304* gene, which subsequently recruits a transcriptional co-repressor complex including *DNMT1* to drive the hyper-methylation of tumour suppressor gene promoters (220). However, despite these studies, there is currently no widely-accepted model that explains the development of CIMP in CRC.

1.5 – Aims and Objectives of the Thesis

1.5.1 – Rationale & Project Scope

Throughout this chapter, it has been made evident that the development of CRC is a complex process of both genetic and epigenetic alterations that drive tumour development and that genetic predisposition plays a key role in tumorigenesis in a significant number of CRCs. The identification of several autosomal-dominant CRC predisposition syndromes, examples of which are described in section 1.3, represent key avenues for early disease detection and improved patient prognosis via the screening of at risk individuals. However, the current list of predisposition syndromes only explains a portion of CRCs with an inherited genetic component – while the underlying cause of a significant proportion of these CRCs remains unclear. It is therefore of critical importance that additional genes associated with CRC risk are identified and characterised to further improve genetic testing and surveillance schemes. Consequently, this thesis will address this by assessing the 4q24 locus of chromosome 4, which was a region of interest recently identified by a Genome-Wide Association Study (GWAS) meta-analysis as potentially being associated with CRC predisposition. In addition to this, this thesis will investigate the role of methyl-CpG binding domain 4 (*MBD4*) in CRC predisposition, the germline inactivation of which has recently been associated with the development of colorectal polyps (221).

Also discussed in this chapter is the key role DNA methylation plays in CRC tumorigenesis. Alterations in the DNA methylation profile of cells have been identified as a key component of the adenoma-carcinoma sequence, but also more generally have been identified in many other cancers. In the context of CRC, promoter hyper-methylation of key tumour suppressor genes has been identified as a potential mechanism of tumorigenesis, with distinct sub-classes of CRC presenting with CIMP – the underlying cause and clinical utility of which is yet to be completely understood. Therefore, this thesis will investigate the role of DNA methylation in CRC tumorigenesis and provide a comprehensive analysis of the mechanisms by which DNA methylation alterations can facilitate colorectal tumorigenesis.

1.5.2 – Objectives of the Thesis

In order to assess the role of the 4q24 locus in CRC tumorigenesis, this thesis will perform the following:

1) Comprehensive *in silico* analysis of the genomic region to investigate the likely mechanisms underlying its association with CRC.
2) Functional validation of the candidate target gene of the GWAS meta-analysis region via further *in silico* analysis and subsequent mouse models of CRC.

The role of *MBD4* in driving intestinal polyp development will also be assessed via:

1) Whole-genome sequencing analysis of several intestinal polyps from a patient with a germline biallelic truncation in *MBD4.*
2) CRC driver gene analysis of these polyps.
3) Comprehensive association of mutations within these polyps with cellular and epigenetic features.

The mechanisms by which DNA methylation affects CRC mutagenesis will be assessed by:

1) Analysis of whole-genome sequencing data of cancers within the 100,000 Genomes Project (100KGP).
2) Driver gene analysis of these cancers.
3) Comprehensive association of mutations within these cancers with DNA methylation, replication timing, transcription strand and replication strand.

Chapter II – Investigating the Association Between the 4q24 Locus of Chromosome 4 & Colorectal Cancer Predisposition

2.1 – Background

2.1.1 – The Principle of Genome-Wide Association Studies

As discussed in Chapter I of this thesis, germline mutations in known cancer predisposition genes underpin the elevated tumour risk of individuals with cancer predisposition syndromes (11,12). Over 100 cancer predisposition genes have been identified across several cancers, but there are still several additional genes whose importance in disease is yet to be defined (222). In order to identify new predisposition genes, Genome-Wide Association Studies (GWAS) have come to the fore as a powerful method to detect genetic variants that are more prevalent in patients with a trait or disease when compared to unaffected controls (223). The common unit measured by GWAS are single-nucleotide polymorphisms (SNPs) – which are small, inherited DNA sequence variations that exist between individuals. While most SNPs have no consequence for the organism as a whole, some SNPs are able to drive alterations in the expression of nearby genes at the same locus or alter the expression of genes several megabases (Mb) away (224). SNPs are able to exert these effects via direct or indirect mechanisms. Direct mechanisms include where a SNP lies within the protein-coding sequence of a gene and therefore potentially represents a pathogenic change to the amino acid structure of the protein (223). SNPs that act via this mechanism are less common, as the majority of SNPs lie outside this protein-coding sequence and may alter gene expression by influencing transcription factor binding to gene promoters or the overall chromatin conformation of the locus, thereby facilitating indirect control of gene transcription (223,224).

In order to identify SNPs which are over-represented in individuals with a particular trait or disease, GWAS genotypes several thousand, or in many cases tens of thousands, of individuals with a trait or disease and a similar number of unaffected controls from the same ethnic background (e.g. European) (223). This genotyping involves the genomic sequencing of each individual using a SNP array, which contains thousands of previously identified SNPs across the genome (225). This primary genotyping data can also be combined with external reference data from other sources, for example the 1,000 Genomes Project, to improve the statistical power of the study (225). Following integration of primary data with external data, stringent quality control measures are required to improve the reliability of the data. Examples of these measures include the removal of poor quality DNA, removal of missing SNP data and the removal of ambiguous DNA base calls for a SNP (226). Following this, SNPs which are significantly associated with a trait or disease can be identified, since GWAS make use of a large sample size and to correct for multiple testing, the p-value threshold ($p_{(GWAS)}$) for a SNP to be deemed genome-wide significant is normally 5×10^{-8}, but this can vary between studies (226).

Following the identification of SNPs that meet the criteria for genome-wide significance, further analysis can be conducted to investigate the effect of a SNP on the expression of a gene. The most common method applied in order to do this is an expression quantitative trait loci (eQTL) study (227). Since the majority of SNPs are not found within the protein-coding sequence of genes and instead are possibly located within regulatory elements, there could be several factors aside from a GWAS SNP that affect the expression of a gene (227). Therefore,

an eQTL is defined as a locus that explains a fraction of the variation in the expression of a gene (227). This has led to several research consortia developing with the purpose of providing a functional annotation for SNPs not found within the coding region of genes, which allows an estimation of their effect on the expression of a particular gene (228). The utility of eQTLs in the functional investigation of GWAS loci will be discussed in section 2.1.4.

2.1.2 – Colorectal Cancer Predisposition SNPs

In recent years, several large-scale GWAS have identified genetic variants associated with a wide range of traits and diseases, including height, hair colour, schizophrenia and several types of cancer (229–233). As a result of this, GWAS have been able to identify dozens of genetic variants associated with increased cancer risk (234–236). Chapter I of this thesis illustrated the pivotal role of genetics in CRC predisposition and the identification of genes that are associated with autosomal-dominant CRC predisposition syndromes (11,237,238). Therefore, a number of GWAS have been performed in the context of CRC with the goal of identifying new genetic variants associated with CRC predisposition, with the subsequent hope of identifying new genes associated with enhanced disease risk (233,239).

These studies have continued to identify new genetic variants significantly associated with enhanced CRC disease susceptibility, with many of the early CRC GWAS identifying the 8q24 region as one of the first risk loci for CRC development (240–242). Tomlinson *et al.* reported the SNP rs6983267 as being significantly associated with CRC development following a GWAS of 930 familial CRCs and 960 controls (243). Tuupanen *et al.* further studied the 8q24 locus and identified rs6983267 as potentially affecting the *Wnt* signalling pathway, which may explain its association with CRC predisposition (232). In this study, it was shown that rs6983267 lies within the binding site of the *Wnt*-regulated transcription factor *TCF7L2* (232). Following the relative success of GWAS at the 8q24 locus, further studies and meta-analysis of previous GWAS by Houlston *et al.* led to the discovery of a total of fourteen common variants significantly associated with CRC risk across several genomic loci – including 1q41, 3q26 and 12q13 (235). Of these fourteen variants, three were located in close proximity to *BMP2, BMP4* and *GREM1* – which are secreted components of the bone morphogenetic protein (BMP) signalling pathway (233). Following further investigation of these loci by Tomlinson *et al.,* using 24,910 CRCs and 26,275 controls, one new variant significantly associated with CRC was identified at both the *BMP2* and *BMP4* loci (233). In addition to this, the previously identified CRC association at the *GREM1* locus was found to be the result of two independent significant variants in close proximity to one another (233).

In addition to these studies in populations of mostly European descent, new variants associated with CRC have also been reported in a variety of ethnic groups – including Han Chinese and African-American populations (236,244). By 2013, the list of known variants significantly associated with CRC grew to twenty-nine variants across twenty-one different genetic loci in European and Asian populations (245). This trend of discovery continued and by 2019, forty-three variants across forty genetic loci had been associated with CRC in European populations alone (234). In Asian populations, significant associations with CRC

were detected for eighteen variants across sixteen loci, many of which are shared with those identified in Europeans (234). Across all ethnicities, a total of fifty-three genomic loci had been significantly associated with CRC risk as a result of GWAS and the number of significantly associated variants in European and Asian populations stood at sixty-one (234). In the same year, Law *et al.* conducted five new GWAS and performed a meta-analysis of ten previously published studies, in a total of 34,627 CRCs and 71,379 controls of European descent, resulting in the identification of 623 variants reaching the threshold for genome-wide significance (234). Of these variants, thirty-one previously unreported loci were identified to be significantly associated with CRC (234). One example of these novel loci was the 4q24 locus of chromosome 4, a locus which has previously been identified in a breast cancer GWAS (246,247).

2.1.3 – Fine-Mapping of GWAS Data can Identify Causal Variants

While the SNPs identified by a GWAS may be correlated with a complex trait or disease, it cannot be assumed that these SNPs are the causal variants which underpin trait or disease susceptibility (228). SNPs represent inherited variations of the genome and the inheritance pattern of SNPs from each parent of an individual was often considered to be a random process. However, it has been identified that nearby SNPs are often inherited in groups, referred to as haplotype blocks, due to the random recombination of the maternal and paternal chromosomes during embryonic development (248). This non-random association of SNPs within the same haplotype block can be referred to as linkage disequilibrium (LD) and is often used to quantitatively measure the correlation of two SNPs via an "r^2" metric – which ranges from zero to one (249). An r^2 value of zero indicates that two SNPs are independent of one another and are likely not inherited together, whereas an r^2 value of one indicates that SNPs are likely to be inherited alongside one another (223).

While the principle of LD blocks has allowed the identification of several additional variants associated with a trait or disease, it also hinders efforts to identify the causal variant(s) at a locus which underpins this association (228). As a result of this, it has now been established that the SNP with the most significant association with a trait or disease at a locus, often referred to as the "lead" SNP, is not necessarily the true causal variant at the locus – but may instead be in LD with the causal variant (228). The confounding nature of LD in determining the causal variant from a list of significantly associated variants identified by GWAS can be overcome by further analysis of existing GWAS data in a process known as fine-mapping (250,251). The goal of this fine-mapping analysis is to reduce the list of variants significantly associated with a trait or disease according to GWAS, which may be an extensive number, into a shorter list of "candidate causal variants" which can then be prioritised for downstream functional investigation in a laboratory setting (228,252).

The lack of publically-available individual-level genotype data from GWAS, often due to infeasibility or data protection constraints, has driven the development of several fine-mapping methodologies which instead make use of GWAS summary statistics – which are much more readily available in the public domain. In order to carry out reliable fine-mapping of a genomic locus, three conditions regarding the data being used must be met (228). The

first of these conditions is that all SNPs being used in fine-mapping analysis must have first passed the stringent GWAS quality control measures, described in more detail in section 2.1.1 (228). Secondly, all common variants from the locus of interest must be included in the GWAS data, either via genotyping of study subjects or imputation from external data and finally, the sample size of the GWAS data must be great enough to accurately discriminate between variants in strong LD (228).

Once the above criteria have been met, there are two predominant fine-mapping methodologies that can be implemented on GWAS data. The first of these methods assesses only SNPs that meet a pre-specified p-value threshold or are in LD with the lead SNP, thereby assuming that only variants that are significantly associated with the trait or disease via GWAS, or variants in LD with the lead SNP, are potentially causal (228). While this method confers some advantages, such as generating a small number of candidate causal variants, there are a number of drawbacks. By placing a large emphasis on the p-value of variants in GWAS data, there runs the risk of data not being accurately comparable between studies or even different loci within the same study – as GWAS p-values can be affected by both the sample size of the study and properties unique to each locus (228).

An alternative fine-mapping methodology employs a more sophisticated Bayesian framework, which assigns a Bayes Factor to each variant at a locus (252,253). This Bayes Factor represents the ratio of evidence for a variant being the underlying causal variant at a locus versus evidence suggesting the causal variant lies elsewhere(252,253). From this Bayes Factor, a quantitative estimate of the likelihood of causality of a variant – referred to as a posterior probability – can be calculated. This posterior probability is subject to a number of assumptions, the most limiting of these being that there is only one causal variant at the locus(228,251–253). After calculating the posterior probabilities of all the variants at a given locus, it is possible to define a "credible set" of potentially causal variants, the size of which depending on both the posterior probabilities of variants at the locus and the threshold used to define the credible set (228,254). The standard cut-off of a credible set is 95%, which suggests that the credible set is 95% likely to contain the causal variant, but can be adjusted to meet study requirements (254). The number of SNPs within a credible set can be used as a measure of how informative Bayesian fine-mapping analysis has been for a given locus. A small credible set provides only a small number of variants – and therefore a small number of candidate causal genes – for functional investigation in the laboratory, while large credible sets are often considered uninformative (254).

As previously discussed, a major limitation of Bayesian fine-mapping strategies is the assumption that there is only one causal variant at a locus, which is unlikely to be true given the complex nature of the control of mammalian gene expression (228). Another weakness of Bayesian fine-mapping is that SNPs identified as part of the credible set at a locus may have no previously established functional importance in the tissue of interest (254). While this limitation is not a direct consequence of the Bayesian fine-mapping framework, it may be considered difficult to calculate an accurate posterior probability for a variant without functional annotation data for the locus of interest which may give an indication of which variants are important in the regulation of gene expression in the tissue of interest(228,252,254). Therefore, subsequent fine-mapping analyses have attempted to address these limitations. An example of this is the study by Kichaev *et al.,* who developed the Probabilistic Annotation Integrator (PAINTOR) pipeline, which aims to integrate GWAS

summary statistics for a locus of interest with both the LD structure of the locus and functional annotation data from the Encyclopaedia of DNA Elements (ENCODE) to improve the reliability of posterior probability calculations (251). Following simulation analysis of 1,000 Genomes Project data comparing the PAINTOR pipeline to other Bayesian fine-mapping strategies, Kichaev *et al.* reported a reduction in the number of SNPs required to create a 90% credible set from 13.3 to 10.4 SNPs per locus (251). In addition to this improvement in credible set generation, the PAINTOR pipeline also seeks to address another major limitation of previous Bayesian fine-mapping strategies by including a feature that allows there to be up to three causal variants at a given locus, as opposed to the assumed one causal variant used in previous studies (227,228,251).

2.1.4 – Integrative GWAS & eQTL Data Analysis to Identify Candidate Target Genes

While Bayesian fine-mapping strategies can be useful in generating a shortlist of candidate causal variants from a locus of hundreds or even thousands of SNPs, there are other downstream analysis techniques that can also be applied to GWAS data. An important limitation of GWAS data in isolation is the lack of information about the effect of a variant significantly associated with a trait or disease on the expression of nearby genes at the locus – a limitation also shared by Bayesian fine-mapping methodologies. This limitation can be addressed by eQTL analysis, which allows a quantitative effect size to be defined for a variant on the expression of nearby genes (227). In order to drive a significant association with a trait or disease, a variant identified by GWAS is assumed to act via either up-regulation or down-regulation of the expression of nearby genes (see Figure 2.1). Therefore, any method which is able to integrate information on the effect of a variant on gene expression (eQTL data) with data which associates a variant with a trait or disease (GWAS data) would potentially allow the identification of candidate causal genes which underpin trait/disease associations identified at a locus by GWAS (255).

Transcriptome-Wide Association Studies (TWAS) have recently emerged as a powerful tool for integrating gene expression data with known GWAS variants in order to extend the identification of genotype-trait associations by GWAS to gene-trait associations (256). Since it is unpractical to perform gene expression analysis on the same scale as SNP genotyping used in GWAS, smaller-scale gene expression datasets are often used, in addition to small external reference datasets (257). The integration of this gene expression data with genotype-trait GWAS data allows inferences to be made between the predicted effect of a variant on gene expression, which is in turn related to the outcome trait or disease (see Figure 2.1) (258). The study by Fernandez-Rozadilla *et al.* integrated CRC GWAS data from European and East-Asian populations with colonic mucosa gene expression data from 1,107 participants and identified novel associations between the expression of fifteen genes and CRC (257). However, TWAS analysis can be hindered by a number of limitations. For example, GWAS analysis, as a consequence of LD, identify several variants significantly associated with a trait or disease at a locus – which can in turn result in several gene-trait associations being identified at a locus by TWAS (256). In addition to this, the expression of

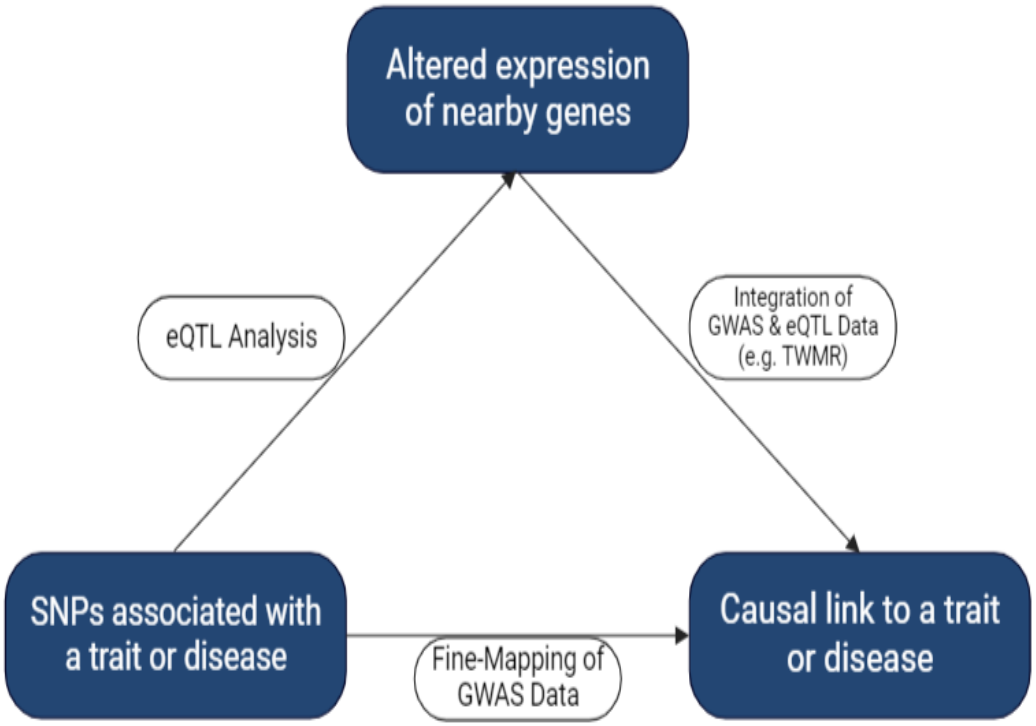


Figure 2.1 – The Mechanism Underlying SNP-Phenotype Associations: A schematic illustration of the mechanism by which a SNP significantly associated with a trait or disease may drive the association. Variants identified by Genome-Wide Association Study (GWAS) analysis are assumed to be associated with a trait or disease by altering the expression of nearby genes at the same locus. Fine-mapping strategies are able to determine the likely causal variant(s) from GWAS data and expression quantitative trait loci (eQTL) studies quantify the effect of a SNP on the expression of nearby genes. Methods have subsequently been developed to integrate both eQTL data with GWAS data to associate alterations to gene expression with a trait or disease. Created with BioRender.com (https://app.biorender.com/).

a number of genes may be correlated with one another, potentially resulting in false positive TWAS results (256).

In addition to TWAS, alternative methods exist that are used to identify novel gene-trait associations. An example of these includes Summary data-based Mendelian Randomisation (SMR), which integrates GWAS summary statistics with eQTL data to quantify the effect of altered expression of a gene of interest on a phenotype of interest (259). A similar method was subsequently developed by Porcu *et al.* known as Transcriptome-Wide Mendelian Randomisation (TWMR) (255). Both SMR and TWMR integrate eQTL and GWAS data and associate the data with a phenotype of interest via Mendelian Randomisation (MR), an epidemiological technique which uses SNPs as "instrumental variables", which act as a proxy for alterations in gene expression, termed the "exposure" and link this exposure to a trait or disease known as an "outcome" (260). However, the advantage of TWMR is the use of

multiple instrumental variables at the same time, compared to SMR which provides a univariate single-instrument methodology that struggles to disentangle causality and pleiotropy of instrumental variables that affect multiple exposures (255,260). The process of MR is subject to a number of assumptions (260). Firstly, the instrumental variables (SNPs) provided in the analysis should be strongly associated with the exposure (expression of the gene of interest), a lack of strong association can result in an over-estimation of the effect size of an exposure on the outcome of interest in what is referred to as weak instrument bias (260). Secondly, these instrumental variables should not be associated with any confounding variables and, thirdly, the instrumental variable(s) should only be associated with the outcome via the exposure(s) being investigated (260).

These assumptions are often difficult to meet, as one instrumental variable can be associated with an outcome via multiple exposures, known as horizontal pleiotropy – which violates the third assumption of MR (260). For example, rs116563317 has been shown to significantly ($p_{(eQTL)} < 0.005$) affect the expression of both TBC1 domain containing kinase (*TBCK*) and eukaryotic transcription elongation factor 1α pseudogene 9 (*EEF1AP9*) in the transverse colon – according to the Genotype-Tissue Expression (GTEx) project data. Furthermore, this third assumption of MR can also be violated by instrumental variables in strong LD with one another leading to the association of multiple exposures with the outcome (260). TWMR aims to address horizontal pleiotropy by replacing this third assumption of MR with the instrument strength independent of direct effect (INSIDE) assumption, which implies that adding additional exposures associated with an instrumental variable in a multi-instrument, multi-exposure model reduces bias owing to pleiotropy – addressing the limitations of the single-instrument, single-exposure methodology of SMR (255). In addition to this, TWMR is able to reduce the impact of LD on MR by accounting for the correlation between instrumental variables included in an analysis (255). Porcu *et al.* implemented TWMR in the context of over forty human traits and identified over 4,000 gene-trait associations – 36% of which had no genome-wide significant GWAS variant at the locus (255). This suggests that the integration of eQTL and GWAS data via TWMR represents a potentially powerful methodology to identify novel gene-trait associations in a variety of complex traits and diseases.

2.1.5 – Chapter Aims

The GWAS meta-analysis performed by Law *et al.,* described in section 2.1.2, identified thirty-one novel loci associated with CRC predisposition (234). One of these novel loci was located at the 4q24 region of chromosome 4 with the lead variant rs17035289 (see Figure 2.2), a variant which has previously been associated with type II diabetes mellitus, according to the GWAS Catalogue, although the risk allele for diabetes in this study was the non-risk allele for CRC predisposition (234,261,262).

In order to characterise the underlying mechanisms that underpin the association with CRC at this locus, several *in silico* analyses of the surrounding genomic region will be performed – these will include:

- Conditional analysis of GWAS meta-analysis data from participants of European heritage to search the region for additional variants significantly associated with CRC development that are independent of the lead variant.
- PAINTOR fine-mapping of the locus around each of the independent signal(s) to identify a credible set of candidate causal variants that underpin the association of the locus with CRC predisposition.
- Functional annotation of the candidate causal variants identified by fine-mapping to determine the likely mechanism(s) by which the candidate causal variants exert their effect on the CRC phenotype.
- TWMR-based analysis integrating eQTL data from GTEx with GWAS data to identify the likely candidate causal gene associate with enhanced CRC predisposition at this locus.
- Genotype-expression analysis using pre-existing CRC TWAS data to determine the effect of GWAS variants on the expression of this candidate causal gene.

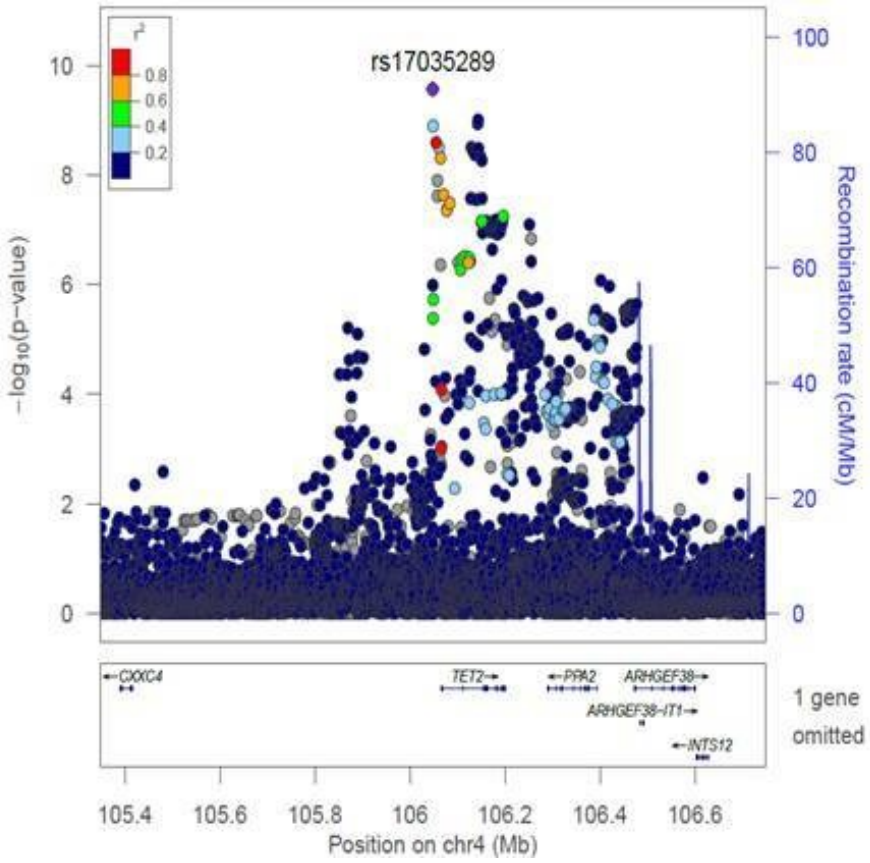


Figure 2.2 – Association of the 4q24 Locus of Chromosome 4 with Colorectal Cancer Predisposition: A representation of the 4q24 locus of chromosome 4 according to a recent Genome-Wide Association Study (GWAS) meta-analysis. Each point on the chart represents a variant and the $-\log_{10}$ p-value of each variant in the context of colorectal cancer is illustrated on the Y-axis. The linkage disequilibrium r^2 statistic of a variant with the lead variant (rs17035289 – illustrated in purple) is also described by the colour of each point. Unpublished data from Ian Tomlinson (reproduced with permission).

2.2 – Materials & Methods

2.2.1 – Conditional Analysis of the 4q24 Locus of Chromosome 4

CRC GWAS meta-analysis data was generated for a total of 185,616 participants. Of these participants, 78,473 cases and 107,143 controls were included across a total of seventeen cohorts of European heritage (257). A detailed description of each cohort is provided in Table 2.1. In addition to this, data from a total of 21,731 cases and 47,444 controls across fourteen cohorts of East-Asian descent were also used to produce a CRC meta-analysis in East-Asian populations. Meta-analysis was performed as described by Fernandez-Rozadilla *et al.,* who used Meta (v1.7) to perform meta-analysis on SNPs with an imputation quality score > 0.4 (225,257). The details of each of these cohorts is presented in Table 2.2. From this data, a region 1Mb upstream of transcription start site and 1Mb downstream of the transcription termination site of the gene closest to the lead SNP was selected for further analysis. Variants that displayed likely heterogeneity between cohorts ($I^2 > 75\%$ or $p_{(heterogeneity)} < 0.01$) were excluded from downstream analysis, leaving a total of 8,166 variants at the locus in European populations and 6,308 variants in East-Asian populations. From these variants. Conditional analysis was performed using Genome-Wide Complex Trait Analysis (GCTA – v1.91.4beta) (263). GCTA conditional analysis was performed conditioning on the lead variant of the locus (rs7679673) using LD reference data generated from European individuals of the 1,000 Genomes Project, which was compiled using PLINK (v1.90) (264). For a variant to be considered as a potential second independent signal at the GWAS locus, a $p_{(conditional)}$ threshold was set at 5×10^{-7}.

2.2.2 – Bayesian Fine-Mapping of the 4q24 Locus of Chromosome 4 via PAINTOR

Following the identification of independent GWAS signal(s) at the locus, windows of the genome were selected either side of the lead variant of each independent signal – the size of which was adjusted in order to capture the complete LD structure of the signal. PAINTOR (V3.0) input files were generated for each signal, which included a locus file with an association statistic (Wald Statistic) for each variant (β/Standard Error). In addition to the locus file, an LD reference matrix of Pearson's correlation coefficients between all variants at the locus was calculated using 1,000 Genomes Project reference data from European individuals. The final input file was a binary matrix indicating overlap with functional annotation data downloaded from ENCODE (251,265,266).

The PAINTOR pipeline allows for the assumption that there are up to three causal variants at a locus which underpins its association with a trait or disease (251). Therefore, results were generated for each locus assuming there were a maximum of either one, two or three causal variant(s) at the locus. PAINTOR recommends including no more than five functional annotations for each locus, therefore multiple annotation sets were generated for each locus

Cohort #:	Cohort Name:	Country/Countries of Origin:	# Cases:	# Controls:	Total:
1	COIN	UK	1,950	2,162	4,112
2	GECCO OmniExpress Exome	Germany/The Netherlands/USA	4,439	4,115	8,554
3	CORECT-EUR	Australia/Canada/Finland/Germany/Israel/Spain/Sweden/UK/USA	19,948	12,124	32,072
4	CORSA-1	Austria	1,460	774	2,234
5	Croatia	Croatia	689	441	1,130
6	GECCO Oncoarray Custom iSelect	Austria/Czech Republic/Spain/Sweden/UK/USA	11,835	11,856	23,691
7	DACHS-4	Germany	1,028	661	1,689
8	FIN2	Finland	1,760	14,132	15,892
9	GECCO Stage I Pooled	Australia/Canada/France/Germany/USA	11,895	14,659	26,554
10	NSCCG	UK	6,596	7,205	13,801
11	SCOT	UK	2,910	4,095	7,005
12	Scotland Phase I	Scotland	932	943	1,875
13	SOCCS-GS	Scotland	4,551	8,804	13,355
14	SOCCS-LBC	Scotland	996	1,297	2,293
15	UK-1	UK	890	900	1,790
16	UK Biobank	UK	4,800	20,289	25,089
17	VQ58	UK	1,794	2,686	4,480

Table 2.1 – Cohort Details of the European GWAS Meta-Analysis: Cohort information of each of the seventeen cohorts that make up the European colorectal cancer Genome-Wide Association Study (GWAS) meta-analysis of 185,616 individuals of European descent. Included are the cohort number (Cohort #), cohort name, the country/countries of origin, the number of colorectal cancer cases (# Cases), number of controls (# Controls) and total number of participants from each cohort. COIN = Combination therapy with or without cetuximab as first-line therapy in treating patients with metastatic colorectal cancer. GECCO = Genetics and Epidemiology of Colorectal Cancer Consortium. CORECT-EUR = Colorectal Cancer Transdisciplinary Study – Europeans. CORSA-1 = Colorectal Cancer Study of Austria. DACHS-4 = Darmkrebs: Chancen der Verhütung durch Screening. FIN2 = Finnish GWAS. NSCCG = National Study of Colorectal Cancer Genetics. SCOT = Short-Course Oncology Treatment. SOCCS-GS = Study of Colorectal Cancer in Scotland (Cases) vs Generation Scotland (Controls). SOCCS-LBC = Study of Colorectal Cancer in Scotland (Cases) vs Lothian Birth Cohort (Controls). UK-1 = Colorectal Tumour Gene Identity Study. VQ58 = VICTOR & QUASAR2 Trials (Cases) vs 1958 Birth Cohort (Controls). UK = United Kingdom. USA = United States of America.

Cohort #:	Cohort Name:	Country/Countries of Origin:	# Cases:	# Controls:	Total:
1	Aichi 1	Japan	401	939	1,340
2	Aichi 2	Japan	224	457	681
3	BBJ	Japan	6,692	27,178	33,870
4	Guangzhou 1	China	638	971	1,609
5	HCES	Republic of Korea	3,130	4,625	7,755
6	HCES 2	Republic of Korea	3,445	2,519	5,964
7	Korea KCPSII	Republic of Korea	325	975	1,300
8	Korea NCC	Republic of Korea	1,313	1,223	2,536
9	Korea NCC 2	Republic of Korea	622	832	1,454
10	Korea Seoul	Republic of Korea	773	619	1,392
11	Shanghai 1	China	474	2,626	3,100
12	Shanghai 2	China	254	650	904
13	Shanghai 3	China	2,575	1,336	3,911
14	Shanghai 4	China	865	2,494	3,359

Table 2.2 – Cohort Details of the Asian GWAS Meta-Analysis: Cohort information of each of the fourteen cohorts that make up the Asian colorectal cancer Genome-Wide Association Study (GWAS) meta-analysis of 69,175 individuals of East-Asian descent. Included are the cohort number (Cohort #), cohort name, the country/countries of origin, the number of colorectal cancer cases (# Cases), number of controls (# Controls) and total number of participants from each cohort. BBJ = BioBank Japan Colorectal Cancer Study. HCES = Hwasun Cancer Epidemiology Study – Colon & Rectum Cancer. KCPSII = Korean Cancer Prevention Study II. NCC = National Cancer Center Colorectal Cancer Study.

of annotations that would be functionally relevant in the context of the colon. Details of each annotation set used in the analysis is provided in Table 2.3. For each locus, a 95% credible set was generated for each condition and summary figures were generated using the Correlation and Annotation Visualisation (CANVIS) tool provided with PAINTOR V3.0.

Annotation #:	Annotation Set 1:	Annotation Set 2:
1	Gencode Exon (hg19)	DNase Peaks (E075 – Colonic Mucosa)
2	Human Permissive Enhancers (Phase I/II)	H3K4ac (E075 – Colonic Mucosa)
3	Promoter Regions (E075 – Colonic Mucosa)	H3K27ac (E075 – Colonic Mucosa)
4	Enhancer Regions (E075 – Colonic Mucosa)	H3K4me1 (E075 – Colonic Mucosa)
5	All Transcription Factor Binding Sites	H3K4me3 (E075 – Colonic Mucosa)

Table 2.3 – Functional Annotations Used in PAINTOR Analysis: A list of functional annotations downloaded from The Encyclopaedia of DNA Elements (ENCODE) used to generate 95% credible sets for Probabilistic Annotation Integrator (PAINTOR) Bayesian fine-mapping analysis.

2.2.3 – Functional Annotation of Credible Variants

Upon identification of the 95% credible sets for each locus, downstream *in silico* functional analysis was performed in RegulomeDB, assigning each variant within the credible set a score depending on proximity to active promoters, enhancers and transcription factor binding sites in a variety of cell types (267). Variants were assessed in the context of intestinal tissues for functionality. Variants were also investigated for previous associations with cancer via a search of previously published GWAS data from the GWAS Catalogue (261). Annotation data for promoters, enhancers, exons, promoter-flanking regions, open chromatin, CCCTC binding factor (*CTCF*) binding sites and transcription factor binding sites were downloaded from Ensembl BioMart (hg19) for the CRC cell line HCT116, large intestine and sigmoid colon (268). Credible set variants were checked for overlap with any of these features using BEDTools (v2.30). To further characterise the credible set, each variant was searched in the Search Candidate Cis-Regulatory Elements by ENCODE (SCREEN – V2) database (269) for details on the epigenetic and transcriptomic annotations for each candidate causal variant.

2.2.4 – Mendelian Randomisation Analysis of the 4q24 Locus of Chromosome 4

In order to implement TWMR on the 4q24 locus of chromosome 4, All SNP-Gene association eQTL data was downloaded from European individuals in the GTEx (v8) database for the transverse colon, breast, prostate, whole blood and Epstein-Barr Virus (EBV) transformed lymphocytes (270). For the genomic region used in PAINTOR analysis and functional annotation, eQTL data was available for thirteen genes. In order to accurately assess the probability of causality for a gene at the locus, TWMR requires an input of conditionally independent eQTL data that have been pruned with an r^2 threshold of < 0.1 in

order to avoid multi-collinearity interference (255). Independent eQTLs were identified as described by Porcu *et al.,* with some modifications to the $p_{(eQTL)}$ threshold due to differences in sample sizes of respective eQTL datasets. For an individual gene at the locus, all eQTLs with a $p_{(eQTL)} < 0.005$ were selected for further analysis. Stepwise conditional analysis via GCTA (see section 2.2.1) was performed starting with the most significant eQTL with a $p_{(conditional)}$ threshold of 0.05 used to determine independence. Following stepwise conditional analysis, each independent eQTL for a gene was checked for overlap with other genes at the same locus – i.e. determining whether a SNP significantly affected the expression of multiple genes. If a SNP was an independent eQTL for more than one gene, all other independent eQTLs for the associated gene(s) were also added to the model. Finally, the list of independent eQTLs was pruned to only include variants with an $r^2 < 0.1$ using plink (v1.90p). A summary of this approach is outlined in Figure 2.3. Following the selection of eQTLs for the associated gene(s), effect sizes of the variant on the overall CRC phenotype were extracted from the GWAS meta-analysis and included in the input matrix. Effect sizes for each variant from the GWAS and eQTL studies were calculated as described in the example input data provided by Porcu *et al.* for the TWMR pipeline and were calculated as:

$$\text{Effect Size} = (\beta/\text{SE})/\sqrt{\text{N}}$$

Where β is the effect size of a variant on gene expression (eQTL data) or CRC (GWAS data), SE is the standard error of this effect size and N is the number of samples. Also included as a part of the TWMR input is a pairwise correlation matrix between each variant, calculated from 1,000 Genomes Project reference data in European individuals. The TWMR programme was then used to calculate the causal effect size of each gene at the 4q24 locus of chromosome 4 on the CRC phenotype.

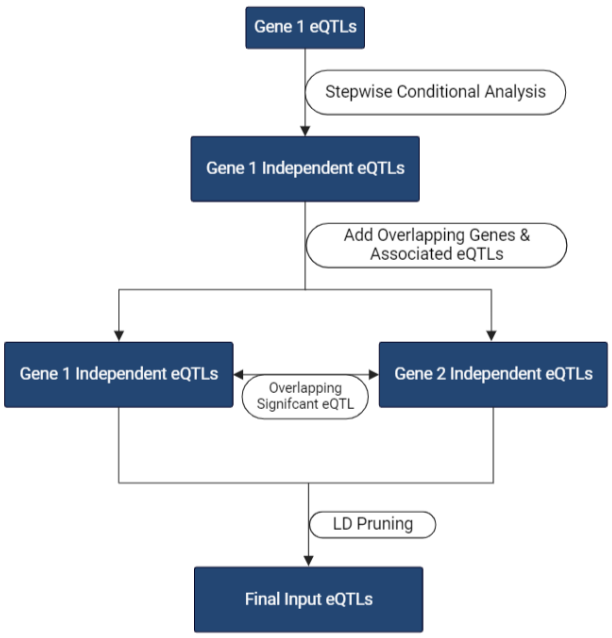


Figure 2.3 – An Overview of TWMR eQTL Selection: An illustration of the procedure by which expression quantitative trait loci (eQTLs) were selected for inclusion in Transcriptome-Wide Mendelian Randomisation (TWMR) analysis of genes at the 4q24 locus of chromosome 4. Firstly, significant eQTLs ($p_{(eQTL)} < 0.005$) were selected for the gene of interest (Gene 1) and reduced to only independent eQTLs, as determined by stepwise conditional analysis ($p_{(conditional)} < 0.05$). If any of the independent eQTLs were also independent eQTLs for other genes at the locus (Gene 2), the eQTLs for this gene were also be included in the final model. Once a complete list of eQTLs has been defined, eQTLs were then pruned to a linkage disequilibrium (LD) r^2 threshold of 0.1. Adapted from Figure 1a in Porcu *et al*. Created with BioRender.com (https://app.biorender.com/).

2.2.5 – Genotype-Expression Analysis of the 4q24 Locus of Chromosome 4

TWAS data was obtained from Fernandez-Rozadilla *et al.,* who performed a GWAS meta-analysis of 100,204 cases and 154,587 controls across the thirty-one cohorts of European or East-Asian descent listed in Tables 2.1 and 2.2 (257). This GWAS data was integrated with gene expression data from 1,107 participants across six cohorts of normal colorectum, as well as GTEx gene expression data from a number of tissues, in order to construct TWAS models using S-PrediXcan and S-MultiXcan (257). Conditional analysis was performed on the TWAS models using summary statistics-based mixed effects score test (sMIST), conditioning on the lead GWAS risk variant at the locus to identify additional genetic variants associated with altered gene expression in the context of CRC (257,271). TWAS data was extracted for genes at the 4q24 locus of chromosome 4.

In addition to this, SNP genotype and gene expression data was available for 109 individuals of self-reported European descent from the INTERMPHEN study (257). Trimmed Mean of M-Values (TMM) gene expression data was available for genes at the 4q24 locus of chromosome 4 and could be coupled with SNP genotype data for the candidate causal variants at the locus. Using this data, plus co-variates including subject age and sex, regressions between candidate causal gene expression and SNP genotype could be calculated. These regressions could therefore be used to identify significant correlations between candidate causal gene expression and the number of risk alleles of previously identified CRC GWAS variants. A non-parametric Jonckheere-Terpstra test was used to test for significant trends between gene expression and number of GWAS risk alleles via the R package *clinfun* (V1.1) (272).

2.3 - Results

2.3.1 – Conditional Analysis Identifies Multiple Independent Signals in European Populations

CRC GWAS meta-analysis data of 185,616 individuals across seventeen cohorts of European descent were used to assess the genomic region located between chr4:105,067,842 –

chr4:107,200,960 (hg19). This region equates to a region 1Mb upstream of the transcription start site of the ten-eleven translocation 2 (*TET2*) gene and 1Mb downstream of the *TET2* transcription termination site. This region was selected to capture the LD structure of all variants that were associated with CRC development at genome-wide significance ($p_{(GWAS)} < 5 \times 10^{-8}$). In Europeans, this region contained 8,166 variants after filtering out variants that displayed evidence of heterogeneity between cohorts. Of these variants, sixty-three were associated with CRC at genome-wide significance and the lead variant was rs7679673 ($p_{(GWAS)} = 2.14 \times 10^{-14}$), a variant which has previously been associated with prostate cancer (273,274). All genome-wide significant variants at the locus were located within 250 kilobases (kb) of the lead variant. This region is shown in Figure 2.4.

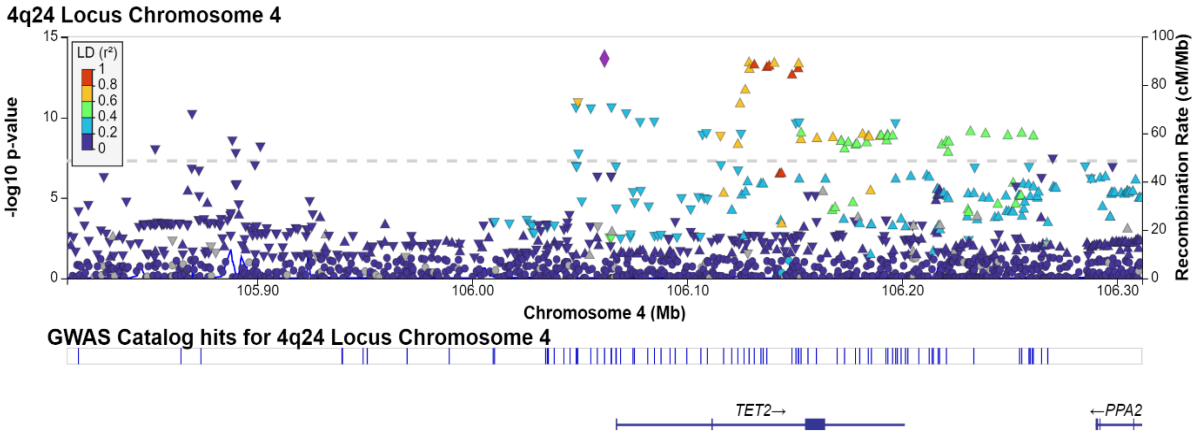


Figure 2.4 – European CRC GWAS Meta-Analysis of the 4q24 Locus of Chromosome 4: A Locus Zoom plot of colorectal cancer Genome-Wide Association study (GWAS) meta-analysis of individuals of European descent. Shown is a 500 kilobase (kb) region of the 4q24 locus of chromosome 4. The Y-axis shows the $-\log_{10}$ GWAS p-value of each SNP at the locus. The colour of each point on the plot represents linkage disequilibrium (LD) measurements in European populations of a variant relative to the lead SNP. The dashed line indicates a $p_{(GWAS)}$ of 5×10^{-8}, used in this study to indicate genome-wide significance. Plot generated using LocusZoom (https://my.locuszoom.org).

As demonstrated in Figure 2.4, the LD structure of the lead variant in Europeans (purple diamond) appears to capture a number of variants that have an association with CRC at genome-wide significance at an r^2 value of at least 0.4, represented by the red, orange and green points in Figure 2.4. This indicates that all of these variants may represent a single GWAS signal, with the significant p-values of many of the variants being in part due to their strong LD with the lead variant. However, there are a number of variants which have genome-wide significant associations with CRC that are in weaker LD with the lead variant, as illustrated by the lighter blue and darker blue points in Figure 2.4. The presence of these variants suggests that there may be more than one GWAS signal significantly associated with CRC development at this locus in Europeans. Therefore, conditional analysis via GCTA was performed, conditioning on the lead variant rs7679673 using LD reference data from European individuals in the 1,000 Genomes Project. In order to identify potentially additional variants independent of the lead SNP significantly associated with CRC development a stringent $p_{(conditional)}$ threshold was set at 5×10^{-7}. The results of this analysis are summarised in Table 2.4 and illustrated in Figure 2.5.

Variant #:	SNP ID:	Chromosome #:	Position (hg19):	$P_{(GWAS)}$:	$P_{(conditional)}$:	r^2 with rs7679673 (Europeans):
1	rs4698928	4	105,852,553	8.48×10^{-9}	4.2×10^{-8}	0.002
2	rs7655284	4	105,869,702	5.44×10^{-11}	2.50×10^{-7}	0.047
3	rs902443	4	105,888,417	2.44×10^{-9}	8.48×10^{-9}	0.001
4	rs1490586	4	105,890,017	1.45×10^{-8}	3.08×10^{-6}	0.036
5	rs2087589	4	105,901,473	5.72×10^{-9}	8.41×10^{-7}	0.007
6	rs17035289	4	106,048,291	2.19×10^{-11}	0.000331	0.224
7	rs4698932	4	106,049,147	1.03×10^{-11}	0.512483	0.698
8	rs10488913	4	106,049,416	1.60×10^{-8}	0.005277	0.192
9	rs6843555	4	106,055,150	2.46×10^{-11}	0.00092	0.246
10	rs7679673	4	106,061,534	2.14×10^{-14}	NA	1
11	rs17035310	4	106,064,754	2.11×10^{-11}	0.000875	0.243
12	rs763480	4	106,071,597	4.98×10^{-11}	0.002034	0.243
13	rs11728350	4	106,078,097	1.74×10^{-11}	0.003658	0.243
14	rs6825684	4	106,084,643	1.64×10^{-10}	0.003713	0.243
15	rs116001054	4	106,107,109	1.07×10^{-9}	0.002896	0.226
16	rs17508261	4	106,108,902	8.48×10^{-10}	0.002579	0.226
17	rs1015521	4	106,115,450	1.26×10^{-9}	0.644034	0.772
18	rs2047408	4	106,123,582	4.83×10^{-9}	0.452257	0.772
19	rs6533182	4	106,124,585	1.44×10^{-11}	0.660957	0.767
20	rs75321784	4	106,125,022	9.47×10^{-10}	0.002089	0.216
21	rs1909122	4	106,127,004	1.99×10^{-12}	0.633154	0.758
22	rs1391441	4	106,128,760	3.97×10^{-14}	0.214409	0.69
23	rs7663401	4	106,128,954	1.02×10^{-13}	0.655983	0.821
24	rs2007403	4	106,131,210	5.52×10^{-14}	0.682046	0.826
25	rs2047409	4	106,137,033	7.85×10^{-14}	0.719808	0.826
26	rs11735256	4	106,138,146	6.54×10^{-14}	0.667229	0.821
27	rs7655890	4	106,140,501	4.31×10^{-14}	0.219726	0.695
28	rs7674220	4	106,148,758	2.33×10^{-13}	0.728197	0.798
29	rs143875052	4	106,150,555	2.08×10^{-10}	0.001585	0.216

Variant #:	SNP ID:	Chromosome #:	Position (hg19):	$P_{(GWAS)}$:	$P_{(conditional)}$:	r^2 with rs7679673 (Europeans):
30	rs1391439	4	106,151,642	9.42×10^{-14}	0.723129	0.821
31	rs1391438	4	106,151,843	4.62×10^{-14}	0.195103	0.686
32	rs75625682	4	106,152,037	1.85×10^{-10}	0.001545	0.216
33	rs7658539	4	106,152,877	2.41×10^{-9}	0.929467	0.601
34	rs7683416	4	106,152,984	9.38×10^{-10}	0.881941	0.585
35	rs7670522	4	106,160,365	2.07×10^{-9}	0.942204	0.604
36	rs62330911	4	106,167,744	1.84×10^{-9}	0.982603	0.597
37	rs2647246	4	106,171,635	3.05×10^{-9}	0.865769	0.577
38	rs2647247	4	106,171,652	3.11×10^{-9}	0.868059	0.577
39	rs2726518	4	106,173,199	9.23×10^{-9}	0.754464	0.526
40	rs2647248	4	106,174,936	3.67×10^{-9}	0.90498	0.581
41	rs2133086	4	106,178,289	5.32×10^{-9}	0.961492	0.6
42	rs7678440	4	106,178,902	4.47×10^{-9}	0.983808	0.6
43	rs1032625	4	106,181,573	1.13×10^{-9}	0.938298	0.594
44	rs2726458	4	106,183,474	3.80×10^{-9}	0.544234	0.47
45	rs2454205	4	106,184,229	1.49×10^{-9}	0.991155	0.597
46	rs2726459	4	106,184,597	1.66×10^{-9}	0.994803	0.597
47	rs2726521	4	106,189,614	1.58×10^{-9}	0.979219	0.594
48	rs2647250	4	106,190,226	1.36×10^{-9}	0.887644	0.578
49	rs2647230	4	106,192,775	2.88×10^{-9}	0.458192	0.454
50	rs2726520	4	106,193,160	1.26×10^{-10}	0.883641	0.583
51	rs2726519	4	106,193,334	1.46×10^{-9}	0.901908	0.583
52	rs2647244	4	106,195,572	1.42×10^{-9}	0.901149	0.587
53	rs34402524	4	106,196,829	2.04×10^{-10}	0.001693	0.229
54	rs2647238	4	106,218,428	5.05×10^{-9}	0.507948	0.439
55	rs2466920	4	106,220,558	3.23×10^{-9}	0.398264	0.434
56	rs114781825	4	106,221,371	1.47×10^{-8}	NA	0.438
57	rs2454203	4	106,221,741	3.37×10^{-9}	0.401969	0.434
58	rs141186521	4	106,231,747	7.71×10^{-10}	NA	0.412

Variant #:	SNP ID:	Chromosome #:	Position (hg19):	$P_{(GWAS)}$:	$P_{(conditional)}$:	r^2 with rs7679673 (Europeans):
59	rs2726507	4	106,238,767	1.05×10^{-9}	0.407171	0.434
60	rs2726475	4	106,249,102	1.38×10^{-9}	0.449815	0.437
61	rs2726479	4	106,255,589	9.75×10^{-10}	0.539811	0.468
62	rs2726482	4	106,261,006	1.45×10^{-9}	0.433964	0.433
63	rs78479210	4	106,270,017	3.21×10^{-8}	0.000535	0.123

Table 2.4 – Conditional Analysis on rs7679673: Conditional analysis of all variants in the European Genome-Wide Association Study (GWAS) colorectal cancer (CRC) meta-analysis that were significantly associated with colorectal cancer at genome-wide significance ($p_{(GWAS)} < 5 \times 10^{-8}$). Variants were conditioned on the lead variant at the locus (rs7679673 – $p_{(GWAS)} = 2.14 \times 10^{-14}$) using linkage disequilibrium estimates from European individuals from the 1,000 Genomes Project data. Columns include the unique identifier of the variant (SNP ID), chromosome number, genome co-ordinates (hg19), GWAS p-value ($P_{(GWAS)}$), conditional p-value ($P_{(conditional)}$) and r^2 correlation with the lead variant in European individuals – calculated from LD-Link (https://ldlink.nci.nih.gov). NA = Not Applicable.

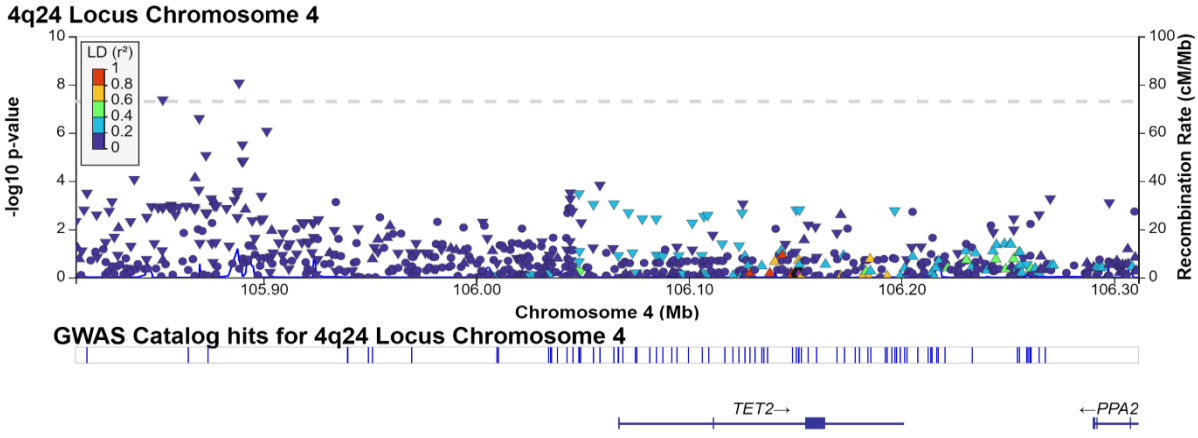


Figure 2.5 – Conditional Analysis on rs7679673: A Locus Zoom plot of colorectal cancer Genome-Wide Association study (GWAS) meta-analysis of individuals of European descent. Shown is a 500 kilobase (kb) region of the 4q24 locus of chromosome 4, comprised of the regions ±250kb relative to the lead SNP rs7679673. The Y-axis shows the $-\log_{10}$ of the conditional GWAS p-value of a variant when conditioned on the lead SNP. The colour of each point on the plot represents linkage disequilibrium (LD) measurements in European populations of a variant relative to the rs1391439, a proxy variant of the lead SNP ($r^2 = 0.821$). Plot generated using LocusZoom (https://my.locuszoom.org).

As seen in Figure 2.5, the majority of variants previously significantly associated with CRC in Figure 2.4 no longer had significant p-values, indicating that they were part of the same GWAS signal on account of their LD with the lead SNP. This is confirmed in Table 2.4, as the $p_{(conditional)}$ of the majority of variants were above the threshold for independence. Two variants, rs114781825 and rs141186521, were not able to have a $p_{(conditional)}$ calculated due to their absence from 1,000 Genomes reference data. However, the r^2 with rs7679673 (0.438 and 0.412 respectively) indicates they were unlikely to be independent signals. Only three variants in Table 2.4 presented with a $p_{(conditional)} < 5 \times 10^{-7}$. These variants were: rs4698928 ($p_{(conditional)} = 4.2 \times 10^{-8}$), rs7655284 ($p_{(conditional)} = 2.5 \times 10^{-7}$) and rs902443 ($p_{(conditional)} = 8.48 \times 10^{-9}$). These variants had an r^2 with the lead SNP in Europeans of 0.002, 0.047 and 0.001 respectively, and an LD ranging between 0.439 – 0.908 with one another, indicating that there was potentially a second signal significantly associated with CRC at this locus. To confirm this, whilst excluding the possibility of more independent signals, conditional analysis was repeated, conditioning on both rs7679673 and the most significant of the three candidate independent variants (rs7655284 – $p_{(GWAS)} = 5.44 \times 10^{-11}$). The results of this additional conditional analysis are summarised in Table 2.5 and Figure 2.6.

As seen in Table 2.5, when conditioning on both rs7679673 and rs7655284, no additional variants meet the $p_{(conditional)}$ threshold to be considered as further independent GWAS signals. However, there were two variants that were close to this threshold – rs10488913 ($p_{(GWAS)} = 1.60 \times 10^{-8}$) and rs78479210 ($p_{(GWAS)} = 3.21 \times 10^{-8}$) presented with $p_{(conditional)}$ values of 8.48×10^{-5} and 7.81×10^{-6} respectively. In conclusion, there appear to be at least two independent GWAS signals at the 4q24 locus of chromosome 4 significantly associated with CRC development, captured by the variants rs7679673 ($p_{(GWAS)} = 2.14 \times 10^{-14}$) and rs7655284 ($p_{(GWAS)} = 5.44 \times 10^{-11}$), the latter representing a potentially novel association with CRC that has not yet been reported. While there is the potential for additional independent signals at this locus, maintaining a stringent $p_{(conditional)}$ threshold improves the reliability of each signal for downstream fine-mapping analysis.

Variant #:	SNP ID:	Chromosome #:	Position (hg19):	$P_{(GWAS)}$:	$P_{(conditional)}$:	r^2 with rs7679673 (Europeans):	r^2 with rs7655284 (Europeans):
1	rs4698928	4	105,852,553	8.48×10^{-9}	0.068298	0.002	0.475
2	rs7655284	4	105,869,702	5.44×10^{-11}	NA	0.047	1
3	rs902443	4	105,888,417	2.44×10^{-9}	0.025925	0.001	0.435
4	rs1490586	4	105,890,017	1.45×10^{-8}	0.680317	0.036	0.661
5	rs2087589	4	105,901,473	5.72×10^{-9}	0.002017	0.007	0.084
6	rs17035289	4	106,048,291	2.19×10^{-11}	0.000563	0.224	0.014
7	rs4698932	4	106,049,147	1.03×10^{-11}	0.908937	0.698	0.081
8	rs10488913	4	106,049,416	1.60×10^{-8}	8.48×10^{-5}	0.192	0.007
9	rs6843555	4	106,055,150	2.46×10^{-11}	0.00182	0.246	0.02
10	rs7679673	4	106,061,534	2.14×10^{-14}	NA	1	0.047
11	rs17035310	4	106,064,754	2.11×10^{-11}	0.000236	0.243	0.003
12	rs763480	4	106,071,597	4.98×10^{-11}	0.000587	0.243	0.003
13	rs11728350	4	106,078,097	1.74×10^{-10}	0.001121	0.243	0.003
14	rs6825684	4	106,084,643	1.64×10^{-10}	0.001139	0.243	0.003
15	rs116001054	4	106,107,109	1.07×10^{-9}	0.001053	0.226	0.004
16	rs17508261	4	106,108,902	8.48×10^{-10}	0.000929	0.226	0.004
17	rs1015521	4	106,115,450	1.26×10^{-9}	0.714656	0.772	0.033
18	rs2047408	4	106,123,582	4.83×10^{-9}	0.497098	0.772	0.033
19	rs6533182	4	106,124,585	1.44×10^{-11}	0.605956	0.767	0.032
20	rs75321784	4	106,125,022	9.47×10^{-10}	0.000661	0.216	0.002
21	rs1909122	4	106,127,004	1.99×10^{-12}	0.574288	0.758	0.03
22	rs1391441	4	106,128,760	3.97×10^{-14}	0.14904	0.69	0.028
23	rs7663401	4	106,128,954	1.02×10^{-13}	0.516718	0.821	0.04
24	rs2007403	4	106,131,210	5.52×10^{-14}	0.56211	0.826	0.04
25	rs2047409	4	106,137,033	7.85×10^{-14}	0.596979	0.826	0.04
26	rs11735256	4	106,138,146	6.54×10^{-14}	0.572632	0.821	0.039
27	rs7655890	4	106,140,501	4.31×10^{-14}	0.165037	0.695	0.029
28	rs7674220	4	106,148,758	2.33×10^{-13}	0.603418	0.798	0.035
29	rs143875052	4	106,150,555	2.08×10^{-10}	0.000466	0.216	0.002

Variant #:	SNP ID:	Chromosome #:	Position (hg19):	$P_{(GWAS)}$:	$P_{(conditional)}$:	r^2 with rs7679673 (Europeans):	r^2 with rs7655284 (Europeans):
30	rs1391439	4	106,151,642	9.42×10^{-14}	0.653964	0.821	0.042
31	rs1391438	4	106,151,843	4.62×10^{-14}	0.155839	0.686	0.03
32	rs75625682	4	106,152,037	1.85×10^{-10}	0.000453	0.216	0.002
33	rs7658539	4	106,152,877	2.41×10^{-9}	0.94929	0.601	0.043
34	rs7683416	4	106,152,984	9.38×10^{-10}	0.821583	0.585	0.037
35	rs7670522	4	106,160,365	2.07×10^{-9}	0.952977	0.604	0.044
36	rs62330911	4	106,167,744	1.84×10^{-9}	0.985081	0.597	0.042
37	rs2647246	4	106,171,635	3.05×10^{-9}	0.770121	0.577	0.037
38	rs2647247	4	106,171,652	3.11×10^{-9}	0.772348	0.577	0.037
39	rs2726518	4	106,173,199	9.23×10^{-9}	0.66609	0.526	0.032
40	rs2647248	4	106,174,936	3.67×10^{-9}	0.817242	0.581	0.037
41	rs2133086	4	106,178,289	5.32×10^{-9}	0.996975	0.6	0.044
42	rs7678440	4	106,178,902	4.47×10^{-9}	0.980695	0.6	0.044
43	rs1032625	4	106,181,573	1.13×10^{-9}	0.893857	0.594	0.041
44	rs2726458	4	106,183,474	3.80×10^{-9}	0.507393	0.47	0.031
45	rs2454205	4	106,184,229	1.49×10^{-9}	0.958853	0.597	0.042
46	rs2726459	4	106,184,597	1.66×10^{-9}	0.972883	0.597	0.042
47	rs2726521	4	106,189,614	1.58×10^{-9}	0.934598	0.594	0.041
48	rs2647250	4	106,190,226	1.36×10^{-9}	0.803351	0.578	0.035
49	rs2647230	4	106,192,775	2.88×10^{-9}	0.440672	0.454	0.033
50	rs2726520	4	106,193,160	1.26×10^{-9}	0.881996	0.583	0.046
51	rs2726519	4	106,193,334	1.46×10^{-9}	0.900259	0.583	0.046
52	rs2647244	4	106,195,572	1.42×10^{-9}	0.908413	0.587	0.047
53	rs34402524	4	106,196,829	2.04×10^{-10}	0.000534	0.229	0.003
54	rs2647238	4	106,218,428	5.05×10^{-9}	0.46616	0.439	0.028
55	rs2466920	4	106,220,558	3.23×10^{-9}	0.364425	0.434	0.028
56	rs114781825	4	106,221,371	1.47×10^{-8}	NA	0.438	0.036
57	rs2454203	4	106,221,741	3.37×10^{-9}	0.367935	0.434	0.028
58	rs141186521	4	106,231,747	7.71×10^{-10}	NA	0.412	0.027

Variant #:	SNP ID:	Chromosome #:	Position (hg19):	$P_{(GWAS)}$:	$P_{(conditional)}$:	r^2 with rs7679673 (Europeans):	r^2 with rs7655284 (Europeans):
59	rs2726507	4	106,238,767	1.05×10^{-9}	0.373449	0.434	0.028
60	rs2726475	4	106,249,102	1.38×10^{-9}	0.436907	0.437	0.03
61	rs2726479	4	106,255,589	9.75×10^{-10}	0.546981	0.468	0.032
62	rs2726482	4	106,261,006	1.45×10^{-9}	0.430626	0.433	0.031
63	rs78479210	4	106,270,017	3.21×10^{-8}	7.81×10^{-6}	0.123	0.005

Table 2.5 – Conditional Analysis on rs7679673 & rs7655284: More conditional analysis of all variants in the European Genome-Wide Association Study (GWAS) colorectal cancer (CRC) meta-analysis that were significantly associated with colorectal cancer at genome-wide significance ($p_{(GWAS)} < 5 \times 10^{-8}$). Variants were conditioned on the lead variant at the locus (rs7679673 – $p_{(GWAS)} = 2.14 \times 10^{-14}$) and the lead variant from a potential independent second signal (rs7655284 – $p_{(GWAS)} = 5.44 \times 10^{-11}$) using linkage disequilibrium estimates from European individuals from the 1,000 Genomes Project data. Columns include the unique identifier of the variant (SNP ID), chromosome number, genome co-ordinates (hg19), GWAS p-value ($P_{(GWAS)}$), conditional p-value ($P_{(conditional)}$) and r^2 correlation with the each variant in European individuals – calculated from LD-Link (https://ldlink.nci.nih.gov). NA = Not Applicable.

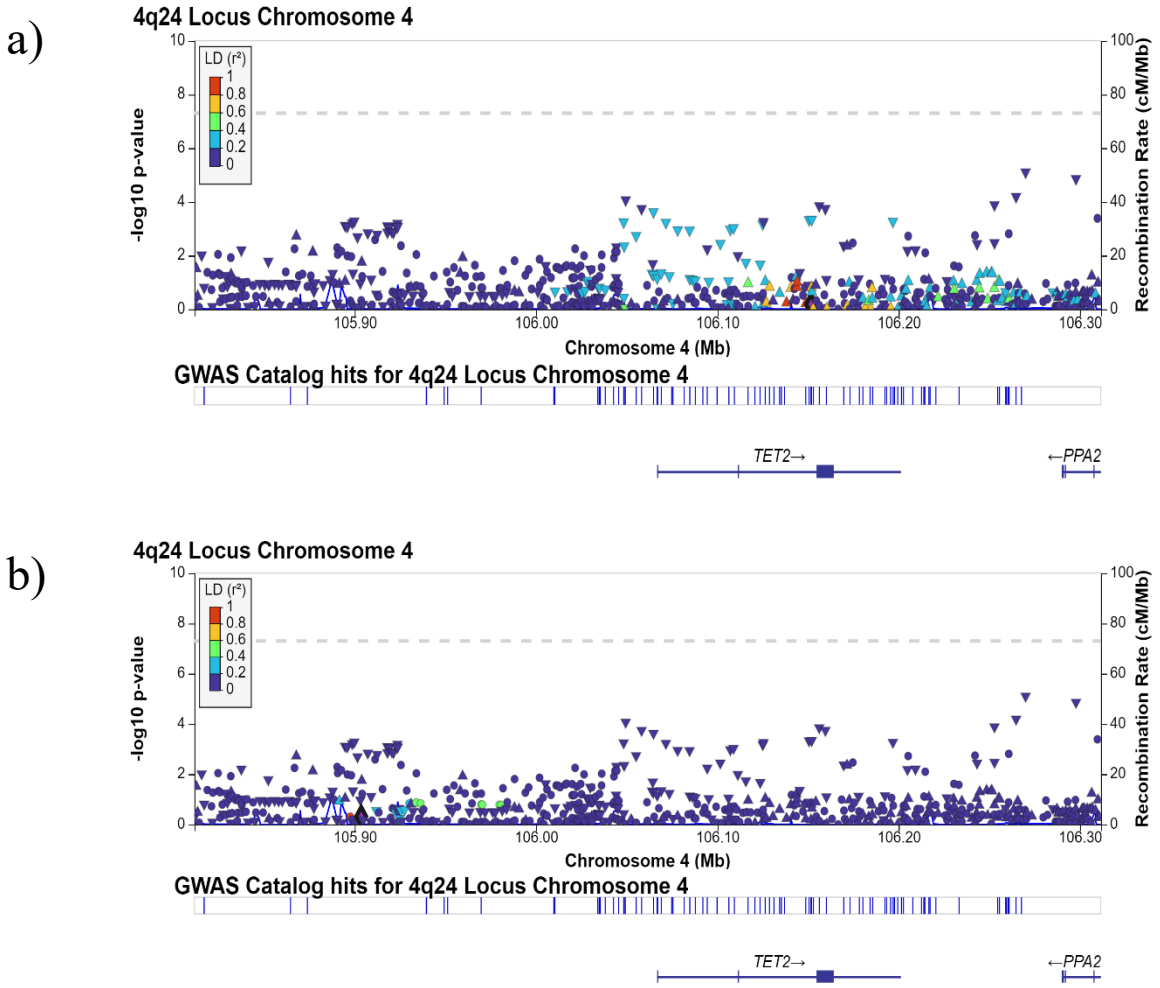


Figure 2.6 - Conditional Analysis on rs7679673 & rs7655284: A Locus Zoom plot of colorectal cancer Genome-Wide Association study (GWAS) meta-analysis of individuals of European descent. Shown is a 500 kilobase (kb) region of the 4q24 locus of chromosome 4, comprised of the regions ±250kb relative to the lead SNP rs7679673. The Y-axis shows the $-\log_{10}$ of the conditional GWAS p-value of a variant when conditioned on both rs7679673 and rs7655284. The colour of each point on the plot represents linkage disequilibrium (LD) measurements in European populations of a variant relative to proxies of each SNP used in the conditional analysis. a) LD in relation to rs1391439 – a proxy of rs7679673 ($r^2 = 0.821$). b) LD in relation to rs1490586 – a proxy of rs7655284 ($r^2 = 0.661$). Plots generated using LocusZoom (https://my.locuszoom.org).

2.3.2 – PAINTOR Analysis Identifies Thirteen Candidate Casual Variants Across All Signals

Following the identification of at least one additional independent GWAS signal at the 4q24 locus of chromosome 4, fine-mapping analysis could be performed on each signal in turn in order to identify the candidate causal variant(s) underlying each signal. To this end, Bayesian fine-mapping was employed on genomic regions capturing each signal using the PAINTOR fine-mapping pipeline. As discussed in section 2.1.3, the lead variant at a locus is not necessarily the causal variant underlying the association with a trait or disease, but may lie

nearby or be in LD with the true causal variant. To this end, regions that captured the LD structure of each signal, while retaining a large enough number of variants to be tested, were selected for PAINTOR fine-mapping. For the GWAS signal associated with rs7679673, a genomic region 150kb upstream of the SNP and 300kb downstream (chr4:105,911,534-106,361,534 – hg19) was required (see Figure 2.7a). For the second signal associated with rs7655284, a region 150kb either side of the lead variant (chr4:105,719,702-106,019,702 – hg19) was selected (see Figure 2.7b).

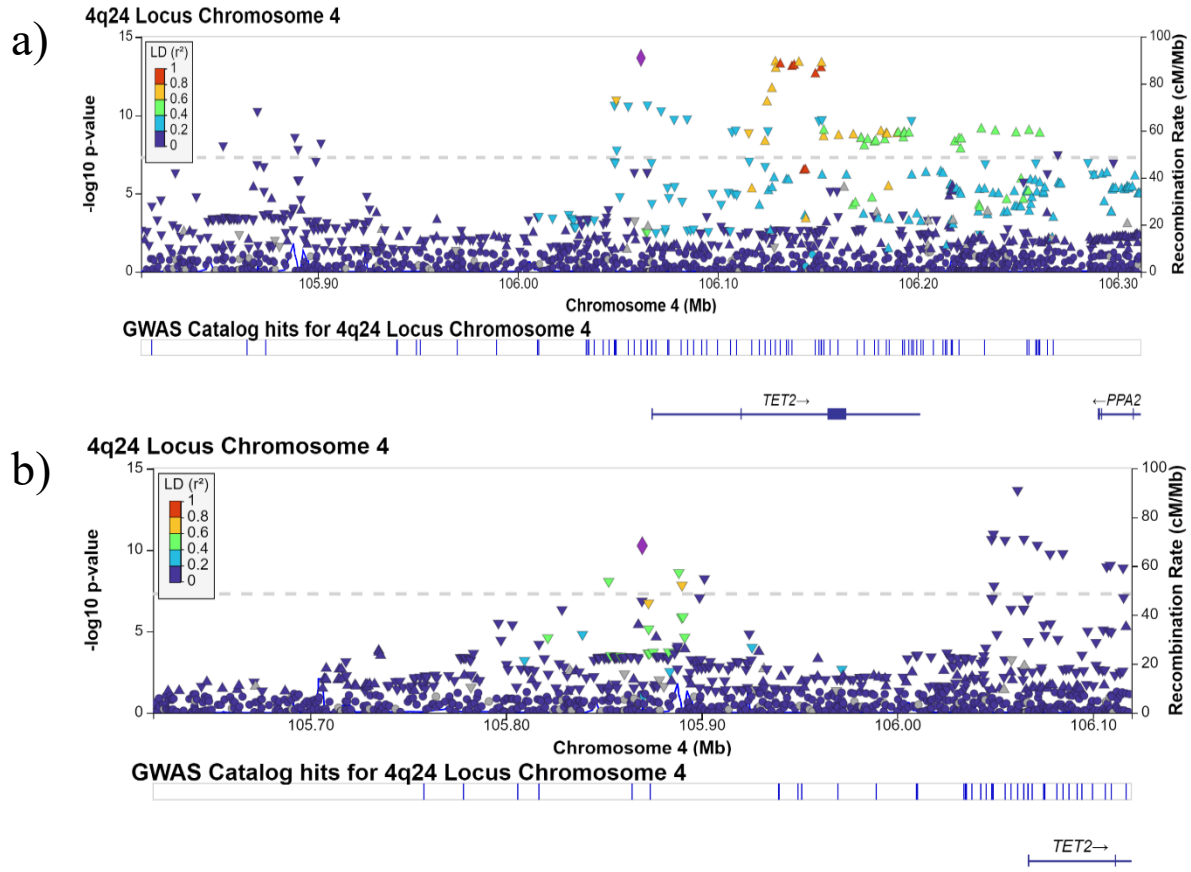


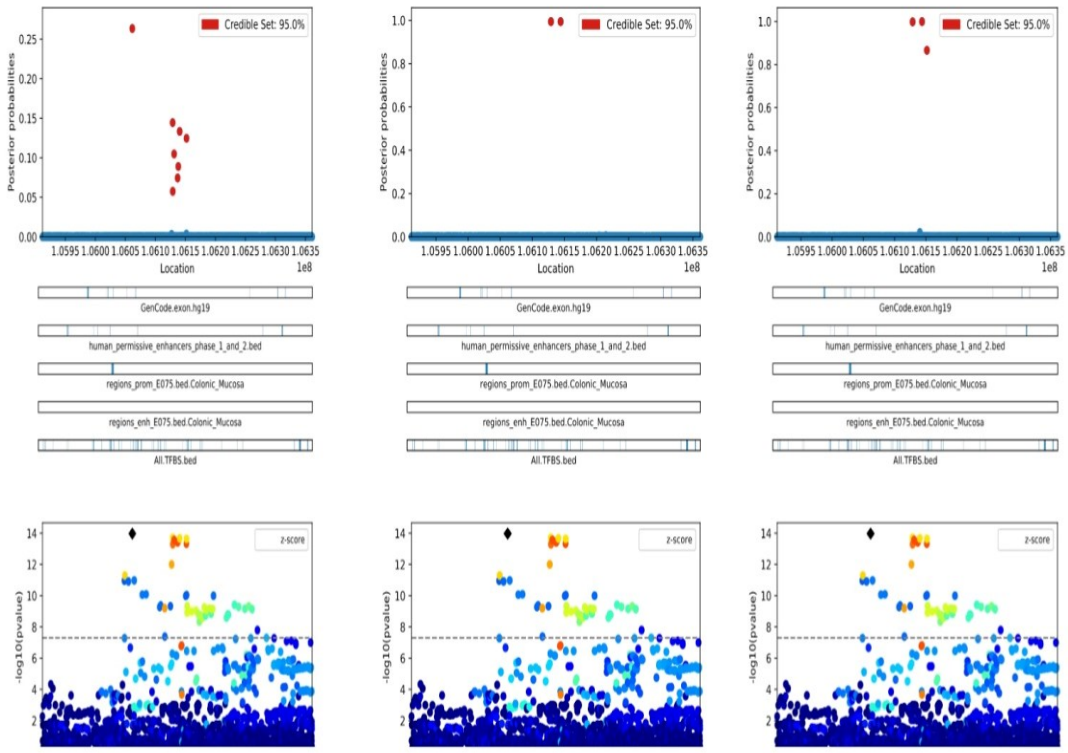
Figure 2.7 – Input Loci for PAINTOR Fine-Mapping: A Locus Zoom plot of colorectal cancer Genome-Wide Association study (GWAS) meta-analysis of individuals of European descent. Shown in each panel 500 kilobase (kb) region of the 4q24 locus of chromosome 4, comprised of the regions ±250kb relative to either rs7679673 (a) or rs7655284 (b). In each panel the named variant is represented by the purple diamond. The Y-axis shows the $-\log_{10}$ of the GWAS p-value of a variant. The colour of each point on the plot represents linkage disequilibrium (LD) measurements in European populations of a variant relative to either rs7679673 (a) or rs7655284 (b). Plots generated using LocusZoom (https://my.locuszoom.org).

For all variants within each region, a Pearson's correlation matrix between variants was computed from 1,000 Genomes Project European reference data and annotation files were generated for each of the annotation sets chosen for the fine-mapping analysis (see Table 2.3). In total, fine-mapping analysis was performed on 1,583 potentially causal variants from the rs7679673 region and 696 from the rs7655284 region. Fine-mapping analysis generated a 95% credible set of candidate causal variants assuming a maximum of either one, two or three causal variants per locus. The results for the rs7679673 region are shown in Table 2.6 and Figure 2.8.

Annotation Set 1:					
One Causal Variant per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs7679673	4	106,061,534	2.14×10^{-14}	0.263
2	rs1391441	4	106,128,760	3.97×10^{-14}	0.144
3	rs7655890	4	106,140,501	4.31×10^{-14}	0.133
4	rs1391438	4	106,151,843	4.62×10^{-14}	0.125
5	rs2007403	4	106,131,210	5.52×10^{-14}	0.105
6	rs11735256	4	106,138,146	6.54×10^{-14}	0.089
7	rs2047409	4	106,137,033	7.85×10^{-14}	0.075
8	rs7663401	4	106,128,954	1.02×10^{-13}	0.0575
Two Causal Variants per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs9884984	4	106,143,948	4.26×10^{-4}	0.994
2	rs7663401	4	106,128,954	1.02×10^{-13}	0.994
Three Causal Variants per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs9884984	4	106,143,948	4.26×10^{-4}	0.99
2	rs7663401	4	106,128,954	1.02×10^{-13}	0.998
3	rs1391438	4	106,151,843	4.62×10^{-14}	0.866
Annotation Set 2:					
One Causal Variant per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs7679673	4	106,061,534	2.14×10^{-14}	0.249
2	rs1391441	4	106,128,760	3.97×10^{-14}	0.136
3	rs7655890	4	106,140,501	4.31×10^{-14}	0.126
4	rs1391438	4	106,151,843	4.62×10^{-14}	0.118
5	rs2007403	4	106,131,210	5.52×10^{-14}	0.099
6	rs11735256	4	106,138,146	6.54×10^{-14}	0.084
7	rs2047409	4	106,137,033	7.85×10^{-14}	0.0703
8	rs1391439	4	106,151,642	9.42×10^{-14}	0.0589
9	rs7663401	4	106,128,954	1.02×10^{-13}	0.0543
Two Causal Variants per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs9884984	4	106,143,948	4.26×10^{-4}	0.994
2	rs7663401	4	106,128,954	1.02×10^{-13}	0.994
Three Causal Variants per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs9884984	4	106,143,948	4.26×10^{-4}	0.982
2	rs7663401	4	106,128,954	1.02×10^{-13}	0.973
3	rs1391438	4	106,151,843	4.62×10^{-14}	0.857

Table 2.6 – PAINTOR Fine-Mapping of the rs7679673 Region: The 95% credible set of variants underlying the significant association with colorectal cancer seen in the region around rs7679673. Credible sets were generated using Probabilistic Annotation Integrator (PAINTOR) and two sets of annotations, with the region being assumed to contain either one, two or three causal variants. Presented in the table is the identifier of the candidate causal variant (SNP ID), chromosomal co-ordinates of the variant (hg19), the Genome-Wide Association study p-value associated with the variant ($p_{(GWAS)}$) and the posterior probability that the variant is the causal variant at the locus

a) One Causal Variant per Locus: Two Causal Variants per Locus: Three Causal Variants per Locus:



b) One Causal Variant per Locus: Two Causal Variants per Locus: Three Causal Variants per Locus:

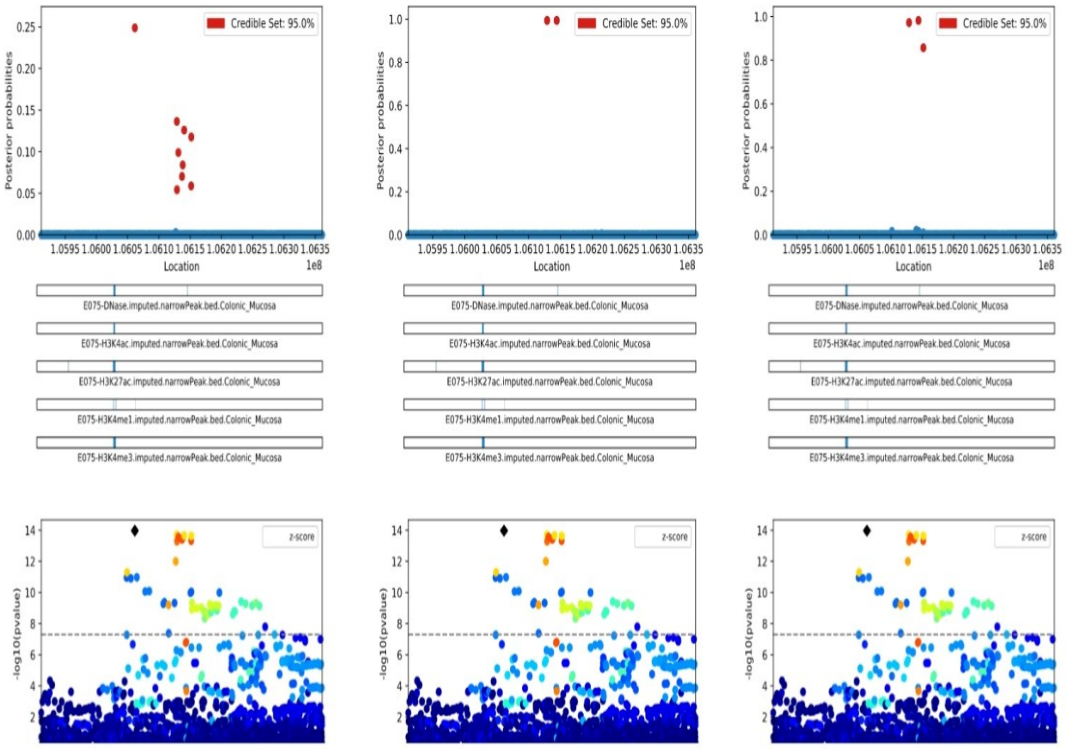


Figure 2.8 – PAINTOR Fine-Mapping of the rs7679673 Region: Graphical representations of 95% credible sets generated via Probabilistic Annotation Integrator (PAINTOR) fine-mapping analysis of the rs7679673 region. Fine-mapping analysis assumed there was either one, two or three causal variants at the locus using either Annotation Set 1 (a) or Annotation Set 2 (b). The X-axis represents the position of a variant on chromosome 4 (hg19) and the Y-axis represents the posterior probability of causality for the variant. Variants included in the 95% credible set are highlighted in red. Overlap of a variant with one of the annotations included in the analysis is indicated by blue shading on the annotation bars. Also included for each plot is the z-score (Wald Statistic) of association of a variant with colorectal cancer according to Genome-Wide Association Study Meta-Analysis. Plots generated using the Correlation and Annotation Visualisation (CANVIS) tool provided with PAINTOR V3.0.

As illustrated in Table 2.6 and Figure 2.8, a total of ten variants were included in the 95% credible sets across both annotation sets. The composition of the credible sets for each annotation set remained remarkably consistent providing a shortlist of candidate causal variants at this region. Credible set variants mostly were already significantly associated with CRC according to the GWAS meta-analysis data, with the exception of rs9884984 ($p_{(GWAS)}$ = 4.26 x 10^{-4}). This variant, alongside all other variants included in the credible set, were in strong LD with rs7679673 in European populations (r^2 ranging from 0.686 – 0.829). This indicates that the causal variant is associated with the single independent signal represented by rs7679673. The same PAINTOR analysis was then applied to the second independent signal at this locus captured by rs7655284. The results of this analysis are presented in Table 2.7 and Figure 2.9.

In total, only three variants were included in the 95% credible sets across all annotation sets for the rs7655284 region, compared to ten variants at the rs7679673 region. The credible sets generated were identical between annotation sets at this region, with two variants included in the credible set that did not have a genome-wide significant $p_{(GWAS)}$. These variants were rs113280693 ($p_{(GWAS)}$ = 0.215) and rs62331067 ($p_{(GWAS)}$ = 2.20 x 10^{-4}). These variants were in complete LD with one another (r^2 = 1 in European populations) and in modest LD with rs7655284, the lead variant at this region (r^2 = 0.14 in European populations). It is possible that the non-significant $p_{(GWAS)}$ of two of the credible set variants could be explained by their low frequencies in comparison to other variants. The variants rs113280693 and rs62331067 had a minor allele frequency of only 5.7%, which may limit the statistical power to assign a more significant $p_{(GWAS)}$ to these variants. Interestingly, the rs7655284 region appears to have a maximum of two causal variants associated with CRC. As seen in Table 2.7 and Figure 2.9, there are no differences between the 95% credible sets of PAINTOR fine-mapping when it is assumed there are a maximum of two or three causal variants at the region.

2.3.3 – *in silico* Functional Annotation of Credible Set Variants Indicates Potential Causality

Following the identification of a total of thirteen candidate causal variants underlying the two independent GWAS associations with CRC at the 4q24 locus of chromosome 4, the next stage of analysis was *in silico* functional annotation of each variant in the context of the colon. The first annotation of each variant was searching the European Bioinformatics Institute GWAS Catalogue (261) for previously identified associations of each credible

Annotation Set 1:					
One Causal Variant per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs7655284	4	105,869.702	5.44×10^{-11}	0.975
Two Causal Variants per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs113280693	4	105,875,730	0.215	0.999
2	rs62331067	4	105,886,228	2.20×10^{-4}	0.999
Three Causal Variants per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs113280693	4	105,875,730	0.215	0.999
2	rs62331067	4	105,886,228	2.20×10^{-4}	0.999
Annotation Set 2:					
One Causal Variant per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs7655284	4	105,869.702	5.44×10^{-11}	0.975
Two Causal Variants per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs113280693	4	105,875,730	0.215	0.999
2	rs62331067	4	105,886,228	2.20×10^{-4}	0.999
Three Causal Variants per Locus:					
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**Posterior Probability:**
1	rs113280693	4	105,875,730	0.215	0.999
2	rs62331067	4	105,886,228	2.20×10^{-4}	0.999

Table 2.7 – PAINTOR Fine-Mapping of the rs7655284 Region: The 95% credible set of variants underlying the significant association with colorectal cancer seen in the region around rs7655284. Credible sets were generated using Probabilistic Annotation Integrator (PAINTOR) and two sets of annotations, with the region being assumed to contain either one, two or three causal variants. Presented in the table is the identifier of the candidate causal variant (SNP ID), chromosomal co-ordinates of the variant (hg19), the Genome-Wide Association study p-value associated with the variant ($p_{(GWAS)}$) and the posterior probability that the variant is the causal variant at the locus.

set variant with cancer. The results are presented in Table 2.8. Three variants had previously published associations with cancer, the CRC risk allele of rs7679673 had previously been associated with an increased risk of prostate cancer (273,274) and the risk alleles of rs1391441 and rs2007403 had previously been associated with CRC (275). The risk alleles of rs7663401 and rs2007403 were also associated with the development of biliary cirrhosis. Affected individuals of biliary cirrhosis have been suggested to have an increased risk of cancer, particularly hepatocellular carcinoma (276). Therefore, given that some of these candidate causal variants have previously been associated with cancer predisposition, there may be some functional relevance of these variants that could be investigated in the context of colonic tissues to explain their possible association with CRC predisposition.

Next, each variant was searched in the RegulomeDB database, which provides each variant with annotation data for regulatory elements across a range of different tissues and cell types (https://regulomedb.org). Each variant was assigned a score ranging from 1 - 7, with a score of 1 indicating a variant is associated with the most regulatory features and 7 indicating the least. The RegulomeDB score of each credible set variant is presented in Table 2.8. In total, three variants had a score of 4, five variants had a score of 5, two variants had a score of 6

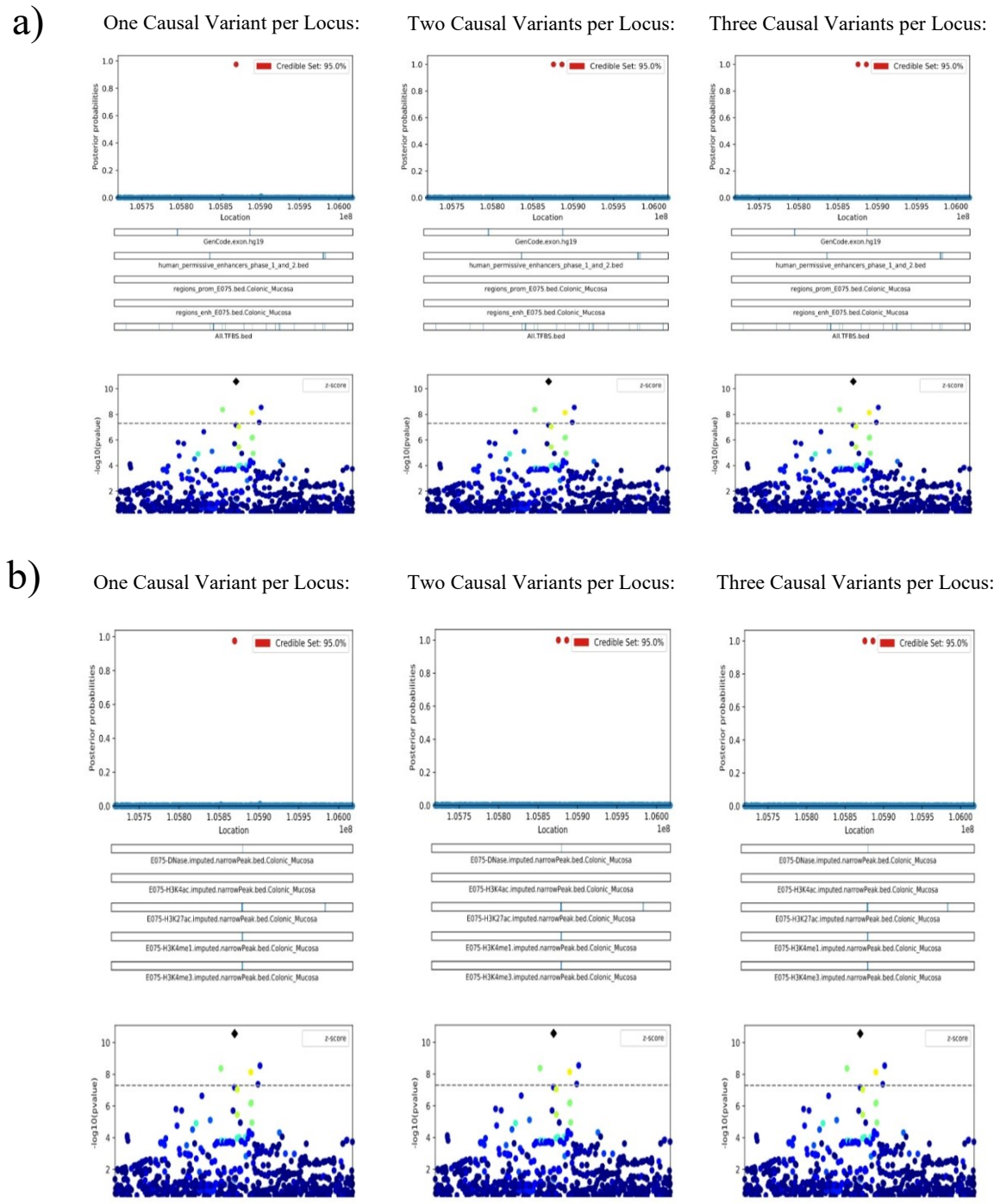


Figure 2.9 - PAINTOR Fine-Mapping of the rs7655284 Region: Graphical representations of 95% credible sets generated via Probabilistic Annotation Integrator (PAINTOR) fine-mapping analysis of the rs7655284 region. Fine-mapping analysis assumed there was either one, two or three causal variants at the locus using either Annotation Set 1 (a) or Annotation Set 2 (b). The X-axis represents the position of a variant on chromosome 4 (hg19) and the Y-axis represents the posterior probability of causality for the variant. Variants included in the 95% credible set are highlighted in red. Overlap of a variant with one of the annotations included in the analysis is indicated by blue shading on the annotation bars. Also included for each plot is the z-score (Wald Statistic) of association of a variant with colorectal cancer according to Genome-Wide Association Study Meta-Analysis. Plots generated using the Correlation and Annotation Visualisation (CANVIS) tool provided with PAINTOR V3.0.

Variant #:	SNP ID:	Chromosome #:	Position (hg19):	$P_{(GWAS)}$:	Previous Cancer Associations:	RegulomeDB Score:
1	rs7655284	4	105,869,702	5.44×10^{-11}	None	5
2	rs113280693	4	105,875,730	0.215	None	7
3	rs62331067	4	105,886,228	2.20×10^{-4}	None	5
4	rs7679673	4	106,061,534	2.14×10^{-14}	Prostate Cancer	7
5	rs1391441	4	106,128,760	3.97×10^{-14}	Colorectal Cancer	4
6	rs7663401	4	106,128,954	1.02×10^{-13}	Biliary Cirrhosis	5
7	rs2007403	4	106,131,210	5.52×10^{-14}	Colorectal Cancer / Biliary Cirrhosis	5
8	rs2047409	4	106,137,033	7.85×10^{-14}	None	4
9	rs11735256	4	106,138,146	6.54×10^{-14}	None	6
10	rs7655890	4	106,140,501	4.31×10^{-14}	None	6
11	rs9884984	4	106,143,948	4.26×10^{-4}	None	7
12	rs1391439	4	106,151,642	9.42×10^{-14}	None	4
13	rs1391438	4	106,151,843	4.62×10^{-14}	None	5

Table 2.8 – Functional Annotation of Credible Set Variants: Functional annotations of 95% credible set variants identified by Probabilistic Annotation Integrator (PAINTOR) fine-mapping of colorectal cancer Genome-Wide Association Study (GWAS) meta-analysis data. Included is the identifier of the credible set variant (SNP ID), the chromosomal co-ordinates of the variant (hg19), GWAS p-value ($p_{(GWAS)}$), previous associations with cancer according to the European Bioinformatics Institute GWAS Catalogue and functional relevance score as determined by RegulomeDB.

and three variants had a score of 7. In the context of colonic tissue, ten of the thirteen (76.9%) credible set variants were associated with regions of strong transcription in either the rectal mucosa, large intestine, colonic mucosa or sigmoid colon (rs62331067, rs1391441, rs7663401, rs2007403, rs2047409, rs11735256, rs7655890, rs988984, rs1391439 and rs1391438). Furthermore, two variants, rs2047409 and rs11735256, were associated with enhancer elements in the large intestine.

In order to further investigate this overlap with regulatory features, the locations of promoters, enhancers, promoter-flanking regions, *CTCF*-binding sites, regions of open chromatin and transcription factor binding sites were downloaded for GRCh37 from Ensembl BioMart for the large intestine, sigmoid colon and the CRC cell line HCT116 (268). These regions were then checked for overlap with the credible set variants using BEDTools (v2.30). Two variants, rs7679673 and rs2047409 displayed overlaps with promoter-flanking regions in all three biological samples, suggesting a functional role in the regulation of gene expression. Surprisingly, there was no overlap of rs2047409 or rs11735256 with enhancer elements in any of the sample types, despite RegulomeDB suggesting otherwise. However, both of these variants were located within 2kb of an enhancer element in the large intestine according to the Ensembl database.

The final functional characterisation carried out on credible set variants was testing for overlap with candidate cis-regulatory elements (cCREs) identified in the SCREEN database (269). cCREs are defined following integration of data for several histone modifications, transcription factor binding sites and DNase binding in several cell types and tissues. From the gathered data, for comparisons to be made between studies and tissue types, z-score normalisation was performed for each annotation within a cCRE – with a z-score of -10 indicating no enrichment for a feature. In order to be classified as highly enriched, the z-score of a feature must lie above the 95th percentile of enrichment for that feature, represented by a z-score of 1.64. When each credible set variant was searched within V2 of the SCREEN database, four of the thirteen variants were within a cCRE (30.8%) and a further seven were within 2kb of a cCRE (53.8%). The functional annotations of each of these cCREs in the context of colonic tissues are presented in Table 2.9. Of the four variants that were located within a cCRE, the cCRE EH38E2317466, associated with rs2047409, showed significant enrichment of DNase binding in a colonic epithelial cell line. Of the cCREs within 2kb of a credible set variant, EH38E2317461, within 2kb of rs7663401, showed significant *CTCF* enrichment in the CRC cell line CACO2. Similarly, EH38E2317467 showed significant DNase enrichment in a colonic epithelial cell line and EH38E2317464 showed significant H3K27ac enrichment within the rectal mucosa. In addition to these features, the SCREEN database provides information on several other histone modifications within multiple tissues and cell types. Of the variants within cCREs, EH38E2317466 showed enrichment of H3K4me^{1} in HCT116 cells and rectal mucosa. Other cCREs associated with H3K4me^{1} included EH38E2317460, EH38E2317461 (both in the sigmoid colon), EH38E2317467 (HCT116 cells) and EH38E2317468 (large intestine).

Overall, the use of publically-available functional annotation data, such as RegulomeDB, SCREEN and Ensembl alongside previously published GWAS associations has provided evidence that the variants presented in the 95% credible set following PAINTOR fine-mapping potentially are of functional importance in colonic tissues. Variants within this credible set have been shown to overlap with regulatory elements including promoter-

Credible Set Variants within cCRE:							
Sample #:	**cCRE ID:**	**cCRE Start (hg38):**	**cCRE End (hg38):**	**Associated SNP:**	**Enriched Feature:**	**Z-Score:**	**Tissue / Cell Type:**
1	EH38E2317330	chr4:104,964,824	chr4:104,965,104	rs62331067	None	NA	NA
2	EH38E2317409	chr4:105,140,373	chr4:105,140,652	rs7679673	None	NA	NA
3	EH38E2317463	chr4:105,207,420	chr4:105,207,717	rs1391441	None	NA	NA
4	EH38E2317466	chr4:105,215,624	chr4:105,215,973	rs2047409	DNase	2.01	Colon Epithelial Cell Line
Credible Set Variants within 2kb of cCRE:							
Sample #:	**cCRE ID:**	**cCRE Start (hg38):**	**cCRE End (hg38):**	**Associated SNP:**	**Enriched Feature:**	**Z-Score:**	**Tissue / Cell Type:**
1	EH38E2317327	chr4:104,948,181	chr4:104,948,392	rs7655284	None	NA	NA
2	EH38E2317328	chr4:104,955,536	chr4:104,955,749	rs11320693	None	NA	NA
3	EH38E2317460	chr4:105,206,035	chr4:105,206,326	rs7663401	None	NA	NA
4	EH38E2317461	chr4:105,206,328	chr4:105,206,513	rs7663401	*CTCF*	1.78	CACO2
5	EH38E2317462	chr4:105,206,527	chr4:105,207,004	rs7663401	None	NA	NA
6	EH38E2317463	chr4:105,207,420	chr4:105,207,717	rs7663401	None	NA	NA
7	EH38E2317464	chr4:105,210,947	chr4:105,211,297	rs2007403	H3K27ac	1.70	Rectal Mucosa
8	EH38E2317466	chr4:105,215,624	chr4:105,215,973	rs11735256	DNase	2.01	Colon Epithelial Cell Line
9	EH38E2317467	chr4:105,215,979	chr4:105,216,280	rs11735256	DNase	1.92	Colon Epithelial Cell Line
10	EH38E2317468	chr4:105,217,179	chr4:105,217,380	rs11735256	None	NA	NA
11	EH38E2317468	chr4:105,217,179	chr4:105,217,380	rs7655890	None	NA	NA
12	EH38E2317469	chr4:105,224,247	chr4:105,224,558	rs9884984	None	NA	NA

Table 2.9 – Credible Set Variants Associated with cCREs: An assessment of variants within the 95% credible set of Probabilistic Annotation Integrator (PAINTOR) fine-mapping of colorectal cancer Genome-Wide Association Study (GWAS) meta-analysis data. Each variant was searched within the Search Candidate cis-Regulatory Elements by ENCODE (SCREEN) database for overlap with Candidate cis-Regulatory Elements (cCREs). Shown is the cCRE identifier (cCRE ID), the chromosomal co-ordinates of a cCRE (hg38), variants that lie within a cCRE or within 2 kilobases of a cCRE (Associated SNP) and regulatory features enriched within colonic tissues. Also shown is the associated z-score of the enriched feature, where a z-score of greater than 1.64 denotes the 95th percentile of enrichment. NA = Not Applicable.

flanking regions, enhancer elements and regions of strong transcription in colonic tissues and cCREs with functional enrichment in colonic tissues – including DNase binding and pro-transcription histone modifications.

2.3.4 – Trans-Ethnic Fine-Mapping as an Additional Approach to Identify Candidate Causal Variants

While many GWAS use populations of largely European descent to identify variants significantly associated with a trait or disease, analysis of other populations represent an opportunity to further validate the results seen in studies of European populations. To this end, a meta-analysis of 21,731 CRC cases and 47,444 controls across fourteen cohorts of East-Asian (EAS) descent (see Table 2.2) was performed on the 4q24 locus of chromosome 4. A total of 6,308 variants were included in the analysis of the same region as was studied in European populations. Interestingly, there were no variants significantly associated with CRC predisposition at this locus in EAS populations (see Figure 2.10). Variants that were significantly associated with CRC risk in European populations had non-significant associations with CRC in EAS populations. Examples of these included rs7655284 ($p_{(Europeans)} = 5.44 \times 10^{-11}$) and rs7679673 ($p_{(Europeans)} = 2.14 \times 10^{-14}$), which represented the lead variants of the two independent GWAS signals identified at this locus in European populations (see section 2.3.1). In EAS populations, these variants presented with $p_{(EAS)}$ of 0.00254 and 0.663202 respectively.

The lack of variants significantly associated with CRC predisposition in EAS populations is unlikely to be a consequence of reduced power compared to the meta-analysis in Europeans or locus-specific effects and therefore represents an opportunity for an additional fine-mapping strategy to identify the candidate causal variant(s) underlying the enhanced CRC risk in Europeans. It is possible that the causal variant(s) in Europeans show very little genotypic variation in EAS populations, consequently leading to a lack of variants associated with CRC at genome-wide significance in this population. The minor allele frequencies (MAF) of the European lead variants rs7655284 and rs7679673 remained fairly consistent across European and EAS populations. For rs7655284, the $MAF_{(Europeans)}$ was approximately 30.2%, whereas the $MAF_{(EAS)}$ was 48.4%. The $MAF_{(Europeans)}$ of rs7679673 was approximately 38.9% and the $MAF_{(EAS)}$ of this variant was estimated to be 19.4%. Therefore, it is unlikely that these variants are the causal variants underlying the GWAS association with CRC at this locus, since these variants also show genotypic variation in EAS populations.

Therefore, trans-ethnic fine-mapping represents an alternative strategy to identify the causal variant(s) underlying these GWAS signals in European populations. In order to achieve this, the variants rs7655284 ($p_{(GWAS)} = 5.44 \times 10^{-11}$), rs17035289 ($p_{(GWAS)} = 2.19 \times 10^{-11}$) and rs1391441 ($p_{(GWAS)} = 3.97 \times 10^{-14}$) were selected as sentinel variants for the European signals. The latter two were selected instead of rs7679673 as both had previously been associated with CRC in European populations and would allow the assessement of more candidate causal vairants than rs7679673 alone (199,240). All sentinel variant proxies with an $r^2 > 0.2$ in European populations were considered to be potentially causal in Europeans due to their

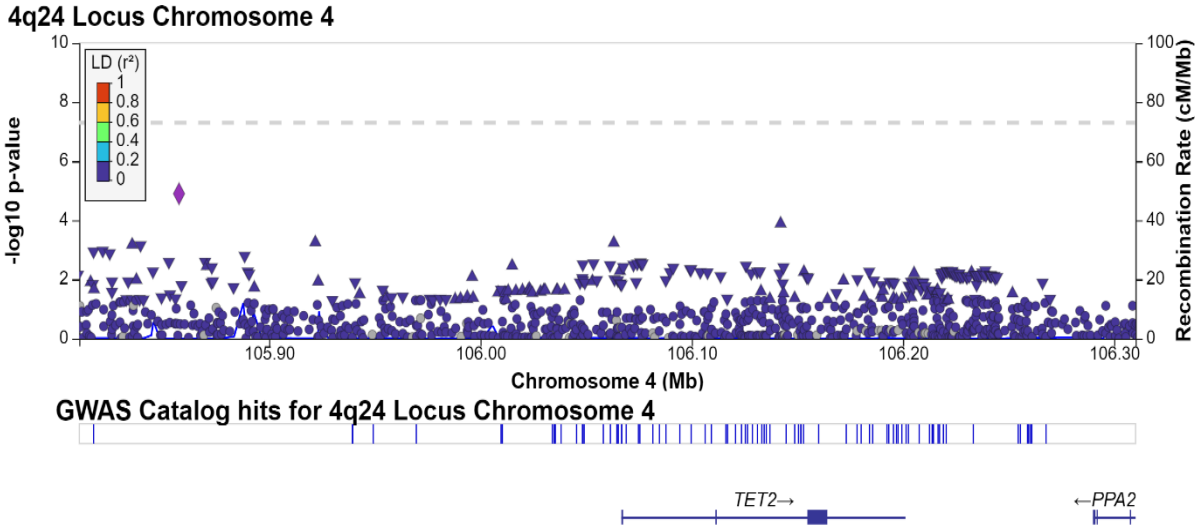


Figure 2.10 – East-Asian CRC GWAS Meta-Analysis of the 4q24 Locus of Chromosome 4: A Locus Zoom plot of colorectal cancer Genome-Wide Association study (GWAS) meta-analysis of individuals of East-Asian descent. Shown is a 500 kilobase (kb) region of the 4q24 locus of chromosome 4. The Y-axis shows the $-\log_{10}$ GWAS p-value of each SNP at the locus. The colour of each point on the plot represents linkage disequilibrium (LD) measurements in East-Asian populations of a variant relative to the lead SNP. The dashed line indicates a $p_{(GWAS)}$ of 5×10^{-8}, used in this study to indicate genome-wide significance. Plot generated using LocusZoom (https://my.locuszoom.org).

LD with one or more of the sentinel variants. This resulted in 22 proxies of rs7655284, 71 proxies of rs17035289 and 286 proxies of rs1391441 being included in downstream analysis in EAS populations. Of these proxy variants, sixty-five presented with a $MAF_{(EAS)}$ and an r^2 with their associated sentinel variant in EAS populations of < 0.1. These variants were therefore considered as variants that showed limited variability in EAS populations and were therefore candidate causal variants underlying the association with CRC in European populations. The details of these candidate causal variants are presented in Table 2.10.

For the rs7655284 region, a total of three candidate causal variants were identified. One of these variants, rs35242239, was not present in the meta-analysis of European populations and therefore could not be the causal variant at this locus. This left two candidate causal variants for this signal. One of these variants, rs35851974, was caused by an insertion/deletion (INDEL) and was not present in 1,000 Genomes Project reference data. Therefore, it was not possible to perform conditional analysis on these candidate causal variants. Of these two candidate causal variants, rs35851974 and rs71599032 had RegulomeDB scores of 6 and 5 respectively. Similarly to the sentinel variant rs7655284, neither variant had previously been associated with any disease according to the GWAS Catalogue (226).

For the rs7679673 locus, a total of forty-seven unique candidate causal variants were identified (fifteen variants were proxies for both rs17035289 and rs1391441). Of these variants, rs17035310 had the strongest association with CRC in the European meta-analysis ($p_{(GWAS)} = 2.11 \times 10^{-11}$). This variant could then be used for conditional analysis of the other candidate causal variants of the rs7679673 signal. As seen in Table 2.11, when conditioned on rs17035310 no other candidate causal variant presented with a significant $p_{(conditional)}$,

rs7655284 Proxies:								
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**$MAF_{(Europeans)}$:**	**$MAF_{(EAS)}$:**	**r^2 with rs7655284 (Europeans):**	**r^2 with rs7655284 (EAS):**
1	rs35242239	4	105,861,835	NA	0.146	NA	0.3488	NA
2	rs35851974	4	105,872,988	7.03×10^{-6}	0.2054	NA	0.5671	NA
3	rs71599032	4	105,883,814	0.00294	0.0965	0.007	0.247	0.007
rs17035289 Proxies:								
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**$MAF_{(Europeans)}$:**	**$MAF_{(EAS)}$:**	**r^2 with rs17035289 (Europeans):**	**r^2 with rs17035289 (EAS):**
1	rs141020046	4	106,044,102	0.0045	0.2116	0.0853	0.2909	0.0136
2	rs141714386	4	106,054,972	5.47×10^{-4}	0.0458	0.0139	0.229	0.0714
3	rs57251748	4	106,058,185	4.31×10^{-7}	0.1522	0.0139	0.8568	0.0714
4	rs201770780	4	106,064,367	0.00928	0.0458	NA	0.229	NA
5	rs17035310	4	106,064,754	2.11×10^{-11}	0.1337	0.0139	0.7362	0.0714
6	rs763480	4	106,071,597	4.98×10^{-11}	0.1337	0.002	0.7362	0.002
7	rs11728350	4	106,078,097	1.74×10^{-10}	0.1337	0.001	0.7362	0.005
8	rs34316731	4	106,078,344	3.34×10^{-6}	0.1337	0.001	0.7362	0.005
9	rs6825684	4	106,084,643	1.64×10^{-10}	0.1337	0.001	0.7362	0.005
10	rs115794885	4	106,094,414	2.85×10^{-4}	0.0545	0	0.2454	NA
11	rs138904640	4	106,101,620	4.72×10^{-5}	0.1324	0.047	0.6483	0.007
12	rs76074589	4	106,101,644	0.00118	0.0483	0.047	0.2129	0.007
13	rs116001054	4	106,107,109	1.07×10^{-9}	0.1324	0.057	0.6483	0.009
14	rs17508261	4	106,108,902	8.48×10^{-10}	0.1324	0.057	0.6483	0.009
15	rs17430251	4	106,115,708	8.30×10^{-8}	0.1324	0.047	0.6483	0.007
16	rs78632895	4	106,123,638	1.80×10^{-7}	0.1324	0.047	0.6639	0.007
17	rs75321784	4	106,125,022	9.47×10^{-10}	0.1312	0.047	0.6562	0.007
18	rs113887651	4	106,125,427	0.00262	0.2302	0.06	0.3022	0.008
19	rs143875052	4	106,150,555	2.08×10^{-10}	0.1312	0.056	0.6562	0.009
20	rs75625682	4	106,152,037	1.85×10^{-10}	0.1312	0.047	0.6562	0.007
21	rs17253672	4	106,156,187	6.83×10^{-6}	0.0619	0	0.2707	NA
22	rs114672787	4	106,159,582	6.75×10^{-6}	0.0656	0	0.2907	NA
23	rs56185013	4	106,160,133	2.84×10^{-4}	0.198	0.047	0.3715	0.007
24	rs201330646	4	106,169,706	0.00395	0.0507	0	0.2258	NA
25	rs4464576	4	106,177,882	2.57×10^{-4}	0.198	0.047	0.3715	0.007
26	rs1498126	4	106,192.457	2.35×10^{-4}	0.2054	0.047	0.3514	0.007
27	rs1498125	4	106,192,563	2.27×10^{-4}	0.2042	0.047	0.3546	0.007
28	rs34402524	4	106,196,829	2.04×10^{-10}	0.1374	0.047	0.6181	0.007
29	rs140527567	4	106,204,863	4.93×10^{-4}	0.0495	0	0.2337	NA
30	rs114358140	4	106,209,264	4.68×10^{-4}	0.0495	0	0.2337	NA

Variant #:	SNP ID:	Chromosome #:	Position (hg19):	$P_{(GWAS)}$:	$MAF_{(Europeans)}$:	$MAF_{(EAS)}$:	r^2 with rs17035289 (Europeans):	r^2 with rs17035289 (EAS):
31	rs78280128	4	106,264,781	5.17×10^{-7}	0.0866	0	0.2042	NA
32	rs78479210	4	106,270,017	3.21×10^{-8}	0.0866	0	0.2042	NA
33	rs77994146	4	106,297,819	1.07×10^{-7}	0.0866	0	0.2042	NA
rs1391441 Proxies:								
Variant #:	**SNP ID:**	**Chromosome #:**	**Position (hg19):**	**$P_{(GWAS)}$:**	**$MAF_{(Europeans)}$:**	**$MAF_{(EAS)}$:**	**r^2 with rs1391441 (Europeans):**	**r^2 with rs1391441 (EAS):**
1	rs12509636	4	106,010,433	2.79×10^{-4}	0.3106	0.0784	0.208	0.0529
2	rs62331120	4	106,017,299	3.64×10^{-4}	0.3144	0.0774	0.2068	0.0551
3	rs72617743	4	106,025,439	2.10×10^{-4}	0.3094	0.0764	0.205	0.0538
4	rs56883672	4	106,026,989	3.74×10^{-4}	0.3144	0.0764	0.2068	0.0538
5	rs17501550	4	106,027,829	5.03×10^{-4}	0.3168	0.0774	0.2077	0.0551
6	rs17429675	4	106,028,073	0.00129	0.3168	0.0774	0.2077	0.0551
7	rs17429682	4	106,028,823	0.00139	0.3168	0.0774	0.2077	0.0551
8	rs62331124	4	106,034,930	6.10×10^{-4}	0.3156	0.0764	0.2099	0.0538
9	rs17429745	4	106,038,169	4.77×10^{-4}	0.3144	0.0774	0.2068	0.0551
10	rs57251748	4	106,058,185	4.31×10^{-7}	0.1522	0.0139	0.2061	0.0127
11	rs17035310	4	106,064,754	2.11×10^{-11}	0.1337	0.0139	0.2597	0.0127
12	rs763480	4	106,071,597	4.98×10^{-11}	0.1337	0.002	0.2597	0
13	rs11728350	4	106,078,097	1.74×10^{-10}	0.1337	0.001	0.2597	0.001
14	rs34316731	4	106,078,344	3.34×10^{-6}	0.1337	0.001	0.2597	0.001
15	rs6825684	4	106,084,643	1.64×10^{-10}	0.1337	0.001	0.2597	0.001
16	rs138904640	4	106,101.620	4.72×10^{-5}	0.1324	0.0466	0.3059	0.0409
17	rs116001054	4	106,107,109	1.07×10^{-9}	0.1324	0.0565	0.3059	0.0214
18	rs17508261	4	106,108,902	8.48×10^{-10}	0.1324	0.0565	0.3059	0.0214
19	rs17430251	4	106,115,708	8.30×10^{-8}	0.1324	0.0466	0.3059	0.0409
20	rs78632895	4	106,123,638	1.80×10^{-7}	0.1324	0.0466	0.3059	0.0409
21	rs75321784	4	106,125,022	9.47×10^{-10}	0.1312	0.0466	0.3111	0.0409
22	rs143875052	4	106,150,555	2.08×10^{-10}	0.1312	0.0556	0.3111	0.0206
23	rs75625682	4	106,152,037	1.85×10^{-10}	0.1312	0.0466	0.3111	0.0409
24	rs34402524	4	106,196,829	2.04×10^{-10}	0.1374	0.0466	0.3282	0.0409
25	rs114832990	4	106,219,656	0.02708	0.1188	0	0.2445	NA
26	rs6419170	4	106,230,888	0.00383	0.3243	0.0665	0.4014	0.0117
27	rs10446713	4	106,233,714	1.19×10^{-7}	0.3267	0.0665	0.3952	0.0117
28	rs375012980	4	106,243,895	NA	0.2054	0	0.4593	NA
29	rs17321073	4	106,258,989	1.07×10^{-7}	0.3218	0.0665	0.4078	0.0117

Table 2.10 – Candidate Causal Variants Identified by Trans-Ethnic Fine-Mapping: A list of variants identified as proxy variants of either rs7655284, rs17035289 or rs1391441 in European populations ($r^2 > 0.2$) while simultaneously presenting with a minor allele frequency (MAF) and r^2 of < 0.1 in East-Asian (EAS) populations. Presented are the identifier of the variant, chromosome, genomic co-ordinates (hg19), $p_{(GWAS)}$ from the meta-analysis in Europeans, MAF in either European or EAS populations and the r^2 with either rs7655824, rs17035289 or rs1391441 in European and EAS populations. NA = Not Applicable. GWAS = Genome-Wide Association study.

indicating that the causal variant(s) at this locus were associated with a single GWAS signal. Of the forty-seven candidate causal variants, five presented with a RegulomeDB score of 3a and one variant presented with a score of 2b (see Table 2.11). In addition to this, three candidate causal variants had been previously reported in the GWAS Catalogue to be associated with cancer predisposition in European populations. rs57251748 ($p_{(GWAS)} = 4.31 \times 10^{-7}$) has been previously associated with breast, ovarian and prostate cancers, rs17035310 ($p_{(GWAS)} = 2.11 \times 10^{-11}$) has been previously associated with CRC and rs34402524 ($p_{(GWAS)} = 2.04 \times 10^{-10}$) has been associated with prostate cancer (277–279). Of the candidate causal variants with a RegulomeDB score of 3a or 2b, rs17035310, rs11728350 and rs34402524 were associated with strong transcription or enhancer elements in colonic tissues, indicating that these variants may be of functional relevance in these tissues. In addition to this, rs17035310 was shown to overlap with the cCRE EH38E2317413 – further indicating its potential functional relevance. Interestingly, rs34402524 represents a missense variant in the *TET2* gene ($TET2^{L1721W}$), which was predicted to be deleterious according to the Sorting Intolerant from Tolerant variant effect predictor and as possibly damaging according to the Polymorphism Phenotyping tool.

Overall, the significant associations between rs17035310, rs34402524 and CRC predisposition in the European meta-analysis and lack of genotypic variation in EAS populations, coupled with the functional annotation data and previously reported associations with cancer, suggest that these variants were the likely causal variants underlying the association with CRC at the rs7679673 signal in Europeans. These variants can be combined with the two candidate causal variants from the rs7655284 signal, rs35851974 and rs71599032, to produce a list of four candidate causal variants across both GWAS signals.

2.3.5 – TWMR Identifies *TET2* as the Candidate Causal Gene Underlying Colorectal Cancer Predisposition

As identified in sections 2.3.1 – 2.3.4, there appear to be multiple independent GWAS signals associated with CRC at the 4q24 locus of chromosome 4. Fine-mapping analysis and *in silico* functional annotation reveals that there were a number of candidate casual variants with potential functional relevance in the colon that may underpin each of these GWAS signals. However, the identity of the candidate causal gene(s) that are associated with the enhanced CRC risk remains uncertain. Therefore, GWAS meta-analysis data was integrated with All SNP-Gene association eQTL data from GTEx (v8) in the transverse colon. GTEx eQTL data contains a quantitative estimate of the effect size of a variant on the expression of nearby genes, which can be combined with the effect size of the variant on the CRC phenotype, provided by GWAS data, and used as an input for MR analysis to predict the likely candidate causal gene underlying CRC predisposition at the 4q24 locus of chromosome 4.

To this end, TWMR analysis was performed on all genes located within the fine-mapping region where significant eQTLs were present ($p_{(eQTL)} < 0.005$), which was a total of thirteen genes. In order to implement TWMR at this locus, each gene was selected in turn and significant eQTLs were identified for each gene. Conditional analysis was then performed on

Variant #:	SNP ID:	Chromosome #:	Position (hg19):	$P_{(GWAS)}$:	$P_{(Conditional)}$:	RegulomeDB Score:	Previously Reported?
1	rs12509636	4	106,010,433	2.79×10^{-4}	0.0194	7	No
2	rs62331120	4	106,017,299	3.64×10^{-4}	0.0231	6	No
3	rs72617743	4	106,025,439	2.10×10^{-4}	0.0165	7	No
4	rs56883672	4	106,026,989	3.74×10^{-4}	0.0236	6	No
5	rs17501550	4	106,027,829	5.03×10^{-4}	0.0337	2b	No
6	rs17429675	4	106,028,073	0.00129	0.0623	4	No
7	rs17429682	4	106,028,823	0.00139	0.0654	4	No
8	rs62331124	4	106,034,930	6.10×10^{-4}	0.035	5	No
9	rs17429745	4	106,038,169	4.77×10^{-4}	0.0272	4	No
10	rs141020046	4	106,044,102	0.0045	0.5488	6	No
11	rs141714386	4	106,054,972	5.47×10^{-4}	0.9511	6	No
12	rs57251748	4	106,058,185	4.31×10^{-7}	0.0608	5	Yes
13	rs201770780	4	106,064,367	0.00928	0.5559	4	No
14	rs17035310	4	106,064,754	2.11×10^{-11}	NA	3a	Yes
15	rs763480	4	106,071,597	4.98×10^{-11}	NA	5	No
16	rs11728350	4	106,078,097	1.74×10^{-10}	NA	3a	No
17	rs34316731	4	106,078,344	3.34×10^{-6}	NA	6	No
18	rs6825684	4	106,084,643	1.64×10^{-10}	NA	7	No
19	rs115794885	4	106,094,414	2.85×10^{-4}	0.9606	6	No
20	rs138904640	4	106,101,620	4.72×10^{-5}	0.1924	5	No
21	rs76074589	4	106,101,644	0.00118	0.9769	6	No
22	rs116001054	4	106,107,109	1.07×10^{-9}	0.9539	6	No
23	rs17508261	4	106,108,902	8.48×10^{-10}	0.9253	6	No
24	rs17430251	4	106,115,708	8.30×10^{-8}	0.5054	4	No
25	rs78632895	4	106,123,638	1.80×10^{-7}	0.4355	6	No
26	rs75321784	4	106,125,022	9.47×10^{-10}	0.9174	7	No
27	rs113887651	4	106,125,427	0.00262	0.2037	7	No
28	rs143875052	4	106,150,555	2.08×10^{-10}	0.9484	7	No
29	rs75625682	4	106,152,037	1.85×10^{-10}	0.9394	6	No
30	rs17253672	4	106,156,187	6.83×10^{-6}	0.6821	7	No
31	rs114672787	4	106,159,582	6.75×10^{-6}	0.6821	5	No
32	rs56185013	4	106,160,133	2.83×10^{-4}	0.1714	5	No
33	rs201330646	4	106,169,706	0.00395	0.3034	5	No
34	rs4464576	4	106,177,882	2.57×10^{-4}	0.1791	7	No
35	rs1498126	4	106,192,457	2.35×10^{-4}	0.2199	5	No
36	rs1498125	4	106,192,563	2.27×10^{-4}	0.2234	5	No

Variant #:	SNP ID:	Chromosome #:	Position (hg19):	$P_{(GWAS)}$:	$P_{(Conditional)}$:	RegulomeDB Score:	Previously Reported?
37	rs34402524	4	106,196,829	2.04×10^{-10}	0.8354	3a	Yes
38	rs140527567	4	106,204,863	4.93×10^{-4}	0.844	5	No
39	rs114358140	4	106,209,264	4.68×10^{-4}	0.8561	3a	No
40	rs114832990	4	106,219,656	0.02708	0.00631	6	No
41	rs6419170	4	106,230,888	0.00383	0.2635	7	No
42	rs10446713	4	106,233,714	1.19×10^{-7}	0.0493	5	No
43	rs375012980	4	106,243,895	NA	NA	7	No
44	rs17321073	4	106,258,989	1.07×10^{-7}	0.053	6	No
45	rs78280128	4	106,264,781	5.17×10^{-7}	0.1178	3a	No
46	rs78479210	4	106,270,017	3.21×10^{-8}	0.0558	4	No
47	rs77994146	4	106,297,819	1.07×10^{-7}	0.0821	7	No

Table 2.11 – Conditional Analysis on rs17035310: Conditional analysis of candidate causal variants identified by trans-ethnic fine-mapping analysis of the 4q24 locus of chromosome 4. Variants were conditioned on the most significant candidate causal variant (rs17035310 – $p_{(GWAS)} = 2.11 \times 10^{-11}$) using linkage disequilibrium estimates from European individuals from the 1,000 Genomes Project data. Columns include the unique identifier of the variant (SNP ID), chromosome number, genome co-ordinates (hg19), GWAS p-value ($P_{(GWAS)}$), conditional p-value ($P_{(conditional)}$), the functional annotation score of a variant according to RegulomeDB and whether the variant has been previously reported to predispose European individuals to cancer development according to the GWAS Catalogue. NA = Not Applicable.

these eQTLs using GCTA and reference LD data from the 1,000 Genomes Project, conditioning on the most significant eQTL for the gene being investigated. In order to identify conditionally independent eQTLs, a $p_{(conditional)}$ threshold was set at 0.05. Following the identification of independent eQTLs, other genes for which these independent eQTLs were shared were added to the model, alongside the independent eQTLs associated with these additional genes. Finally, eQTLs were pruned to only include variants with an r^2 of < 0.1, this was done using PLINK (v1.90p) in European populations. See Figure 2.3 for a summary of the TWMR approach. A Pearson's correlation matrix between all variants included in the model was also generated using 1,000 Genomes Project data in Europeans. Using this approach, each candidate causal gene at the locus could be assessed in turn, the results of this are presented in Table 2.12.

As seen in Table 2.12, of the thirteen genes investigated at this locus, only three displayed a $p_{(TWMR)} < 0.05$. These genes were *TET2* ($p_{(TWMR)} = 0.0072462$) – a member of the TET family of proteins involved in DNA de-methylation (245), *AC105391.1* ($p_{(TWMR)} = 1.048 \times 10^{-9}$) – a pseudogene and *TET2-AS1* ($p_{TWMR)} = 0.040044535$) – an antisense transcript of the *TET2* gene. Of the three genes with the significant gene-trait associations as determined by TWMR, it seems likely that the candidate causal gene underlying the GWAS signals at this locus is *TET2,* given the extensively-documented role of DNA methylation in CRC (as discussed in Chapter I of this thesis). TWMR analysis also provides a quantitative estimate of causality for a gene in the form of an α statistic. For *TET2,* this α statistic was -0.006591237, implying that a reduction in *TET2* expression in the transverse colon may potentially drive its association with CRC development. However, the significant gene-trait association for *TET2* in the transverse colon does not mean that there are no other potential target tissues where altered *TET2* expression may influence CRC risk. For example, it is conceivable that altered expression of a gene in the blood may indirectly drive the development of CRC. Therefore, TWMR analysis was repeated for the *TET2* gene using GTEx (v8) All SNP-Gene association eQTL data for the breast, prostate, whole blood and EBV-transformed lymphocytes to identify additional target tissues associated with CRC risk. These tissues were chosen as the 4q24 locus of chromosome 4 has previously been implicated in breast cancer, the lead variant in the CRC GWAS meta-analysis has previously been associated with prostate cancer and *TET2* has previously been reported to be mutated in leukaemia (280,281). As seen in Table 2.13, there were no significant gene-trait associations between *TET2* expression and CRC in any of the tissues investigated. This implies that the association with CRC at the 4q24 locus may be driven by reduced *TET2* expression in colonic tissues only.

In order to further investigate *TET2* as the candidate causal gene at this locus, variants from the 95% credible set identified by PAINTOR fine-mapping analysis and candidate causal variants identified by trans-ethnic fine-mapping analysis (see section 2.3.4) were searched for their effect on *TET2* expression in the transverse colon according to GTEx (v8) All SNP-Gene association data. The results of this are presented in Table 2.14. Of the seventeen candidate causal variants, fifteen had available eQTL data for the *TET2* gene in the transverse colon. The effect allele of most candidate causal variants appears to suggest a reduction in *TET2* expression increases CRC risk, or conversely an increase in *TET2* expression reduces CRC risk. Of the fifteen variants, rs71599032 had a significant effect on *TET2* expression ($p_{(eQTL)} = 0.008$) and rs7655284 had a nominally significant effect ($p_{(eQTL)} = 0.029$). Four other variants had $p_{(eQTL)}$ approaching nominal significance ($p_{(eQTL)} < 0.1$). Overall, this data

Gene #:	Ensembl ID:	Gene Symbol:	# SNPs:	# Genes:	α:	SE:	Z-Score:	$P_{(TWMR)}$:
1	ENSG00000138777.19	*PPA2*	4	1	0.001476626	0.002577199	0.57295772	0.566673322
2	ENSG00000138780.14	*GSTCD*	2	1	-0.003171686	0.00356736	-0.899085109	0.373957341
3	ENSG00000138785.14	*INTS12*	3	1	0.004764608	0.003719343	1.28103506	0.200181352
4	ENSG00000145348.16	*TBCK*	4	2	-0.000935245	0.001948835	-0.479899414	0.631298918
5	ENSG00000168743.12	*NPNT*	3	1	-0.000409286	0.002355193	-0.17378022	0.86203819
6	ENSG00000168769.13	*TET2*	3	1	-0.006591237	0.002454551	-2.685312583	0.0072462
7	ENSG00000168772.10	*CXXC4*	3	1	-0.006552217	0.003479158	-1.883276339	0.059662928
8	ENSG00000236699.8	*ARHGEF38*	5	1	0.000135679	0.001377015	0.098531314	0.921510408
9	ENSG00000248778.1	*AC105391.1*	4	1	-0.067315013	0.011031775	-6.101920629	1.048×10^{-9}
10	ENSG00000249264.1	*EEF1AP9*	4	2	0.000958767	0.002844921	0.337010093	0.736109292
11	ENSG00000250740.1	*AC105391.2*	1	1	0.000427277	0.006482193	0.065915541	0.947445067
12	ENSG00000251259.1	*AC004069.1*	5	1	0.002129052	0.00619239	1.314847132	0.188561284
13	ENSG00000251586.1	*TET2-AS1*	3	1	-0.003668563	0.001786676	-2.05328923	0.040044535

Table 2.12 – TWMR Analysis of the 4q24 Locus of Chromosome 4: The results of Transcriptome-Wide Mendelian Randomisation (TWMR) analysis of the 4q24 Locus of chromosome 4. Expression quantitative trait loci (eQTL) data from the transverse colon was integrated with Genome-Wide Association Study (GWAS) meta-analysis data for colorectal cancer in thirteen genes at the locus. Included in the table are the Ensembl ID of the gene, its symbol, the number of SNPs (# SNPs) and genes (# Genes) included in the TWMR model for the gene, the quantitative effect size of causality as estimated by TWMR (α) and its standard error (SE), the z-score for comparisons of effect sizes between genes and the p-value of TWMR analysis ($p_{(TWMR)}$).

GTEx Breast:								
Gene #:	**Ensembl ID:**	**Gene Symbol:**	**# SNPs:**	**# Genes:**	**α:**	**SE:**	**Z-Score:**	**$P_{(TWMR)}$:**
1	ENSG00000168769.13	*TET2*	4	1	-0.00197996	0.002552756	-0.77562	0.4379754
GTEx Prostate:								
Gene #:	**Ensembl ID:**	**Gene Symbol:**	**# SNPs:**	**# Genes:**	**α:**	**SE:**	**Z-Score:**	**$P_{(TWMR)}$:**
1	ENSG00000168769.13	*TET2*	3	1	-0.003344346	0.002541966	-1.315653666	0.1882903
GTEx Whole Blood:								
Gene #:	**Ensembl ID:**	**Gene Symbol:**	**# SNPs:**	**# Genes:**	**α:**	**SE:**	**Z-Score:**	**$P_{(TWMR)}$:**
1	ENSG00000168769.13	*TET2*	5	1	-0.002957112	0.002513849	-1.176328079	0.2394638
GTEx EBV-Transformed Lymphocytes:								
Gene #:	**Ensembl ID:**	**Gene Symbol:**	**# SNPs:**	**# Genes:**	**α:**	**SE:**	**Z-Score:**	**$P_{(TWMR)}$:**
1	ENSG00000168769.13	*TET2*	8	3	0.003635801	0.00225109	1.61512918	0.1062827

Table 2.13– TWMR Analysis of Extra-Colonic Tissues: The results of Transcriptome-Wide Mendelian Randomisation (TWMR) analysis of the 4q24 Locus of chromosome 4. Expression quantitative trait loci (eQTL) data from breast tissue, prostate tissue, whole blood or Epstein-Barr Virus (EBV) transformed lymphocytes was integrated with Genome-Wide Association Study (GWAS) meta-analysis data for colorectal cancer to investigate the ten-eleven translocation 2 (*TET2*) gene. Included in the table are the Ensembl ID of the gene, its symbol, the number of SNPs (# SNPs) and genes (# Genes) included in the TWMR model, the quantitative effect size of causality as estimated by TWMR (α) and its standard error (SE), the z-score of the effect size and the p-value of TWMR analysis ($p_{(TWMR)}$).

Variant #:	SNP ID:	Effect Allele:	$\beta_{(GWAS)}$:	$SE_{(GWAS)}$:	$P_{(GWAS)}$:	*TET2* $\beta_{(eQTL)}$:	*TET2* $SE_{(eQTL)}$:	*TET2* $P_{(eQTL)}$:
1	rs7655284	G	-0.05132	0.007825	5.44×10^{-11}	0.09883	0.04500	0.029
2	rs35851974	C	-0.0771	0.017161	7.03×10^{-6}	NA	NA	NA
3	rs113280693	C	-0.09012	0.072665	0.215	NA	NA	NA
4	rs71599032	T	-0.03681	0.012381	0.00294	0.20261	0.07522	0.008
5	rs62331067	C	-0.06021	0.016294	2.20×10^{-4}	0.08719	0.08464	0.304
6	rs7679673	A	-0.05825	0.007623	2.14×10^{-14}	0.07365	0.04270	0.086
7	rs17035310	T	-0.0729	0.010884	2.11×10^{-11}	-0.06416	0.06226	0.304
8	rs1391441	A	0.05918	0.007826	3.97×10^{-14}	-0.02750	0.04345	0.527
9	rs7663401	T	0.056482	0.007594	1.02×10^{-13}	-0.07207	0.04269	0.093
10	rs2007403	T	0.05707	0.00759	5.52×10^{-14}	-0.07025	0.04285	0.102
11	rs2047409	A	0.56669	0.007583	7.85×10^{-14}	-0.07150	0.04251	0.094
12	rs11735256	C	0.056854	0.007584	6.54×10^{-14}	-0.07150	0.04251	0.094
13	rs7655890	T	0.059104	0.007827	4.31×10^{-14}	-0.02750	0.04345	0.527
14	rs9884984	A	0.075689	0.021481	4.26×10^{-4}	-0.06427	0.04190	0.126
15	rs1391439	A	0.056541	0.007591	9.42×10^{-14}	-0.06919	0.04253	0.105
16	rs1391438	C	0.059126	0.007839	4.62×10^{-14}	-0.03157	0.04379	0.472
17	rs34402524	G	-0.069989	0.011069	2.04×10^{-10}	-0.03985	0.0643	0.536

Table 2.14 – Effect of Credible Set Variants on *TET2* Expression: The effect of candidate causal variants identified by Bayesian or trans-ethnic fine-mapping of Genome-Wide Association Study (GWAS) meta-analysis data for colorectal cancer on the expression of the ten-eleven translocation 2 (*TET2*) gene. Presented are the variant identifier (SNP ID), the effect allele associated with the variant, the effect size of the effect allele according to GWAS data ($\beta_{(GWAS)}$), as well as its standard error ($SE_{(GWAS)}$) and p-value ($p_{(GWAS)}$). Additionally, the effect size of the effect allele on *TET2* expression in the transverse colon ($\beta_{(eQTL)}$), as well as its association standard error ($SE_{(eQTL)}$) and p-value ($p_{(eQTL)}$). NA = Not Applicable.

suggests that *TET2* is potentially the causal gene underlying GWAS associations with CRC at the 4q24 locus of chromosome 4 and the candidate causal variants identified by either Bayesian or trans-ethnic fine-mapping may have the effect of reducing *TET2* expression in the colon, thus potentially elevating CRC risk.

2.3.6 – Genotype-Expression Analysis Reveal Down-Regulation of *TET2* is Significantly Associated with CRC

The TWMR analysis presented above suggests that down-regulation of *TET2* expression is the causal factor underlying CRC predisposition at the 4q24 locus of chromosome 4. When this hypothesis was investigated in a TWAS dataset obtained from Fernandez-Rozadilla *et al., TET2* was the only gene whose expression was associated with CRC risk at the level of the defined Bonferroni threshold ($p_{(TWAS)} < 4.6 \times 10^{-6}$) (257). Significant associations were identified between *TET2* expression and CRC risk in both the colorectum ($p_{(TWAS)} = 7.69 \times 10^{-20}$) and normal gastrointestinal tissues ($p_{(TWAS)} = 1.19 \times 10^{-29}$) (257). However, significant associations were not identified between *TET2* expression and CRC risk in the muscle of the sigmoid colon, mesenchymal tissues and immune cells (257). This data supports what has previously been presented in Tables 2.12 and 2.13, suggesting that altered *TET2* expression in colonic tissues – but not other tissues – is significantly associated with CRC risk. In addition to this, when the TWAS model between *TET2* expression and CRC risk in the colorectum was conditioned on the genotype of the lead GWAS variant at this locus via sMIST, the TWAS association was reduced but still significant ($p_{(conditional)} = 1.46 \times 10^{-7}$), further indicating the presence of multiple independent GWAS signals at this locus – supporting what has been presented in section 2.3.1 (257).

When the genotypes of rs7679673 and rs7655284, the lead SNPs at each of the independent European GWAS signals identified in section 2.3.1, were combined with distal colonic *TET2* gene expression data from 109 European individuals of the INTERMPHEN study, significant correlations were identified between the number of risk alleles and reduced *TET2* expression. As seen in Figure 2.11a, the expression of *TET2* was significantly reduced in individuals who were homozygous for the rs7655284 risk allele (AA) compared to those homozygous for the non-risk allele (GG) ($p = 0.0023$). There was a significant association between the number of risk alleles and *TET2* expression in these individuals according to both the Jonckheere-Terpstra trend test ($p = 0.003$) and regression model ($p = 0.001$). Furthermore, neither age nor sex were significantly correlated with *TET2* expression in the regression model ($p = 0.274$ and $p = 0.103$ respectively). When the same analysis was performed on rs7679673, the expression of *TET2* was also significantly reduced in individuals homozygous for the CRC risk allele (CC) compared to individuals homozygous for the non-risk allele (AA) ($p = 0.011$, Figure 2.11b). While this trend may have been significant in the regression model ($p = 0.032$), the trend between the number of risk alleles and *TET2* expression was non-significant in the Jonckheere-Terpstra trend test, although the trend was nearly significant ($p = 0.08$). Similarly to rs7655284, age and sex were not significantly associated with *TET2* expression ($p = 0.344$ and $p = 0.05$ respectively). When the genotypes of both rs7655284 and rs7679673 were compared with *TET2* expression (Figure 2.11c), *TET2* expression was significantly

reduced in individuals with three or four risk alleles compared to individuals with zero risk alleles ($p = 0.036$ and $p = 0.0077$ respectively). Furthermore, the number of risk alleles was significantly correlated with reduced *TET2* expression in both the Jonckheere-Terpstra trend test ($p = 0.004$) and regression model ($p < 0.0001$), while age and sex were not significantly associated with *TET2* expression ($p = 0.268$ and $p = 0.064$ respectively).

When the same analysis was repeated using the candidate causal variants identified by trans-ethnic fine-mapping analysis (see section 2.3.4), similar results were obtained. Genotype data for the variant rs35851974 was not available from the INTERMPHEN study. Individuals with the CC genotype for rs71599032 presented with lower *TET2* expression than individuals with the TC genotype – however this difference was not significant ($p = 0.13$, Figure 2.12a). This lack of significance may be a consequence of only a few individuals in the population presenting with the TC genotype. Regression analysis also indicated that the number of rs71599032 risk alleles was inversely correlated with the expression of *TET2,* but again this was not significant ($p = 0.299$). Individuals with the CC rs17035310 genotype had reduced *TET2* compared to those with the TT genotype, however this difference was not significant ($p = 0.23$, Figure 2.12b). However, regression analysis and the Jonckheere-Terpstra trend test found a significant negative correlation between *TET2* expression and the number of rs17035310 risk alleles ($p = 0.012$ and $p = 0.025$ respectively). As seen in Figure 2.12c, individuals with the rs34402524 genotype TT had significantly reduced *TET2* expression compared to individuals with the GG genotype ($p = 0.048$). Regression analysis and the Jonckheere-Terpstra trend test also found a significant association between the number of rs34402524 risk alleles and reduced *TET2* expression ($p = 0.004$ and $p = 0.015$ respectively). When the genotypes for the above variants were combined, individuals with a total of six risk alleles presented with significantly reduced *TET2* expression compared to individuals with only two risk alleles ($p = 0.045$, Figure 2.12d). Regression analysis and the Jonckheere-Terpstra trend test identified a significant negative correlation between the total number of risk alleles and *TET2* expression in these individuals ($p = 0.002$).

Overall, the data from the previously published TWAS suggests that altered *TET2* expression in colonic tissues is associated with CRC risk, consistent with what has been reported in section 2.3.5. In addition to this, it appears that risk alleles of the lead GWAS variants for each independent signal at this locus were associated with a reduction in *TET2* gene expression, potentially driving the association with CRC risk. Furthermore, the number of risk alleles of candidate causal variants identified by trans-ethnic fine-mapping analysis also appear to correlate with reduced *TET2* expression in the distal colon – further supporting the notion that *TET2* is the candidate causal gene underlying the association between the 4q24 locus of chromosome 4 and CRC predisposition.

2.4 - Discussion

Over the past twenty years, GWAS have emerged as a powerful tool for identifying novel associations between genetic variants and a trait or disease. In the context of CRC, a total of

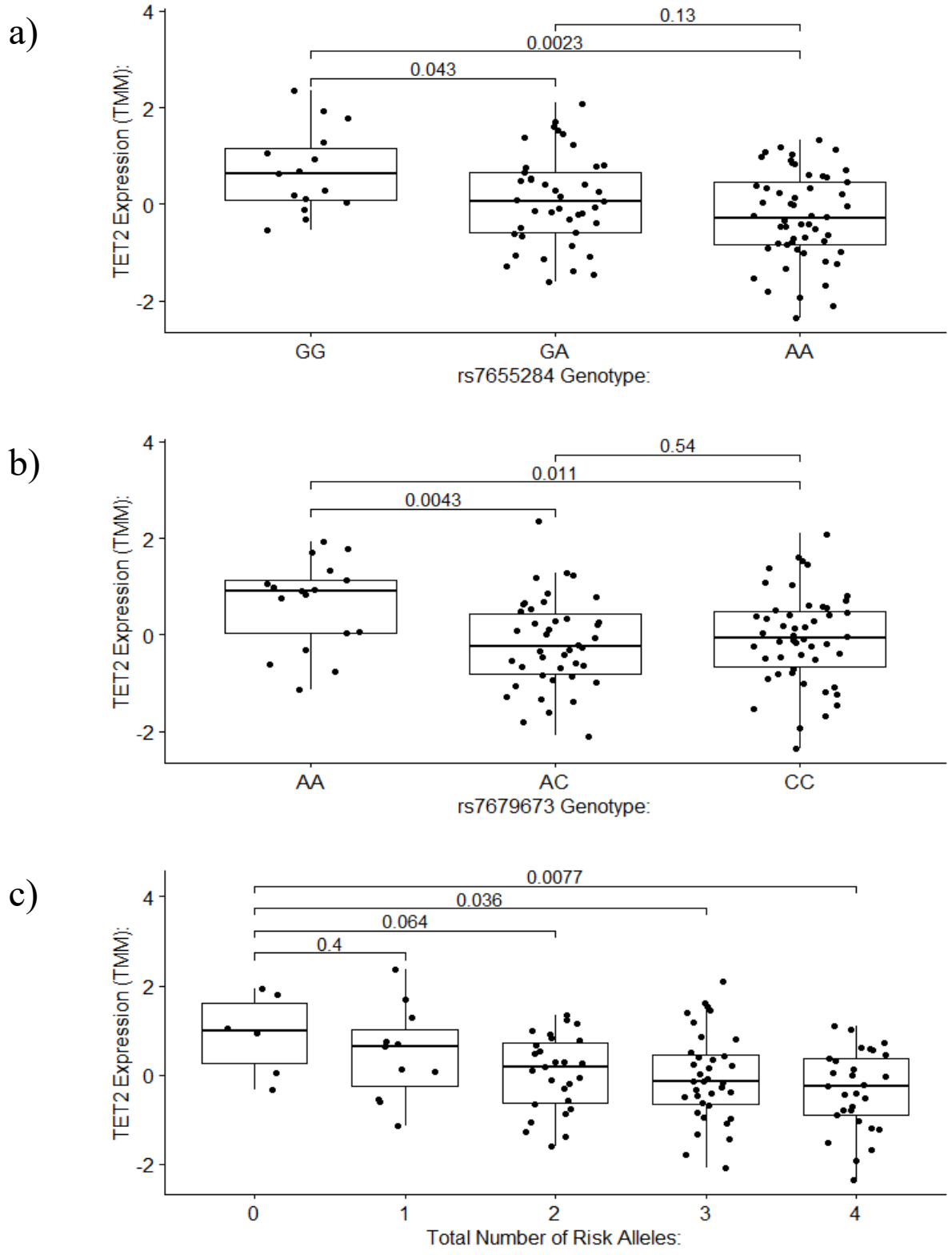


Figure 2.11 – Genotype-Expression Analysis of *TET2* Expression in the Distal Colon: Boxplots indicating the expression of the *TET2* gene in relation to the genotype(s) of variants associated with colorectal cancer predisposition according to a Genome-Wide Association Study meta-analysis in European individuals. Presented are the genotype-expression relationships for rs7655284 (a), rs7679673 (b) and *TET2* gene expression compared to the number of risk alleles of both of these SNPs together (c).

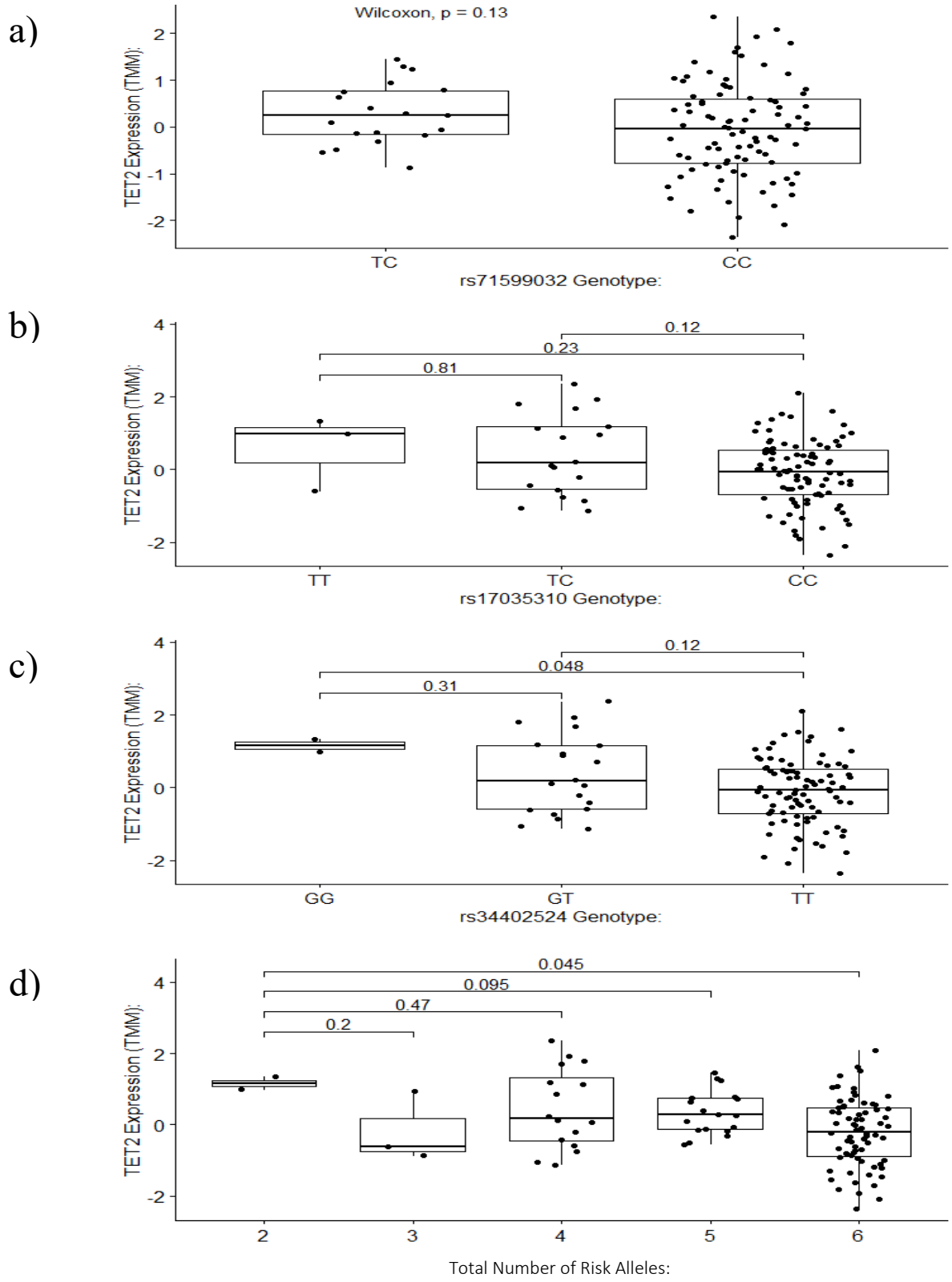


Figure 2.12 – Candidate Causal Variant Effect on *TET2* Expression in the Distal Colon: Boxplots indicating the expression of the *TET2* gene in relation to the genotype(s) of variants associated with colorectal cancer predisposition according to a Genome-Wide Association Study meta-analysis in European individuals. Presented are the genotype-expression relationships for rs71599032 (a), rs17035310 (b), rs34402524 (c) and the *TET2* gene expression compared to the number of risk alleles of all of these SNPs together (d).

205 variants across multiple ethnicities have been significantly associated with disease risk (257). One example of such loci is the 4q24 locus of chromosome 4, characterised by the significant association of rs17035289 with CRC risk according to the meta-analysis by Law *et al.* (234). In this chapter, this locus was more closely investigated in the context of CRC using an expanded GWAS meta-analysis dataset. This dataset included a total of 78,473 cases and 107,143 controls across seventeen cohorts of European descent. The ultimate goal of this chapter was to identify the candidate causal variant(s) at this locus, in order to identify functionally relevant variants which may drive the increased CRC risk seen at this locus. This analysis could also be extended to identify the most likely candidate causal gene, whose altered expression may be the driving force underpinning CRC development.

Conditional analysis was performed on 8,166 variants across a 2.13Mb genomic region of the 4q24 locus of chromosome 4. This locus was characterised by the lead SNP rs7679673, which has previously been associated with prostate cancer but its role in CRC has yet to be fully defined. In addition to this, conditional analysis revealed a second, independent, signal significantly associated with CRC risk. This additional signal was characterised by rs7655284 which, to the best of my knowledge, represents a novel CRC risk variant. Despite the significant association of rs7655284 with CRC ($p_{(GWAS)} = 5.44 \times 10^{-11}$), there are no entries in any of the 6,401 publications in the GWAS Catalogue for this variant in the context of any trait or disease. This is despite this variant having a MAF of 33.5% according to CRC GWAS meta-analysis data. Furthermore, GWAS meta-analysis data identifies rs7655284 in all seventeen cohorts and in all individuals, with an I^2 of 0 and a $p_{(heterogeneity)}$ of 0.611. This indicates that the variant is truly associated with CRC and has no heterogeneity of effect between any of the cohorts included in the meta-analysis. It may also be possible that there were additional independent signals at this locus being captured rs7679673. For example, there are two variants at this locus that have been previously reported as CRC risk variants – including rs17035289 and rs1391441 (234,275). These variants may represent two independent signals captured together by their association with rs7679673 (r^2 with rs7679673 in European populations of 0.224 and 0.69 respectively).

When performing conditional analysis, a $p_{(conditional)}$ threshold of $< 5 \times 10^{-7}$ was set to determine independence of variants from one another. Other studies have used different thresholds, for example the studies by Huyghe *et al.* and Law *et al.* set a more stringent $p_{(conditional)}$ threshold of $< 5 \times 10^{-8}$ whereas other studies, for example the study by Knight *et al.* use a less stringent $p_{(conditional)}$ threshold for independence ($p_{(conditional)} < 1 \times 10^{-6}$) (234,275,282). This makes setting the $p_{(conditional)}$ threshold difficult, when set to $< 5 \times 10^{-7}$, the r^2 of the three variants below this threshold in relation to rs7679673 ranged from 0.001 to 0.047 in European populations, further highlighting the likely independence of the second GWAS signal. It should also be noted that the conditional analysis performed by Huyghe *et al.* and Law *et al.* was on a whole-genome scale, as opposed to a small locus of 8,166 variants. Overall, the indication of the data is that the second independent GWAS signal centred around rs7655284 represents a genuine novel association with CRC at this locus.

Following the identification of a second, novel, GWAS signal significantly associated with CRC at this locus, Bayesian fine-mapping was performed independently on each signal in order to identify the underlying candidate causal variant(s). Following the implementation of PAINTOR, including two separate colon-specific annotation sets and several iterations assuming either one, two or three causal variants at each locus, a total of thirteen candidate

causal variants were identified across both signals. Ten candidate causal variants were associated with the rs7679673 region, whereas only three variants were associated with the rs7655284 region. In the rs7655284 region, two of the three variants, rs113280693 ($p_{(GWAS)}$ = 0.215) and rs62331067 ($p_{(GWAS)} = 2.20 \times 10^{-4}$), identified in the credible sets had non-significant GWAS p-values. As previously discussed, this may be a result of their low frequencies within European populations. Unsurprisingly, most variants present within the 95% credible set of the rs7679673 region were significantly associated with CRC, the only exception being rs9884984 ($p_{(GWAS)} = 4.26 \times 10^{-4}$). This phenomenon may represent a limitation of this analysis, where association statistics between a variant and the phenotype may carry more weight when assigning variants into the credible set than overlaps with functional annotations in the analysis. In addition to this, there also appears to be no way of prioritising functional annotations in terms of importance or functional relevance in the fine-mapping analysis, despite an overlap with exonic regions of the genome potentially carrying more causal weight in an analysis than overlap with other annotations, for example histone modifications. However, PAINTOR still represents an important fine-mapping methodology due to its novelty of including functional annotation data and the ability to assume that a locus contains multiple causal variants.

In addition to PAINTOR, there are other Bayesian fine-mapping methodologies that can account for the possibility of multiple causal variants at each locus of interest. Hormozdiari *et al.* developed Causal Variants Identification in Associated Regions (CAVIAR) which, similarly to PAINTOR, makes use of GWAS association statistics and the LD structure of a locus to determine the likely candidate causal variant(s) underpinning the association with a trait or disease (283). However, the lack of integrated functional annotation data in CAVIAR, however limited the utility of this may be, made PAINTOR seem the best tool to use for fine-mapping analysis. The use of fine-mapping to identify candidate causal variants which underpin significant associations between a SNP and a trait or disease is highly complex, affected by a number of cell, tissue and disease-specific factors which can make definitive results difficult to obtain (284). With the exception of a few tools, including CAVIAR and PAINTOR, many fine-mapping strategies have had their utility hampered by the assumption that there is only one causal variant at a locus, which may contain multiple independent signals and causal variants (284). Fine-mapping is also limited by the GWAS data on which the analysis is being performed. One of the critical assumptions of Bayesian fine-mapping is that all variants that could be causal are included in the analysis – which may not be achievable in rare diseases or where the true causal variant is rare, especially when using smaller-scale GWAS data.

Interestingly, GWAS meta-analysis data from EAS populations showed no significant associations with CRC predisposition at the 4q24 locus of chromosome 4. This was possibly a result of the lower sample size of this meta-analysis compared to the study in European populations. However, this is unlikely given that previous studies have identified significant associations with traits and diseases with GWAS data from a similar number of participants (285). The study by Lu *et al.* performed a meta-analysis on 22,775 cases and 47,731 controls in the context of CRC in EAS populations and identified thirteen novel CRC risk loci at genome-wide significance (285). Therefore, a more likely explanation for the lack of genome-wide significant associations at CRC at this locus was that the causal variant(s) were genetically invariant in EAS populations but show a greater degree of variability in

Europeans and therefore show significant association with CRC risk. For each of the two independent GWAS signals, two candidate causal variants were identified by trans-ethnic fine-mapping analysis. For the rs7655284 signal, these included variants which were in modest LD with rs7655284 and included an INDEL variant as well as a single base substitution. This INDEL variant was not present in 1,000 Genomes Project reference data or subsequent eQTL data, making downstream functional characterisation of this variant difficult. The rs7679673 signal presented with two candidate causal variants, both of which had been associated with cancer development in previous studies (277,278). A strength of this trans-ethnic fine-mapping analysis as opposed to Bayesian fine-mapping methodologies implemented previously includes the lack of bias towards a candidate causal variant based on the association statistic of the GWAS meta-analysis data. This trans-ethnic fine-mapping analysis has also generated a much shorter list of credible variants that are more strongly associated with functional annotations in colonic tissues than the variants identified by Bayesian fine-mapping strategies.

Following the identification of these candidate causal variants via both Bayesian and trans-ethnic fine-mapping strategies, most were revealed to have some functional relevance in colonic tissues as determined by *in silico* functional annotation. Examples of these functional roles include a variant being previously associated with CRC predisposition, lying within enhancer elements or regions of strong transcription in colonic tissues, the presence of nearby cCREs with functional relevance in the colon and overlap with promoter-flanking regions in HCT116 cells, sigmoid colon and the large intestine. Each of these factors suggest a potential role for these candidate causal variants in the regulation of the expression of nearby genes at the locus. However, the role of these candidate causal variants in longer-range regulation of gene expression has not been assessed. In addition to regulation of nearby genes at the same locus as the SNP, referred to as *cis*-eQTLs, it has been shown that genetic variants are able to also regulate the expression of genes up to 5Mb away – acting as *trans*-eQTLs (286). In recent years, techniques have evolved to allow long-range genomic interactions to be studied via the investigation of chromatin conformation. The folding of chromatin within the nucleus of cells can result in regions of DNA located several Mb from one another linearly being brought into close proximity, allowing long-range interactions between genomic regions and, consequently, the long-range regulation of gene expression (287,288). The study by Baxter *et al.* employed novel Capture Hi-C (CHi-C) protocol, an extension of previously employed chromatin conformation capture methodologies, to assess interactions between regulatory elements and target genes in breast cancer (289,290). In this study, CHi-C interaction peaks were identified at thirty-three breast cancer risk loci, encompassing 110 potential target genes (254). In addition to this, there were also examples of long-range interaction between genomic elements more than 4Mb apart (289). It would therefore be of interest to perform similar CHi-C experiments on the candidate causal variants identified by fine-mapping in order to provide information about potential long-range interactions with distant regions of the genome and also provide further information on the role these variants in the regulation of gene expression.

Following the identification and functional annotation of likely causal variants at the locus, TWMR analysis of the 4q24 locus of chromosome 4 using GTEx eQTL data for the transverse colon identified three significant gene-trait associations with CRC. These significant associations included *TET2, AC105391.1* and *TET2-AS1*. Of the three, *TET2*

represents the most likely candidate causal gene underlying CRC predisposition at the locus due to the important role of altered DNA methylation patterns in CRC pathogenesis. However, while *AC105391.1* represents a pseudogene that likely has little biological relevance, *TET2-AS1* represents the antisense transcript of the *TET2* gene. In recent years, there has been increasing evidence for the role of antisense transcripts, a class of non-coding RNA, in the regulation of mammalian gene expression (291). Therefore it is feasible that altered *TET2-AS1* expression may, in turn, affect the expression of the *TET2* gene – thus driving CRC pathogenesis.

The *TET2* gene has previously been implicated in the development of myelodysplastic syndrome (MDS) and acute myeloid leukaemia (AML), with germline mutations in *TET2* and subsequent somatic loss of heterozygosity (LoH) being reported in a number of myeloid malignancies (292). Physiologically, the role of the *TET2* protein is to facilitate the first step in active DNA de-methylation by catalysing the conversion of 5-mC to 5-hmC, initiating a process of stepwise modifications resulting in the replacement of 5-mC with unmodified cytosine (293). Therefore, reductions in *TET2* expression, which was the predicted causal effect of TWMR analysis, may drive CRC pathogenesis via the pathogenic accumulation of DNA methylation within cells. In the context of CRC, Huang *et al.* reported a reduction in nuclear localisation of *TET2* in CRC tissue, but mutations in the *TET2* gene have only been identified in a small proportion of CRCs (294). Recently, an analysis of the landscape of colorectal cancer driver gene mutations in the 100KGP identified *TET2* as a novel candidate CRC driver gene, with nearly 3% of CRCs harbouring an oncogenic mutation in the *TET2* gene (295). Further TWMR-based analysis revealed the transverse colon as the likely target tissue underlying the significant gene-trait association between *TET2* and CRC. Despite the well-described role of *TET2* mutations driving a hyper-methylated phenotype in haematological cancers (296,297), the integration of whole-blood eQTL data with CRC GWAS data concluded there was no significant gene-trait association between *TET2* and CRC. As previously described in Chapter I of this thesis, alterations to the DNA methylation pattern of cells represents a key feature of the adenoma-carcinoma sequence of CRC tumorigenesis. Therefore, it could be feasible that mutations in *TET2* may drive a CRC phenotype via alterations to the DNA methylation profiles of colonic cells.

Previous fine-mapping analysis of the 4q24 locus of chromosome 4 in the context of other cancers identified multiple independent signals associated with enhanced disease risk and also concluded that *TET2* was the likely candidate causal gene underlying these signals (247). Subsequently, Kim *et al.* generated *Tet2*-knockout animal models of breast cancer, where *Tet2* expression was abrogated in the mammary epithelium of transgenic animals (298). These animals were shown to display abnormal mammary development compared to controls and impaired luminal differentiation (298). When these animals were crossed with mice harbouring a mutation in the polyoma middle T (*PyMT*) oncogenic protein, which predisposes animals to breast cancer development, mammary carcinomas were detected in as little as five weeks in animals which also harboured a *Tet2* mutation – compared to animals with a *PyMT* mutation only, which displayed pre-malignant lesions only (298). Furthermore, *PyMT*-mutant animals that also harboured a *Tet2* mutation developed larger and more aggressive cancers when compared to animals with *PyMT* mutations alone (298). This provides evidence that fine-mapping analysis of the 4q24 locus in other cancers and downstream investigation of the candidate causal gene can potentially identify a novel gene

associated with disease predisposition, which may have clinical relevance if the novel predisposition gene were to be added to modern genetic screening panels.

Interestingly, TWMR analysis also revealed a near-significant gene-trait association between *CXXC4* and CRC ($p_{(TWMR)} = 0.0597$). This gene encodes the protein *IDAX,* the *CXXC*-type zinc finger domain of the *TET2* gene (299). Other members of the TET family have this domain incorporated into their own coding sequence, however evolutionary gene fission of the *TET2* gene resulted in the separation of *IDAX* from the *TET2* gene (299). It has been suggested that the *IDAX* protein is able to regulate the expression of the *TET2* gene, facilitating its degradation via binding (300). However, the causal estimate for the *CXXC4* gene in TWMR analysis suggests a down-regulation of *CXXC4* expression drives the near-significant association with CRC. This implies that the near-association was not a consequence of *CXXC4*-mediated *TET2* down-regulation and instead perhaps a consequence of increased *Wnt* signalling – as *CXXC4* has been suggested to act as an inhibitor of the *Wnt* signalling pathway (300).

While the implementation of TWMR provides a useful indication into the identity of the causal gene at the 4q24 locus of chromosome 4, there are limitations associated with its use. The first of these limitations is the need for highly significant eQTLs to act as an input. Porcu *et al.* in the original TWMR publication were able to set a stringent significance threshold of $p_{(eQTL)} < 1.83 \times 10^{-5}$, compared to the threshold of $p_{(eQTL)} < 0.005$ used in this analysis. This is in part a result of the large eQTL study sample size available to the authors of the original study – which combined GTEx and eQTLGen Consortium data to achieve a total of 31,684 individuals in the eQTL study, compared to the 246 individuals with available data for the region of interest in the transverse colon of GTEx v8. Therefore, in order to reach the significance of the original authors, a much larger eQTL dataset would be required, a dataset which is not available in the context of colonic tissues. As previously discussed, long-range interactions between distant regions of the genome may have a role in the regulation of gene expression, therefore acting as *trans*-eQTLs. It would therefore be of interest to extend this TWMR analysis to include *trans*-eQTL data to investigate if any of these long-range interactions may underpin the association with CRC at this genomic locus. Despite these drawbacks, the advantages of TWMR over other MR strategies, primarily the circumvention of horizontal pleiotropy via a multi-instrument, multi-exposure model, made the tool the logical choice for MR analysis of the 4q24 locus of chromosome 4.

Finally, previous TWAS analysis also suggested that altered *TET2* expression was significantly associated with CRC risk in colonic tissues (257). In addition to this, conditional TWAS analysis revealed this significant association was not abolished when conditioned on the genotype of the lead GWAS variant at the locus (257). This data provides support for the conclusions made from the data presented in this chapter. Firstly, there is additional evidence in this TWAS study to suggest that there are multiple independent GWAS signals at this locus associated with CRC predisposition and, secondly, that altered *TET2* expression in colonic tissues only is correlated with CRC risk. As described in section 2.1.4, the utility of TWAS analysis can be limited by correlated expression of genes at the same locus (256). Despite this and the high density of genes at the 4q24 locus of chromosome 4, only the expression of *TET2* was significantly associated with CRC in this dataset – indicating that correlated gene expression has not limited the utility of this data. Furthermore, the number of risk alleles of both the lead variants underpinning the GWAS signals at this locus and the

number of risk alleles of the candidate causal variants identified by trans-ethnic fine-mapping were shown to correlate with reduced expression of *TET2* in the distal colon. Overall, this suggests that these variants, potentially alongside other correlated variants, are associated with altered *TET2* expression in the intestinal compartment, consequently driving enhanced CRC predisposition.

In conclusion, extensive analysis of the 4q24 region of chromosome 4 has revealed multiple independent signals associated with CRC. The fine-mapping of these signals has revealed multiple candidate causal variants that have functional relevance in colonic tissues and the integration of this GWAS meta-analysis data with eQTL data from the transverse colon has revealed *TET2* as the likely candidate causal gene underlying the significant association with CRC seen at this locus. It has also been suggested that colonic tissue is the underlying target tissue that underpins this gene-trait association. A previous study of *Tet2*-deficient mouse models identified abnormalities in the breast tissue and enhanced tumorigenesis of *Tet2*-mutant animals on a genetic background predisposed to breast cancer development (298). Therefore, extending the analysis performed in this chapter to *Tet2*-deficient mouse models of CRC may identify a novel association between *TET2* and CRC predisposition, whilst also providing an insight into the mechanisms by which this association may be driven.

Chapter III – The Role of *TET2* & Isocitrate Dehydrogenase in Driving CIMP & Colorectal Cancer Tumorigenesis

3.1 – Background

3.1.1 – The Role of TET Proteins in DNA De-Methylation

Throughout Chapter I of this thesis, the critical role of DNA methylation in CRC pathogenesis was described. In addition to this, Chapter II of this thesis suggested that down-regulation of ten-eleven translocation 2 (*TET2*) expression, a gene involved in the conversion of 5-methylcytosine (5-mC) to 5-hydroxymethylcytosine (5-hmC) was associated with CRC predisposition. Therefore, it is plausible that pathogenic *TET2* mutations could drive CRC development via alterations to cellular DNA methylation profiles. As briefly discussed in Chapter I of this thesis, *TET2* is a member of the TET family of proteins, involved in catalysing active DNA de-methylation via the oxidation of 5-mC to 5-hmC (301). For many years, DNA methylation was thought to be a permanent and irreversible epigenetic modification, suggesting that promoter methylation of a gene would have the consequence of permanent transcriptional silencing (302). However, the discovery of DNA de-methylation in zygotes and primordial germ cells challenged this paradigm, leading to the discovery of TET-mediated DNA de-methylation (see Figure 3.1) (303,304). This de-methylation process involves the removal of 5-mC and its replacement with unmodified cytosine via a series of intermediates (200). The first step of TET-mediated DNA de-methylation is the oxidation of 5-mC to 5-hmC by one of the TET proteins, followed by the conversion of 5-hmC to 5-formylcytosine (5-fC), which is then subsequently converted to 5-carboxylcytosine (5-caC) (258). Following the conversion of 5-fC to 5-caC, Thymine DNA Glycosylase (*TDG*), in combination with DNA base excision repair (BER) proteins, excises 5-caC from DNA and replaces it with unmodified cytosine (293). The process of TET-mediated DNA de-methylation is summarised in Figure 3.1

Members of the TET family are dioxygenase enzymes which are dependent on both iron and α-ketoglutarate (α-KG) (200,293). *TET1* was the first protein in the family to be identified, which was originally described as a fusion partner of the mixed-lineage leukaemia gene in AML (305). Subsequently, *TET2* and *TET3* were identified and functionally characterised to have the same catalytic activity as *TET1* (200). Previous studies of the TET family identified an important role for these proteins in embryonic development (302). They study by Dawlaty *et al.* demonstrated that loss of both *Tet1* and *Tet2* in mice results in depletion of 5-hmC within mouse embryos (306). While these double-mutant mice were viable, a large proportion presented with abnormalities during gestation and perinatal lethality (306). In addition to this, the study by Koh *et al.* demonstrated the importance of *Tet1* and *Tet2* in mouse embryonic stem cells (mESCs) (307). Depletion of both *Tet1* and *Tet2* reduced 5-hmC expression by 75-80% and the reprogramming of mouse embryonic fibroblasts into induced pluripotent stem cells resulted in a significant up-regulation of both *Tet1* and *Tet2* mRNA (307). A subsequent study by Dawlaty *et al.* characterised the triple knockout of *Tet1, Tet2* and *Tet3* in mESCs (308). Loss of all three TET proteins depleted 5-hmC within mESCs and resulted in a subtle increase in 5-mC (308). Furthermore, knockout of all three TET proteins impaired differentiation and de-regulated the expression of genes critical for proper embryonic development (308).

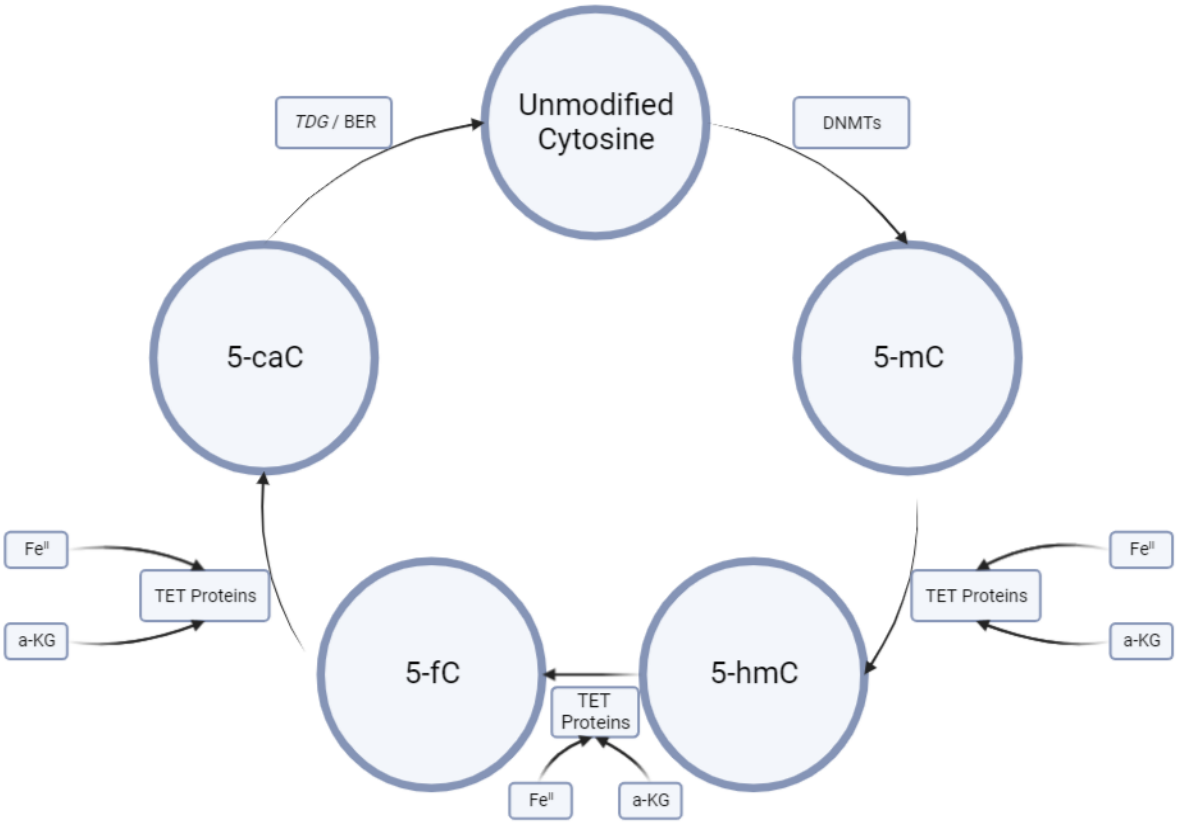


Figure 3.1 – The Process of TET-Mediated DNA De-Methylation: An illustration of the process by which DNA de-methylation is catalysed by members of the ten-eleven translocation (TET) family. Unmodified cytosine can be methylated to 5-methylcytosine (5-mC) by DNA methyltransferases (DNMTs). The initial step of DNA de-methylation is the conversion of 5-mC to 5-hydroxymethylcytosine (5-hmC) by TET proteins in a process dependent on iron (Fe^{II}) and α-ketoglutarate (a-KG). 5-hmC is then converted to 5-formylcytosine (5-fC) and subsequently 5-carboxylcytosine (5-caC). Finally, 5-caC is removed from DNA and replaced by unmodified cytosine in a process dependent on thymine DNA glycosylase (*TDG*) and DNA base excision repair (BER) proteins. Created with BioRender.com (https://app.biorender.com/).

In addition to TET-mediated DNA de-methylation, other mechanisms also exist within cells to convert 5-mC to unmodified cytosine (309). Examples of these depend on the deamination of 5-mC or 5-hmC by AID/APOBEC enzymes (310,311). Deamination of 5-mC produces thymine, resulting in a T:G mismatch in DNA, whereas the deamination of 5-hmC produces 5-hydroxymethyluracil (5-hmU) (310,311). These T:G mismatches generated by the deamination of 5-mC can be repaired by several mechanisms, including either methyl-CpG Binding Domain 4 (*MBD4*), *TDG* or uracil DNA glycosylases acting in combination with BER proteins (310–312). Similarly, 5-hmU may be converted back to unmodified cytosine via the BER pathway (310). The deamination of 5-mC and the role of *MBD4* will be discussed in Chapter IV of this thesis. Furthermore, it has been hypothesised that DNA replication represents a passive form of DNA de-methylation (313). During DNA replication, the newly-synthesised DNA strand is initially unmethylated before *DNMT1*-mediated methylation occurs (313). He *et al.* showed that expression of *DNMT1* is up-regulated in response to cellular proliferation rates, while the inhibition of *DNMT1* resulted in both a

reduced cellular proliferation rate and increased DNA de-methylation (313). Therefore, it is also apparent that failure to re-establish DNA methylation profiles either during or immediately after DNA replication represents another mechanism of passive DNA de-methylation.

As well as these roles in DNA de-methylation, the TET proteins may also possess non-canonical roles involved in transcriptional silencing (314). It has been suggested that *TET1* is able to recruit polycomb repressive complex 2 (PRC2) to DNA, driving the deposition of repressive histone modifications and subsequent transcriptional silencing of certain bivalent promoter regions (see section 3.1.3 for a brief description of bivalent promoters) (314,315).

3.1.2 – *TET2* Protein Structure & Interactions with DNA

Following the initial discovery of the TET proteins, their DNA-modifying function were predicted due to their sequence similarities to J (β-D-glucosyl-hydroxymethyl-uracil) binding proteins 1 and 2 (*JBP1*/*JBP2*), which produce base J via the oxidation of methyl-thymine to 5-hmU (316,317). As previously described, the *TET2* gene is located at the 4q24 locus of chromosome 4 and is comprised of eleven exons – with exon three representing the first protein-coding exon. While *TET1* and *TET3* possess an N-terminal CXXC-type zinc finger domain, evolutionary fission resulted in this domain of the *TET2* protein being encoded by a separate gene known as *IDAX* (see Chapter II of this thesis), which is transcribed in the opposite direction to the *TET2* gene (299).

The crystal structure of the *TET2* protein was determined by Hu *et al.,* who investigated the *TET2* protein while in a complex with DNA (280). In order to facilitate the oxidation of 5-mC to 5-hmC and other subsequent intermediates produced during DNA de-methylation, each member of the TET family possess a C-terminal catalytic domain (280). The catalytic domain of *TET2* is comprised of two sub-domains, a Cys-rich domain and a double-stranded β-helix (DSBH) domain, which itself forms a central DSBH core surrounded by two or three-stranded β-sheets of the DSBH domain and the N-terminal and C-terminal sub-domains of the Cys-rich domain (280). The Cys-rich domain can be further sub-divided into the highly-conserved loop domains L1 and L2, which form the DNA interaction groove of the *TET2* protein – with L1 acting as a DNA support structure and L2 binding to the minor groove of DNA (280). Other studies have suggested that the Cys-rich domain acts as the interface for DNA-protein interactions by binding to the mono or di-methylated H3K36 residue of histone tails (318). Mutations in the Cys-rich domain are suggested to decrease *TET2* binding to methylated histone residues, thereby diminishing *TET2* enzymatic activity (318). The Cys-rich and DSBH regions of the catalytic domain are brought into proximity to one another via zinc chelating residues, which stabilise the *TET2* catalytic domain through co-ordinated interaction with Zn^{II} and Zn^{III} ions (280).

TET2, like other members of the TET family, requires a number of co-factors in order to perform its function. These include the recruitment of α-KG and iron ions (Fe^{II}) to the catalytic site of the protein (280). The binding domains of these co-factors are highly evolutionarily conserved, with mutations in the sites of α-KG interaction or Fe^{II} binding potentially abolishing *TET2* enzymatic activity (280). Unsurprisingly, mutations in these co-

factor interaction sites, examples of which include *TET2*^{R1261G}, *TET2*^{H1382Y}, *TET2*^{R1896M} and *TET2*^{S1898F}, result in reduced *TET2* activity and have been associated with several human cancers, including AML and chronic myelomonocytic leukaemia (CMML) (280,319). An illustration of the structure of the *TET2* protein is provided in Figure 3.2.

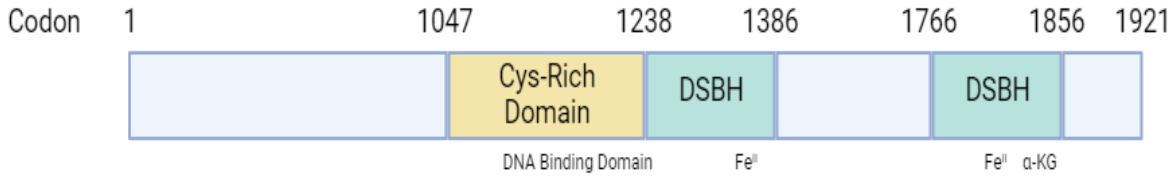


Figure 3.2 – The Structure of the *TET2* Protein: An illustration of the structure of the human *TET2* protein. Indicated are the Cys-rich domains and double-stranded β-helix (DSBH) domains, as well as which regions of the protein interact with DNA, iron ions (Fe^{II}) and α-ketoglutarate (α-KG). Adapted from Garcia-Outeiral *et al.* (320). Created with BioRender.com (https://app.biorender.com/).

3.1.3 – *TET2* Mutations in Human Cancer

3.1.3.1 – Haematological Cancers

Pathogenic mutations in *TET2* have been implicated in the development of several types of human cancer and pre-malignant conditions, resulting in *TET2* being described as a novel tumour suppressor gene in recent years (321). Examples of these pathogenic mutations include frameshift, stop-gained or substitutions at critical domains (see section 3.1.2) (322). As discussed in Chapter II of this thesis, the most well-characterised incidences of *TET2* mutations are in MDS and AML, where it is estimated that *TET2* mutations are present in up to 15% of all myeloid malignancies and 22% of all AMLs (319,321). Mutations in *TET2* are more commonly reported in elderly patients and have been reported in other haematological malignancies, including CMML and angioimmunoblastic T-cell lymphoma (323). However, mutations in *TET1* or *TET3* are rarely seen in these cancers (323,324). Duployez *et al.* reported a family with a germline heterozygous *TET2* frameshift mutation, with each affected

sibling developing AML, CMML and polycythaemia vera at the ages of sixty-one, sixty and fifty-three respectively (325).

Weissmann *et al.* reported the clinical characteristics of *TET2*-mutant AML as being associated with an older age, higher white blood cell count, lower platelet count and inferior event-free survival (6.7 months vs 18.7 months) than *TET2* wild-type (WT) cancers (326). However, despite the high prevalence of *TET2* mutations in myeloid cancers, the exact clinical and prognostic implications of this remain unclear (281). In contrast to Weissmann *et al.,* Nibourel *et al.* investigated the characteristics of patients with *TET2* mutations compared to *TET2*-WT in normal karyotype AML patients and found no significant differences in three-year overall survival (51% vs 54% for *TET2*-mutant and *TET2*-WT respectively) (327). Similarly, the study by Gaidzik *et al.* identified *TET2* mutations in 60 out of 783 (7.6%) AML patients and found no differences in the event-free survival, response to induction chemotherapy, relapse-free survival or overall survival of *TET2*-mutant AML compared to *TET2*-WT counterparts (328). In order to definitively determine the clinical and prognostic utility of *TET2* mutations in AML, Wang *et al.* performed a meta-analysis of sixteen previous studies and determined that mutations in *TET2* were associated with a poorer overall survival and event-free survival in comparison to *TET2*-WT AMLs (329). Poorer overall survival and event-free survival was also observed in AML patients under sixty-five years of age and in AML patients with a normal karyotype, suggesting that mutations in *TET2* are associated with a worse prognosis in AML (329).

In addition to being associated with clonal haematopoiesis, AML and MDS, mutations in *TET2* often drive a DNA hyper-methylation phenotype (297). This is perhaps unsurprising given the role of *TET2* in catalysing active DNA de-methylation. It is plausible that DNA hyper-methylation in *TET2*-mutant cancers is driven by an accumulation of DNA methylation in cells as a result of impaired DNA de-methylation pathways. In the study by Moran-Crusio *et al., Tet2* deletion in mouse haematopoietic tissues resulted in abnormal haematopoiesis and an expansion of the progenitor cell compartment in the bone marrow (296). Similarly, the study by Cimmino *et al.* developed a reversible *Tet2* knockout in mice and observed a similar phenotype of abnormal haematopoiesis and an expansion of the progenitor cell compartment (297). Interestingly, the restoration of *Tet2* to these progenitor cells when cultured *in vitro* promoted cell death or differentiation down a myeloid lineage, as well as a reduction in aberrant DNA hyper-methylation resulting from the initial *Tet2* knockout (297). In addition to this, the treatment of *Tet2*-deficient cells with vitamin C, a co-factor of α-KG dependent dioxygenases known to promote DNA de-methylation, mimicked *Tet2* restoration by having the same effect of reducing progenitor cell self-renewal and promoting DNA de-methylation (297). The hyper-methylation phenotype of *TET2*-mutant AML was also characterised in the study by Figueroa *et al.* who identified 129 significantly hyper-methylated regions in *TET2*-mutant AML patients compared to normal bone marrow cells (330).

3.1.3.2 – Solid Cancers

In addition to their high prevalence in myeloid cancers, rare pathogenic *TET2* mutations have also been reported in solid tumours – including chondrosarcoma, cholangiocarcinoma, glioma, thyroid cancer, hepatocellular carcinoma and CRC (294,321,331,332). The study by

Sajadian *et al.* identified an increase in 5-mC and decrease in 5-hmC in hepatocellular carcinoma via immunohistochemical analysis and an associated reduction in *TET2* and *TET3* mRNA within the cancerous tissue (331). Treatment of cancer cells with 5-aza was shown to reduce proliferation rates in a dose-dependent manner and also appeared to increase the levels of 5-hmC in comparison to untreated cells in a potentially *TET2*-dependent mechanism, thereby suggesting that active DNA de-methylation can be driven by 5-aza treatment (331).

The reduction of 5-hmC within cancer tissues has also been reported in CRC. The study by Uribe-Lewis *et al.* performed 5-hmC profiling in colorectal adenomas and adenocarcinomas – reporting markedly reduced 5-hmC enrichment in cancer tissue in comparison to normal tissue independent of the mRNA levels of any of the TET genes (333). Interestingly, it was also reported that there was a gradient in 5-hmC expression along the crypt-villus axis within intestinal tissue – with 5-hmC more commonly identified in differentiated tissue of the villus compared to progenitor cell populations of the crypt (334). A subsequent study also reported that gene promoters usually characterised by 5-hmC were rarely hyper-methylated in cancer tissue (333). Following *TET2* knockout in HCT116 cells, reductions to 5-hmC was observed, but promoter hyper-methylation was not – suggesting that *TET2* was not responsible for preventing promoter hyper-methylation in cancer cells (333). Furthermore, 65% of promoters marked by 5-hmC in CRC were found to overlap with regions identified as bivalent promoters in human embryonic stem cells compared to 30% of 5-hmC marked promoters in normal tissue (333). Bivalent promoters are promoters simultaneously marked with the activating histone modification H3K4me^3 and the repressive histone mark H3K27me^3 (335). These bivalent promoters are commonly found at developmental genes in ESCs, allowing tight transcriptional regulation of these genes during development (335). The loss of H3K4me^3 from bivalent promoters is thought to drive aberrant hyper-methylation in cancer, potentially leading to the silencing of tumour suppressor genes (335). While this suggestion has led to interest in the hyper-methylation of bivalent promoters in cancer, the study by Dunican *et al.* noted that most bivalent promoter hyper-methylation in cancer was secondary to transcriptional repression of the associated gene(s) – suggesting that this cancer-associated hyper-methylation was not causative of transcriptional silencing in the majority of cases (336). Therefore, the exact role of bivalent promoter hyper-methylation remains controversial, with some suggesting that the mechanisms underpinning H3K4me^3 loss are the true driver of disease (336). As previously reported in Chapter II of this thesis, *TET2* nuclear localisation has been reported to be reduced in CRC, as well as reduced *TET2* mRNA being identified in both CRC tissue and CRC cell lines (294). However, *TET2* mutations appear to have no clinical or prognostic implications for CRC patients (333). Overall, it is clear that the prevalence of *TET2* and 5-hmC is reduced in solid tumours, but the importance of this in disease pathogenesis remains unclear.

3.1.4 – Inhibition of *TET2* by Mutant Isocitrate Dehydrogenase

In addition to *TET2*, mutations in other genes in AML drive a hyper-methylated phenotype. Mutations in Isocitrate Dehydrogenase (IDH) 1 or 2 were first reported in glioblastoma following next-generation sequencing of twenty-two patients (337–339). Pathogenic mutations in the IDH active site have been reported in ~12% of glioblastomas and more than

75% of Grade II/III glioblastomas or secondary glioblastomas (338,340). In addition to glioblastoma, mutations in IDH have also been identified in chondrosarcoma, cholangiocarcinoma, AML and MDS (323,330,341,342). The homodimeric *IDH1* and *IDH2* proteins, which share up to 70% structural similarity, play an important role in cellular metabolism – whereas the heterotrimeric *IDH3* is less similar to other members of the family (343,344). All three of the IDH proteins perform a critical role in the Tricarboxylic Acid (TCA) cycle by catalysing the conversion of isocitrate to α-KG, which may subsequently be used in cellular metabolism or as a co-factor for as many as sixty iron-dependent dioxygenase enzymes (340,343,345–347). The *IDH1* protein primarily operates within peroxisomes, whereas *IDH2* is restricted to the mitochondrial matrix (343).

Recurrent point mutations in the isocitrate-binding domains of *IDH1* and *IDH2* have been reported in IDH-mutant cancers, with $IDH1^{R132H}$ estimated to account for up to 85% of *IDH1* mutations (346). In addition to this mutation, $IDH1^{R132C}$, $IDH1^{R132G}$, $IDH1^{R132L}$, $IDH2^{R140Q}$ and $IDH2^{R172K}$ have also been reported in cancers, with $IDH2^{R172}$ representing the homologue of the $IDH1^{R132}$ codon (348,349). Instead of being conventional loss-of-function mutations, these IDH mutations result in a neomorphic gain-of-function, whereby instead of converting isocitrate to α-KG, mutant IDH instead produces 2-hydroxyglutarate (2-HG) – a known oncometabolite with tumour-promoting effects (350–352). It has been suggested that the production of 2-HG by IDH-mutant cells acts as a competitive inhibitor of α-KG dependent dioxygenases, including *TET2,* which may explain the similar hyper-methylated phenotype identified in both *TET2*-mutant and IDH-mutant AML (330,352,353). Interestingly, it has been reported that mutations in IDH and *TET2* are mutually exclusive in AML. The study by Figueroa *et al.* studied 385 AML patients and identified pathogenic *IDH1* or *IDH2* mutations in 14.8% of patients, pathogenic *TET2* mutations in 7.3% of patients and found no overlap between *IDH1, IDH2* or *TET2* mutations (330). In addition to this, mutations in *IDH1* or *IDH2* resulted in elevated 2-HG levels and a concomitant increase in 5-mC (330). Transfection of HEK293T cells with flag-tagged *TET2* resulted in an increase in nuclear 5-hmC as expected, however, co-transfection of *TET2* with $IDH1^{R132H}$ reversed the increase in 5-hmC – further suggesting that 2-HG produced by mutant IDH inhibits *TET2*, thereby driving an accumulation of 5-mC (330). When compared to normal bone marrow, *TET2*-mutant AMLs presented with 129 genomic regions that were significantly hyper-methylated and of these, 61% were also seen in IDH-mutant AML (330). Furthermore, 93% of genes aberrantly expressed in *TET2*-mutant AMLs were also aberrantly expressed in IDH-mutant cancers, further highlighting the phenotypic similarities between *TET2*-mutant and IDH-mutant AML (330).

Pathogenic IDH mutations have also been reported to drive a malignant phenotype in other cancers. The study by Bardella *et al.* constructed a mouse model with a conditional knock-in of $Idh1^{R132H}$ in cells of the sub-ventricular zone (SVZ) – a neural progenitor cell niche (354,355). Animals began displaying adverse effects four to six weeks post-induction of knock-in and by two months of age presented with an expanded SVZ with a concomitant increase in progenitor cell numbers compared to control animals (354). This displays similarities with the previous studies of *TET2*-mutant AML animals which also displayed an expanded progenitor cell population (297,354). *Idh1*-mutant animals also displayed a two-fold increase in 2-HG and a 30% reduction to α-KG – alongside this, *Idh1*-mutant animals displayed a significant reduction in 5-hmC compared to controls and a non-significant

increase in 5-mC, suggesting that the production of 2-HG drives the reduction in 5-hmC via the inhibition of TET proteins (354). A summary of this proposed mechanism is presented in Figure 3.3.

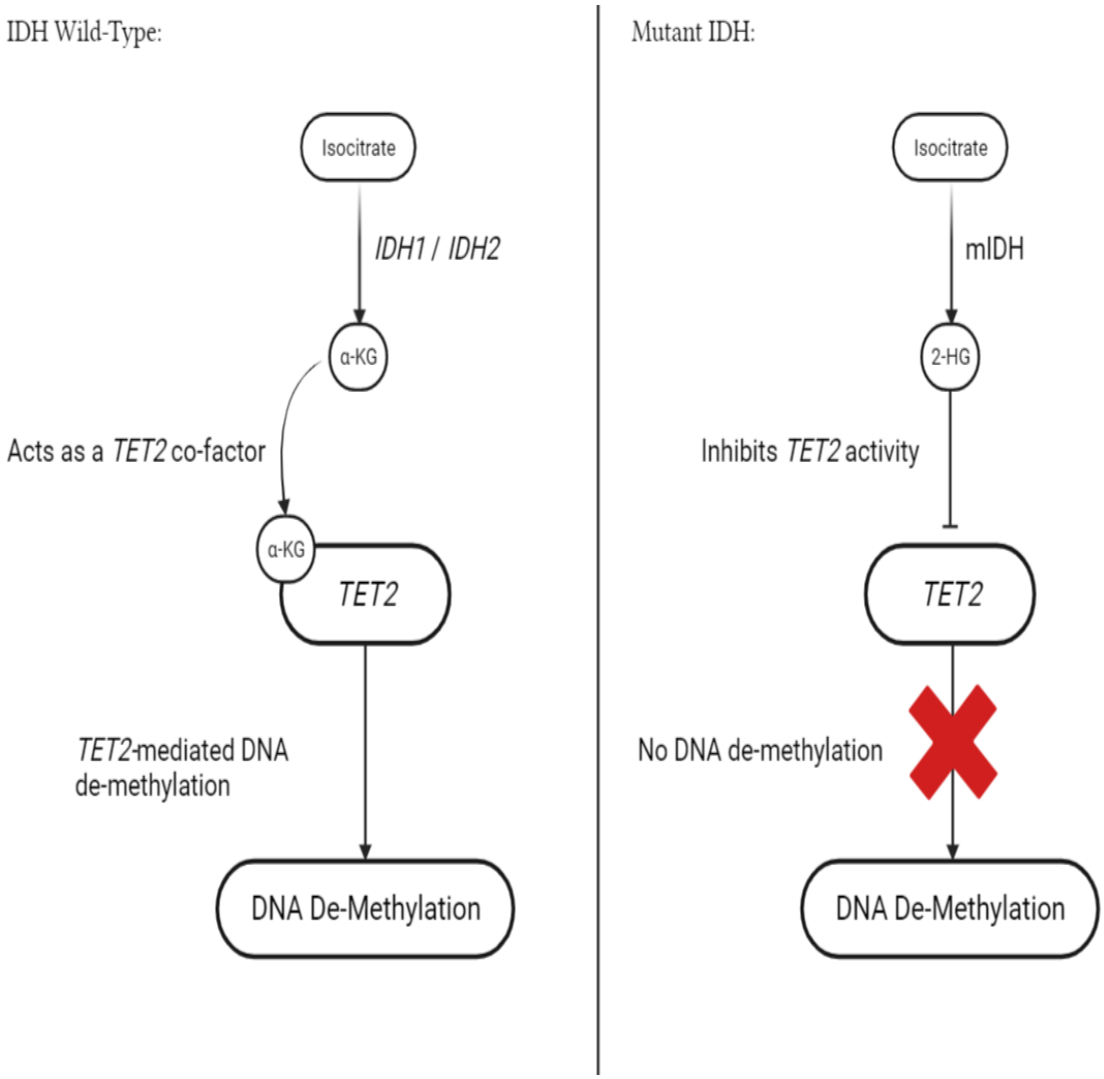


Figure 3.3 – The Proposed Mechanism of DNA Hyper-Methylation Driven by Mutant Isocitrate Dehydrogenase: The mechanism by which mutations in isocitrate dehydrogenase (IDH) 1 or 2 may drive DNA hyper-methylation in cancer. Shown on the left is a schematic for wild-type IDH catalysing the conversion of isocitrate to α-ketoglutarate (α-KG), which subsequently acts as a co-factor for the ten-eleven translocation 2 (*TET2*) protein which drives active DNA de-methylation. However, in cancers with mutant IDH (mIDH – right), isocitrate is instead converted to 2-hydroxyglutarate (2-HG), which inhibits the *TET2* protein and therefore downstream DNA de-methylation. This inhibition of DNA de-methylation pathways may subsequently drive DNA hyper-methylation seen in cancers with IDH mutations. Created with BioRender.com (https://app.biorender.com/).

3.1.5 – Chapter Aims

It is apparent that *TET2* mutations play a key role in myeloid malignancies, driving disease pathogenesis via an expansion of hyper-methylated stem cell populations (297,330). Similar phenotypes have been observed in IDH-mutant AML and, to an extent, glioblastoma (330,354). Therefore, it is plausible that the two phenotypes share a common pathway, driving hyper-methylation by *TET2* inactivation via either pathogenic mutation or competitive inhibition by 2-HG produced by mutant IDH (321,330,356).

Chapter I of this thesis discussed the extensive DNA hyper-methylation presented by $CIMP^+$ CRCs and also described how the mechanisms underlying the development of $CIMP^+$ CRCs remains poorly understood. Given the previously described hyper-methylation and loss of 5-hmC in other cancers harbouring mutations in *TET2* or IDH, it is feasible that the same mechanism of either *TET2* pathogenic mutation or inhibition by 2-HG may drive DNA hyper-methylation (and therefore CIMP) in CRC. This is further supported by evidence presented in Chapter II of this thesis, identifying a possible association between reduced *TET2* expression and CRC predisposition. In addition to this, *IDH1* was, alongside *TET2,* identified as a potentially novel CRC driver gene following analysis of 100KGP data (295). Therefore, this chapter will:

- Assess the role of *TET2* in colorectal tumorigenesis via *Tet2*-knockout mouse models.
- Assess the impact of *TET2, IDH1* and *IDH2* mutations on DNA methylation and the development of CIMP in CRC using methylation array data of CRCs from the TCGA database.
- Assess the characteristics of $CIMP^+$ cancers in this cohort.
- Assess if *TET2*-mutant or IDH-mutant CRCs with CIMP have distinct molecular features compared to other hyper-methylated CRCs.

These aims are accompanied by the following hypotheses:

- *Tet2*-knockout mouse models of CRC will develop intestinal tumours characterised by a loss of 5-hmC and gain of 5-mC.
- Mutations in *TET2, IDH1* and *IDH2* will significantly affect DNA methylation in CRC.
- Mutations in *TET2, IDH1* and *IDH2* will be associated with hyper-methylation and CIMP.

3.2 – Materials & Methods

3.2.1 – *Tet2*-Knockout Mouse Models of Colorectal Cancer

3.2.1.1 – Mice

An intestinal-specific knockout of *Tet2* was achieved using the established Cre-LoxP system (357). Briefly, Cre-recombinase represents a tyrosine site-specific recombinase, recognising specific 34 base-pair locus of x-over P1 (LoxP) sites (357). Recognition of these LoxP sites inserted into the DNA results in the excision of the DNA between the two LoxP sites, which can be manipulated to produce tissue-specific knockout animals (357). *Tet2*$^{fl/fl}$ and constitutive *Vil1-cre* mice were obtained from the Jackson Laboratory (catalogue #017573 and #035595 respectively). The *Tet2*$^{fl/fl}$ animals were engineered to contain LoxP sites either side of exon three of the *Tet2* gene, while *Vil1-cre* mice contained Cre-recombinase under the control of the *Vil1* promoter, a gene strongly expressed in the intestinal epithelium (296,358). Therefore, crossing these animals together would result in the deletion of exon three in the intestinal epithelium, resulting in a tissue-specific knockout of *Tet2.* All mouse strains were backcrossed for three generations onto a C57BL/6J background. For genotyping, ear snips taken from mice at fourteen days were incubated for forty minutes at 95°C in 50µl of HotShot buffer (25mM NaOH, 0.2mM disodium ethylenediaminetetraacetic acid – pH = 12). The solution was then neutralised with 50µl neutralising buffer (40mM Tris-HCl – pH = 5). For *Tet2*$^{fl/fl}$, *Vil1-cre* and a secondary Cre primer set, a polymerase chain reaction (PCR) mix of 10µl EconoTaq PLUS Green 2x MasterMix (VWR – catalogue #95024-004), 0.8µl of each required primer at 12.5µM and 2µl of the extraction mix described above was made up to 20µl with ddH_2O. The sequences of each genotyping primer are provided in Table 3.1. An illustration of the *Tet2*$^{fl/fl}$ mouse construct, as well as the expected bands for each *Tet2* and *Vil1-cre* genotype, is provided in Figure 3.4.

Primer Name:	Primer Sequence:	Expected WT Band (bp):	Expected Mutant Band (bp):
Tet2-fl-FW	5'-AAGAATTGCTACAGGCCTGC-3'	249	427 / ~620
Tet2-fl-RV	5'-TTCTTTAGCCCTTGCTGAGC-3'	249	427
Tet2-LoxP3R	5'-TAGAGGGAGGGGGCATAAGT-3'	None	~620
Vil1-cre-Control-FW	5'-AGTGGCCTCTTCCAGAAATG-3'	521	None
Vil1-cre-Control-RV	5'TGCGACTGTGTCTGATTTCC-3'	521	None
Vil1-cre-Transgene-FW	5'-CCAGTTTCCCTTCTTCTTCTG-3'	None	~280
Vil1-cre-Transgene-RV	5'-CGGTTATTCAACTTGCACCA-3'	None	~280
Cre-FW	5'-TTACCGGTCGATGCAACGAG-3'	None	~550
Cre-RV	5'-CCACCGTCAGTACGTCAGAT-3'	None	~550

Table 3.1 – Primer Sequences for Mouse Genotyping: The 5' to 3' sequence of each primer used for mouse genotyping. Included are the name of the primer and its sequence. Adapted from Moran-Crusio *et al.* (296).

Genotyping PCR conditions for *Tet2*$^{fl/fl}$ and the secondary Cre primers comprised of 94°C initial denaturation for two minutes, thirty-five cycles of 94°C denaturation (fifteen seconds), 55°C annealing (thirty seconds) and 72°C extension (three minutes) and a final 72°C extension for seven minutes. For the *Vil1-cre* specific primers, PCR conditions were 98°C initial denaturation (thirty seconds), followed by ten cycles of a 65°C to 60°C touchdown, twenty-eight cycles of 94°C denaturation (seven seconds), 60°C annealing (thirty seconds) and 72°C extension (thirty seconds) and a final 72°C extension for two minutes.

An illustration of the breeding plan used for this study is provided in Figure 3.5. Briefly, *Tet2*$^{fl/fl}$ animals were crossed with *Vil1-Cre* animals, resulting in the generation of *Tet2;Vil1-cre* animals, which were born at the expected Mendelian ratios (data not shown). The resulting *Tet2*$^{fl/+}$*;Vil1-cre*^{Positive} animals were self-crossed to produce the range of genotypes presented in Figure 3.5. Five homozygous animals (*Tet2*$^{fl/fl}$*;Vil1-cre*^{Positive}), five heterozygous animals (*Tet2*$^{fl/+}$*;Vil1-cre*^{Positive}) and five control animals (either *Tet2*$^{+/+}$*;Vil1-cre*^{Positive}, *Tet2*$^{fl/+}$*;Vil1-cre*^{Negative} or *Tet2*$^{fl/fl}$*;Vil1-cre*^{Negative}) were aged for six months before sacrifice and subsequent harvesting of intestinal tissues. Intestinal tissue was separated from fat tissue, flushed with phosphate-buffered saline (PBS) and opened longitudinally. The colonic tissue and three equally-sized sections of the small intestine were then prepared according to the standard "Swiss-Roll" technique and left overnight at room temperature in 10% neutral-buffered formalin. Samples were then embedded into paraffin blocks for subsequent histological analysis.

3.2.1.2 – *in situ* Hybridisation

In order to confirm the recombination and subsequent knockout of *Tet2* in *Tet2;Vil1-cre* animals, 5µm sections were cut from blocks of *Tet2*$^{fl/fl}$*;Vil1-cre*^{Positive}, *Tet2*$^{fl/+}$*;Vil1-cre*^{Positive} and control animals using a rotary microtome. Sections were baked at 70°C for one hour, de-paraffinised in xylene for ten minutes and submerged in 100% ethanol for a further two minutes. Slides were rinsed with ddH_2O and left to air dry. The RNA-Scope *in situ* hybridisation protocol (Advanced Cell Diagnostics – ACD) was then performed for *Tet2* according to the manufacturer's instructions. Briefly, slides were incubated for ten minutes at room temperate with an endogenous blocking agent (ACD – catalogue #322330), rinsed in ddH_2O and boiled for fifteen minutes in target retrieval buffer (ACD – catalogue #322000). Slides were then rinsed in ddH_2O followed by 100% ethanol and incubated for thirty minutes at 40°C with protease plus (ACD – catalogue #322330).

RNA-Scope probes targeted to base-pairs 401-1,373 of mouse *Tet2* (ACD – catalogue #511591) were applied to the slides and left to hybridise at 40°C for two hours. Slides were then washed twice with RNA-Scope wash buffer (ACD – catalogue #3100911). The slides were then incubated at 40°C with AMP1 (ACD – catalogue #322310) for thirty minutes and

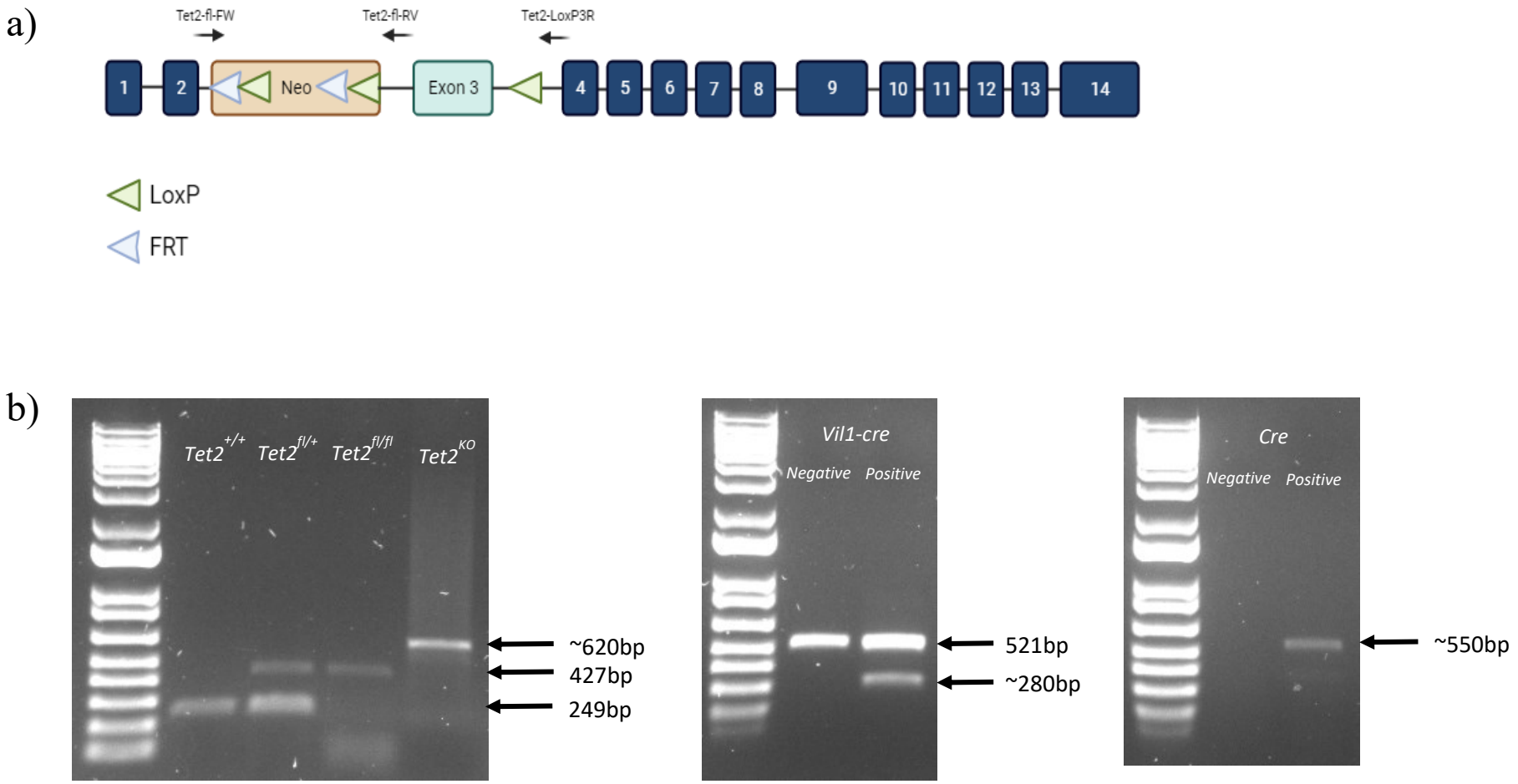


Figure 3.4 – The Development of *Tet2*-Knockout Mouse Models: A diagrammatic illustration of the *Tet2*$^{fl/fl}$ mice used to produce an intestinal-specific *Tet2* knockout (a). Also presented are where each of the genotyping primers presented in Table 3.1 bind to the construct. Adapted from Moran-Crusio *et al.* (296). Created with BioRender.com (https://app.biorender.com/). Also shown are the expected bands for each genotype (b).

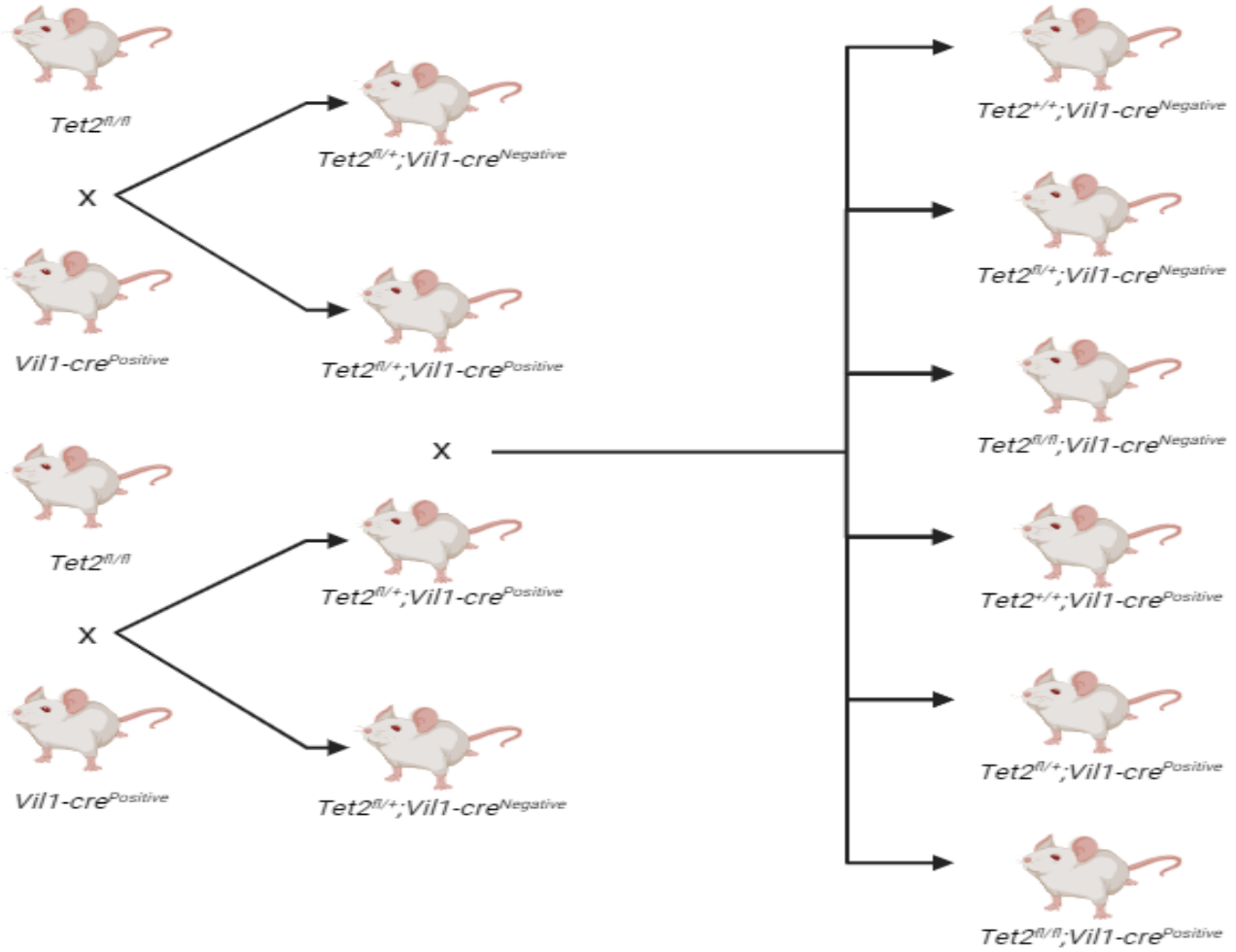


Figure 3.5 – *Tet2;Vil1-cre* Mouse Breeding Plan: An illustration of the breeding plan to produce intestinal-specific *Tet2*-knockout mouse models. Created with BioRender.com (https://app.biorender.com/).

washed twice with RNA-Scope wash buffer. This process was repeated with AMP2 (fifteen minutes), AMP3 (thirty minutes) and AMP4 (fifteen minutes). Slides were incubated with AMP5 (forty-five minutes) and AMP6 (fifteen minutes) at room temperature. Signal was detected by incubating slides at room temperature for ten minutes with an equal mix of DAB-A and DAB-B (ACD – catalogue #322310), counterstained for two minutes in 50% Gill's Haematoxylin (ThermoFisher – catalogue #6765005) and mounted to glass coverslips with EcoMount mounting medium (BioCare Medical – catalogue #EM897L).

3.2.1.3 – Haematoxylin & Eosin Staining

Haematoxylin and Eosin (H&E) staining was performed using standard techniques. Briefly, 4µm sections of mouse intestinal tissue were de-paraffinised in xylene for five minutes at room temperature and subsequently dehydrated through a sequence of 100%, 90% and 70% ethanol for five minutes each at room temperature. Slides were rinsed in ddH_2O and submerged in 50% Gill's Haematoxylin (see section 3.2.1.2) for forty-five seconds, rinsed in ddH_2O followed by acid alcohol (0.05M hydrochloric acid in 70% ethanol) and submerged in eosin (Sigma-Aldrich – catalogue #HT110232) for three minutes. Slides were then rinsed a final time in ddH_2O, rehydrated in ethanol and mounted to glass coverslips using EcoMount mounting medium (see section 3.2.1.2).

3.2.1.4 – Immunohistochemistry

Immunohistochemistry for 5-mC and 5-hmC in *Tet2;Vil1-cre* animals was performed using standard techniques. Briefly, 4µm sections of intestinal tissue were de-paraffinised in xylene at room temperature for six minutes and sequentially transferred through 100%, 90% and 70% ethanol solutions for six minutes each at room temperature. These slides were then rinsed in ddH_2O, placed into a pressure cooker and boiled at for fifteen minutes in citrate antigen retrieval buffer (Agilent – catalogue #S2369). Slides were cooled for thirty minutes and incubated for five minutes in peroxidase blocking solution (Agilent – catalogue #S2023). Slides were incubated with antibodies against mouse 5-mC (Abcam – catalogue #214727) or mouse 5-hmC (Active Motif – catalogue #39769) diluted 1:500 in antibody diluent (Agilent – catalogue #3022) overnight at 4°C and washed four times in phosphate-buffered saline with 0.1% tween (PBS-T). Slides were then incubated with a 1:500 dilution of the polyclonal goat anti-rabbit horse radish peroxidase-conjugated secondary antibody (Agilent – catalogue #P0448) for one hour at room temperature. Signal was visualised using a standard DAB incubation (Agilent – catalogue #K3467). Slides were counterstained in 50% Gill's Haematoxylin and mounted to glass coverslips using EcoMount mounting medium as described in section 3.2.1.2.

3.2.2 – *in silico* DNA Methylation Analysis

3.2.2.1 – Data Availability

DNA methylation data was downloaded from the TCGA GDC portal via the R package *TCGAbiolinks* (359) for 526 participants in either the colorectal adenocarcinoma (COAD) or rectal adenocarcinoma (READ) domains (69). Methylation data was available from either Illumina 27K or Illumina 450K methylation arrays, which were batch corrected to allow comparisons to be made for 20,618 methylation probes common to both arrays via ChAMP by Dr James Wood and Melissa Morgan (360). CIMP statuses for 388 of these cancers were provided by Dr James Wood, Dr Enric Domingo and Melissa Morgan, who determined cancers as $CIMP^{-}$, $CIMP^{Low}$ or $CIMP^{High}$ using previously published studies alongside a recursive-partitioning algorithm on the most variable 10% of probes outside the sex chromosomes (361,362). Pathogenic mutations in candidate CIMP driver genes were identified in TCGA-COAD and TCGA-READ cancers via a search of the GDC and subsequent confirmation via a search of the associated whole-exome sequencing variant call file (VCF). In addition to this TCGA data, other CRC datasets were available with the associated CIMP status of each cancer. Additional data was available for 666 participants of the Stratification in Colorectal Cancer: From Biology to Treatment Prediction (S:CORT) consortium and 618 participants from the Dana-Farber Cancer Institute (DFCI) (363,364).

3.2.2.2 – Characterisation of Cancers with Pathogenic Mutations in Candidate CIMP Driver Genes

The average DNA methylation β-value was taken across the 20,618 methylation probes for each cancer in the TCGA-COAD and TCGA-READ cohort as a measure of average DNA methylation in that cancer. Boxplots were generated for samples with pathogenic mutations in the candidate CIMP driver genes versus WT counterparts using the R package *ggpubr* and tested for significant differences using a Wilcoxon Test (365).

In order to investigate correlations between CIMP and pathogenic mutations in the aforementioned candidate CIMP driver genes, a linear mixed-effects model was constructed using CIMP status as the outcome variable. For simplicity, in this model $CIMP^{Low}$ and $CIMP^{High}$ cancers were considered together as $CIMP^{+}$, with the fixed effects included in the model being pathogenic mutations in candidate CIMP driver genes, MSI status, age, sex, tumour stage and tumour location – separated into proximal and distal colon. Of the three studies with available CIMP status data, 338 cancers from TCGA-COAD and TCGA-READ had all the above clinical data, as well as 472 cancers from the S:CORT study and 446 cancers from DFCI data. In order to correct for any inter-study differences that could influence the model, the study to which a cancer belonged was included as a random effect in the linear mixed-effects model. This model was constructed using the *lme4* R package and statistical significance was determined for each variable in the model using *lmerTest* (366,367).

In order to characterise probes that were differentially methylated in *TET2*-mutant, *TET2*-WT, IDH-mutant and IDH-WT $CIMP^{+}$ cancers compared to their WT $CIMP^{-}$ counterparts,

probe β-values were converted to M-values according to the following formula described by Du *et al.* (368):

$$M = log_2(\frac{\beta}{1-\beta})$$

A linear model was fitted for each probe using the R package *limma* with the help of Dr Juan Fernández-Tajes (369). Probes were characterised as significantly differentially methylated in $CIMP^+$ cancers compared to $CIMP^-$ cancers if the Benjamini-Hochberg corrected p-value ($p_{(BHC)}$) was less than 0.05. Hyper-methylated probes ($p_{(BHC)} < 0.05$) with a $\log_2$(Fold Change) > 1 were considered to be extensively hyper-methylated and the mean β-values of these probes in *TET2*-mutant or IDH-mutant cancers were compared to their WT counterparts using a paired Wilcoxon test via the R package *PairedData* (370). The locations of 5,766 bivalent promoter regions were obtained from the study by Court & Arnaud, who produced a consensus map of these regions in five human embryonic stem cell lines (371). These bivalent promoter regions were lifted from hg18 to hg19 using the UCSC LiftOver tool (372). An additional map of 32,677 promoter regions was obtained from the large intestine dataset in Ensembl BioMart (hg19) (268). Probe annotations from the batch-corrected TCGA-COAD and TCGA-READ datasets also listed 15,665 probes located within CpG islands. Differentially methylated probes in $CIMP^+$ cancers were mapped to each of these regions and a chi-squared (χ^2) test was performed with one degree of freedom to test if the number of differentially methylated probes mapping to each feature was greater than would be expected.

3.3 – Results

3.3.1 – Loss of *Tet2* in Mouse Intestinal Tissues Results in No Adverse Phenotype

Following the analyses performed in Chapter II of this thesis, reduced expression of the *TET2* protein has been suggested to be the causal element underpinning the GWAS association with CRC found at the 4q24 locus of chromosome 4. In addition to this, previous TWAS models have also implicated reduced *TET2* expression in CRC development (257). Therefore, in order to explore this experimentally, $Tet2^{fl/fl}$ mice were crossed with the constitutive *Vil1-cre* in order to delete exon three of the *Tet2* gene specifically in the intestinal tissue of these animals, consequently abolishing expression of the *Tet2* protein in the intestinal tissues. $Tet2^{fl/fl}$;$Vil1\text{-}cre^{Positive}$, $Tet2^{fl/+}$;$Vil1\text{-}cre^{Positive}$ and control animals (see section 3.2.1.1) were born at the expected Mendelian ratios. Intestinal tissues were harvested from these animals at six months of age and prepared for subsequent histological examination using standard techniques (see section 3.2.1 for details). As seen in Figure 3.6a, there was a reduction of *Tet2* expression in the epithelial cells of the small intestine of $Tet2^{fl/+}$;$Vil1\text{-}cre^{Positive}$ animals compared to controls. In addition to this, $Tet2^{fl/fl}$;$Vil1\text{-}cre^{Positive}$ animals presented with a near-complete knockout of *Tet2* within the small intestine epithelium (see Figure 3.6a). However, as seen in Figure 3.6b, H&E analysis of the small intestines of both $Tet2^{fl/+}$;$Vil1\text{-}cre^{Positive}$ and $Tet2^{fl/fl}$;$Vil1\text{-}cre^{Positive}$ animals appeared histologically normal, with no noticeable abnormalities or tumours compared to WT controls.

In order to characterise any potential changes in DNA methylation that may have arisen as a consequence of *Tet2* loss, IHC analysis was also performed for 5-hmC (Figure 3.6c) and 5-mC (Figure 3.6d). As seen in Figure 3.6c, the expression of 5-hmC appears to be follow the same gradient previously reported by Uribe-Lewis *et al.,* where the expression of 5-hmC appeared to be strongest in the villus and absent in intestinal crypts (334). However, there appears to be no substantial difference in the expression of 5-hmC in the intestines of control animals and $Tet2^{fl/fl}$;$Vil1\text{-}cre^{Positive}$ animals, possibly due to the rarity of 5-hmC within intestinal tissues (334). Furthermore, there appears to be little difference in the overall DNA methylation in the intestinal compartment of *Tet2*-knockout animals compared to controls, according to 5-mC staining (see Figure 3.6d).

A similar lack of abnormalities was observed upon histological examination of the colonic tissues of these *Tet2*-knockout animals (see Figure 3.7). As also seen in the small intestine, the expression of *Tet2* was reduced in the colonic crypts of the $Tet2^{fl/+}$;$Vil1\text{-}cre^{Positive}$ and $Tet2^{fl/fl}$;$Vil1\text{-}cre^{Positive}$ animals compared to controls (see Figure 3.7a). These colonic crypts presented with no gross abnormalities and no tumours were detected in the colons of *Tet2*-knockout animals (Figure 3.7b). Unlike the small intestine, expression of 5-hmC was detected in the colonic crypts of control animals and *Tet2*-knockout animals, although there was no obvious reduction in 5-hmC in the *Tet2*-knockout animals compared to the controls (see Figure 3.7c). Finally, expression of 5-mC was detected in the colonic crypts of control animals and *Tet2*-knockout animals (Figure 3.7d). However, there was not a noticeable difference in the *Tet2*-knockout animals compared to the controls.

Overall, this indicates that targeted deletion of *Tet2* in the intestinal epithelium does not, on its own, drive colorectal tumorigenesis in these animals. There also appears to be no gross differences in the DNA methylation profiles of control and *Tet2*-knockout animals. However, it cannot be excluded that there may be more subtle changes in the DNA methylation profiles of *Tet2*-deficient animals that could not be detected via IHC analysis. Therefore, analysis of

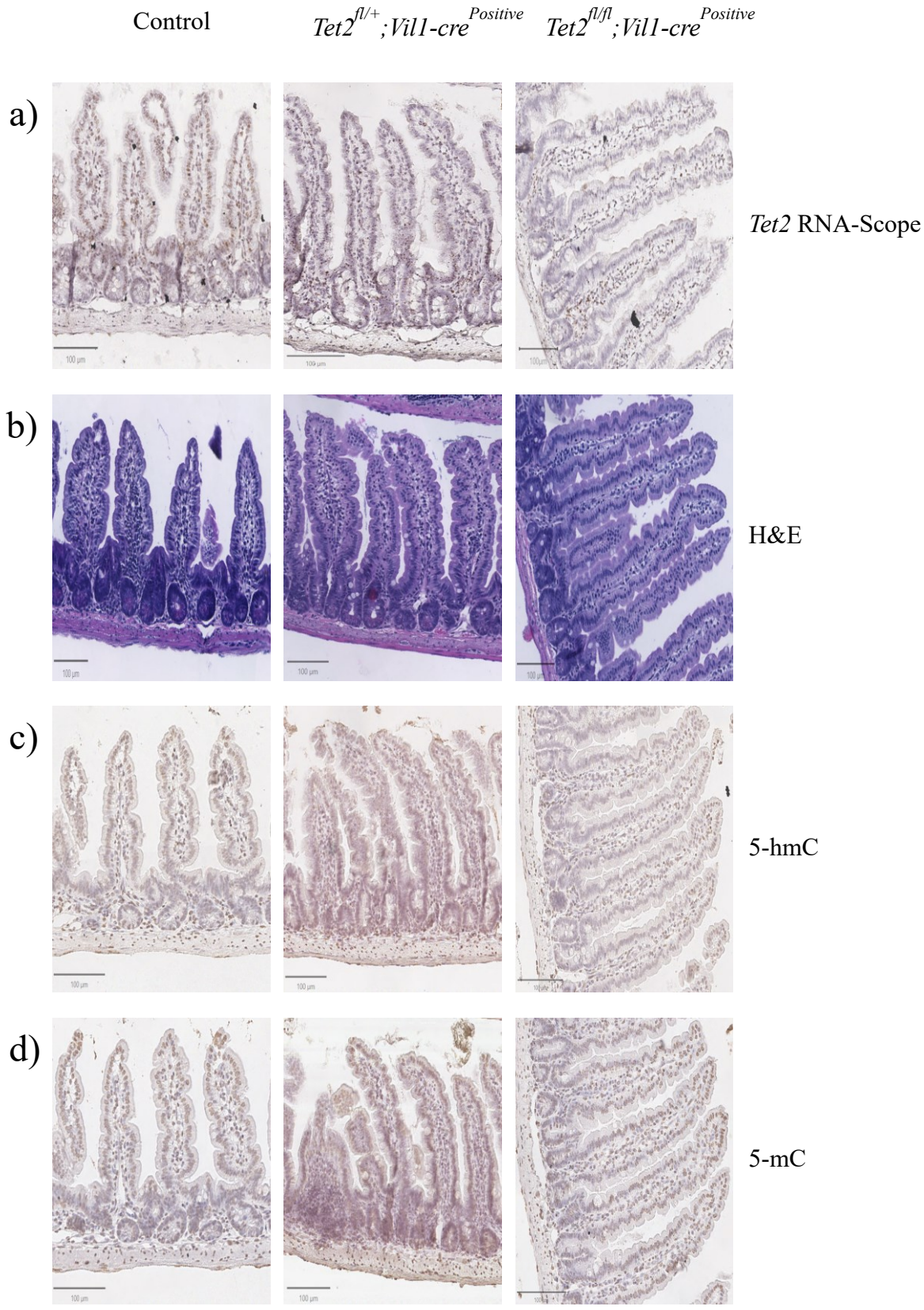


Figure 3.6 – Histological Analysis of the Small Intestine of *Tet2*-Knockout Animals: Images of the small intestine of control, *Tet2fl/+;Vil1-cre*^{Positive} and *Tet2*$^{fl/fl}$*;Vil1-cre*^{Positive} mice. Included are images of RNA-Scope staining for *Tet2* (a), a haematoxylin & eosin (H&E) stain (b), immunohistochemistry analysis for 5-hydroxymethylcytosine (5-hmC) (c) and 5-methylcytosine (5-mC) (d). Brown = positive. Purple = negative

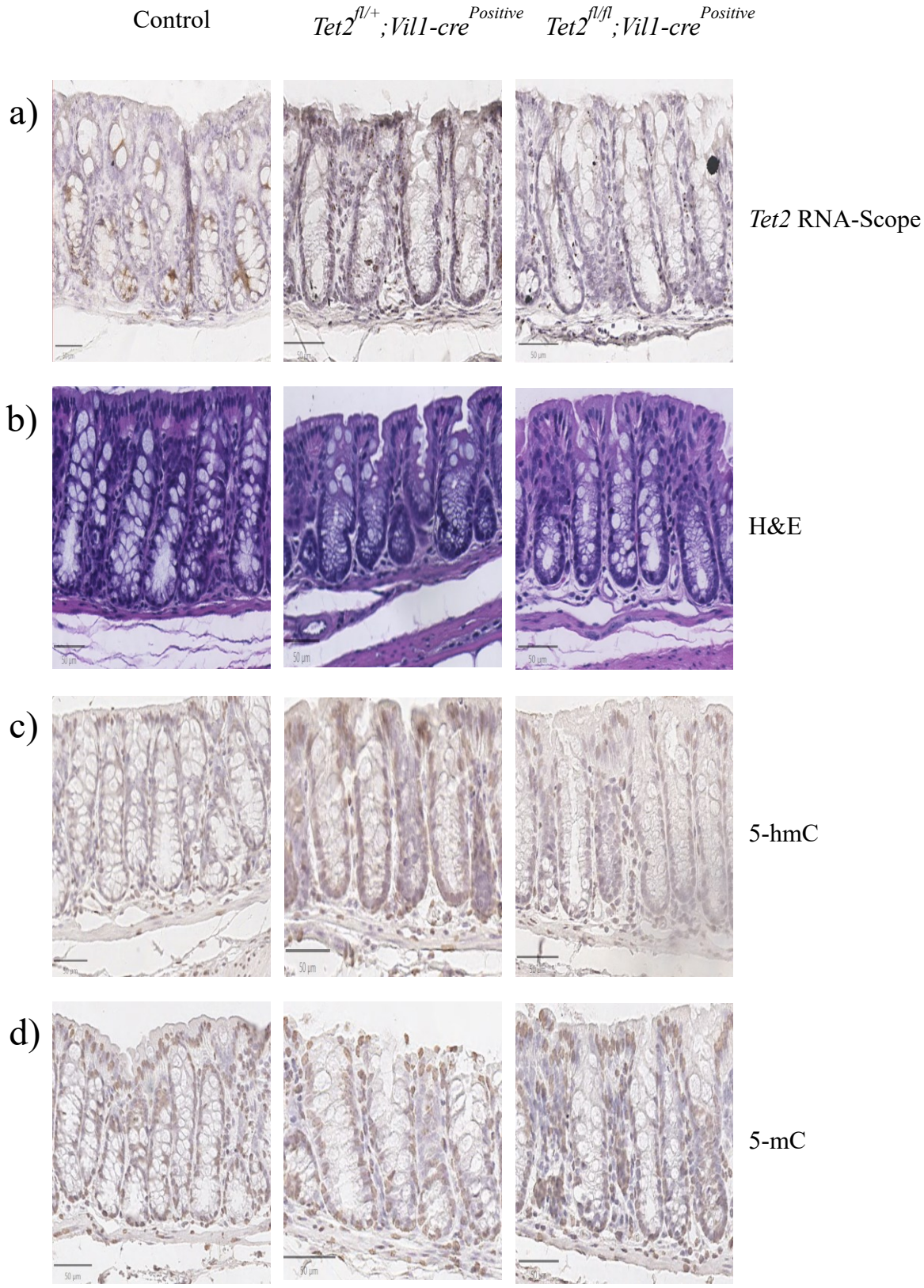


Figure 3.7 – Histological Analysis of the Colon of *Tet2*-Knockout Animals: Images of the colon of control, *Tet2*$^{fl/+}$;*Vil1-cre*^{Positive} and *Tet2*$^{fl/fl}$;*Vil1-cre*^{Positive} mice. Included are images of RNA-Scope staining for *Tet2* (a), a haematoxylin & eosin (H&E) stain (b), immunohistochemistry analysis for 5-hydroxymethylcytosine (5-hmC) (c) and 5-methylcytosine (5-mC) (d). Brown = positive. Purple = negative.

publically-available DNA methylation array data from human CRCs may provide a more quantitative assessment of the consequences of *TET2* mutations on DNA methylation within this tissue. As discussed in section 3.1, mutations in IDH are thought to drive hyper-methylation via 2-HG mediated inhibition of *TET2,* so IDH-mutant CRC methylation array data should also be investigated.

3.3.2 – Pathogenic Mutations in *TET2* and IDH are Associated with Increased DNA Methylation

In order to characterise the effects of mutations in *TET2, IDH1* or *IDH2* on DNA methylation in CRC, DNA methylation array data from TCGA-COAD and TCGA-READ was collected from cancers with predicted pathogenic mutations in these genes (including missense mutations at $IDH1^{R132}$, $IDH2^{R140}$, $IDH2^{R172}$ and *TET2* loss-of-function mutations), alongside their WT counterparts. In order to correct for the differences between the methylation arrays used for these cancers, which included both Illumina 27K and 450K arrays, batch correction was performed to produce a common set of 20,618 methylation probes between all samples, allowing comparisons to be made between all samples of the datasets. The mean methylation β-value of these probes was calculated for each cancer and used as a measure of average DNA methylation. As presented in Figure 3.8a, cancers with *TET2* truncations (n = 7) had a significantly higher average DNA methylation than *TET2*-WT cancers (n = 519) (p = 0.014).

In addition to this, the TCGA-COAD and TCGA-READ datasets also included five cancers with pathogenic mutations in either *IDH1* or *IDH2.* These included two cancers with $IDH1^{R132C}$ mutations, one cancer with an $IDH1^{R132G}$ mutation, one cancer with an $IDH2^{R140W}$ mutation and one cancer with an $IDH2^{R172K}$ mutation. Similarly to cancers with *TET2* mutations, CRCs with pathogenic IDH mutations had increased average DNA methylation in comparison to their WT counterparts. As seen in Figure 3.8b, this increase in average DNA methylation was near-significant in cancers with *IDH1* mutations (p = 0.069). For cancers with pathogenic *IDH2* mutations, this average DNA methylation increase was significant (p = 0.034, Figure 3.8c). When *IDH1* and *IDH2* mutations were combined in Figure 3.8d, the increase in average DNA methylation in comparison to IDH-WT cancers was more significant than either *IDH1* or *IDH2* alone (p = 0.0058).

Overall, this data suggests that pathogenic mutations in either *TET2* or IDH in CRC are associated with an increase in genomic DNA methylation – therefore suggesting that mutations in these genes may be, at least in part, responsible for DNA hyper-methylation in CRC. In order to further investigate this association with DNA hyper-methylation, correlations between mutations in *TET2, IDH1* or *IDH2* with $CIMP^{+}$ disease in CRC should be investigated.

3.3.3 – Pathogenic Mutations in IDH are Correlated with the Development of CIMP in Colorectal Cancer

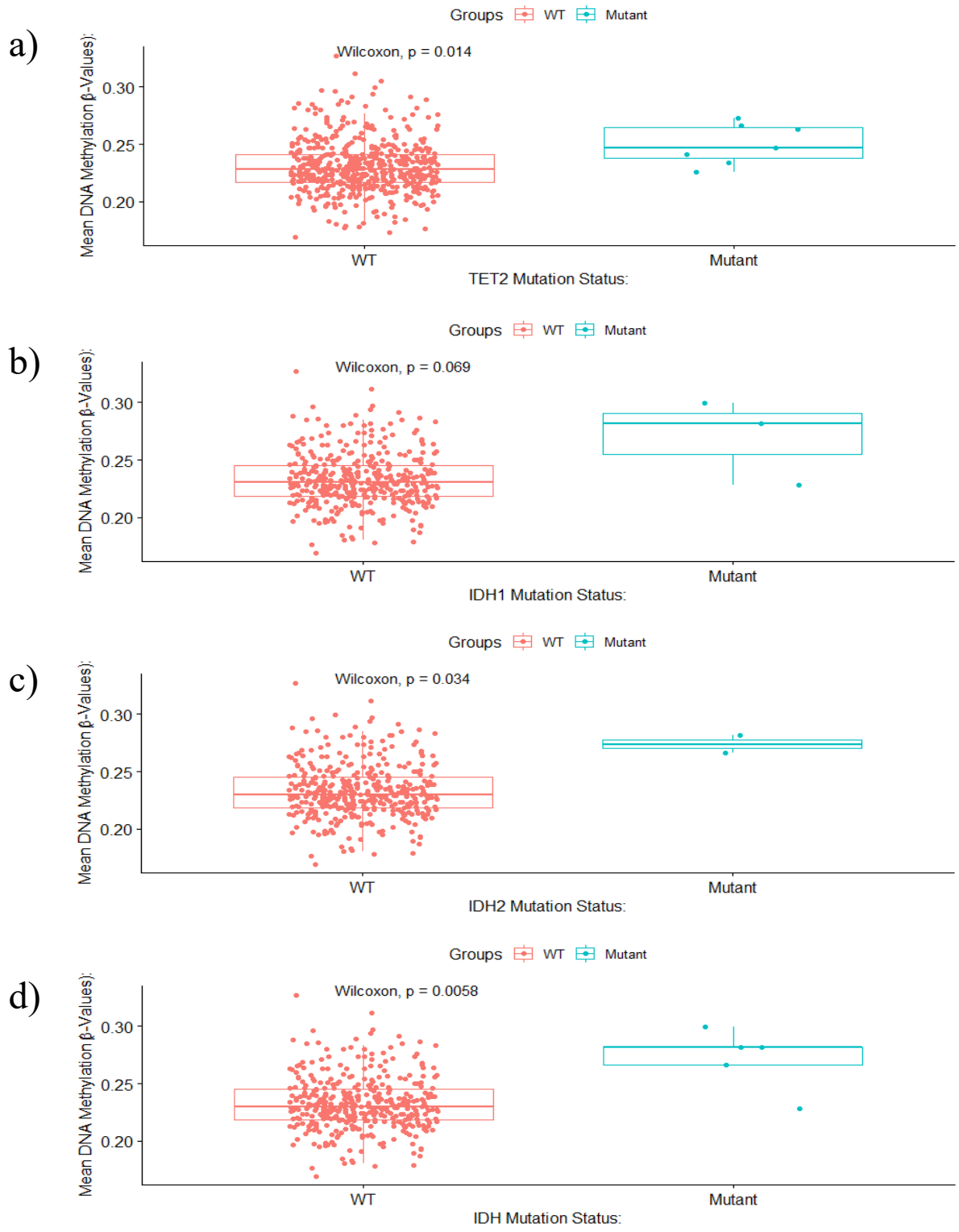


Figure 3.8 – Average DNA Methylation of *TET2*-Mutant or IDH-Mutant Colorectal Cancers: The average DNA methylation β-value of cancers taken from the TCGA-COAD and TCGA-READ datasets. Shown are the average DNA methylation β-values of either wild-type (WT – red) cancers or cancers with pathogenic mutations (blue) in *TET2* (a), *IDH1* (b), *IDH2* (c) or either *IDH1* or *IDH2* (d).

As discussed in Chapter I of this thesis, CIMP is characterised in CRC by the aberrant hyper-methylation of CpG sites, potentially resulting in the silencing of key tumour suppressor genes, thus promoting tumorigenesis. In order to characterise the effect of CIMP on DNA methylation in CRC, the CIMP statuses of 388 of the 526 cancers were available, allowing these cancers to be characterised as $CIMP^{-}$ (n = 234), $CIMP^{Low}$ (n = 86) or $CIMP^{High}$ (n = 68). Unsurprisingly, the average DNA methylation of both $CIMP^{Low}$ and $CIMP^{High}$ cancers was higher than that of $CIMP^{-}$ cancers ($p < 2.2 \times 10^{-16}$). Also as expected, the average DNA methylation of $CIMP^{High}$ cancers was also significantly higher than the average DNA methylation of $CIMP^{Low}$ cancers ($p = 1.1 \times 10^{-12}$). The average DNA methylation of each of these groups is presented in Figure 3.9.

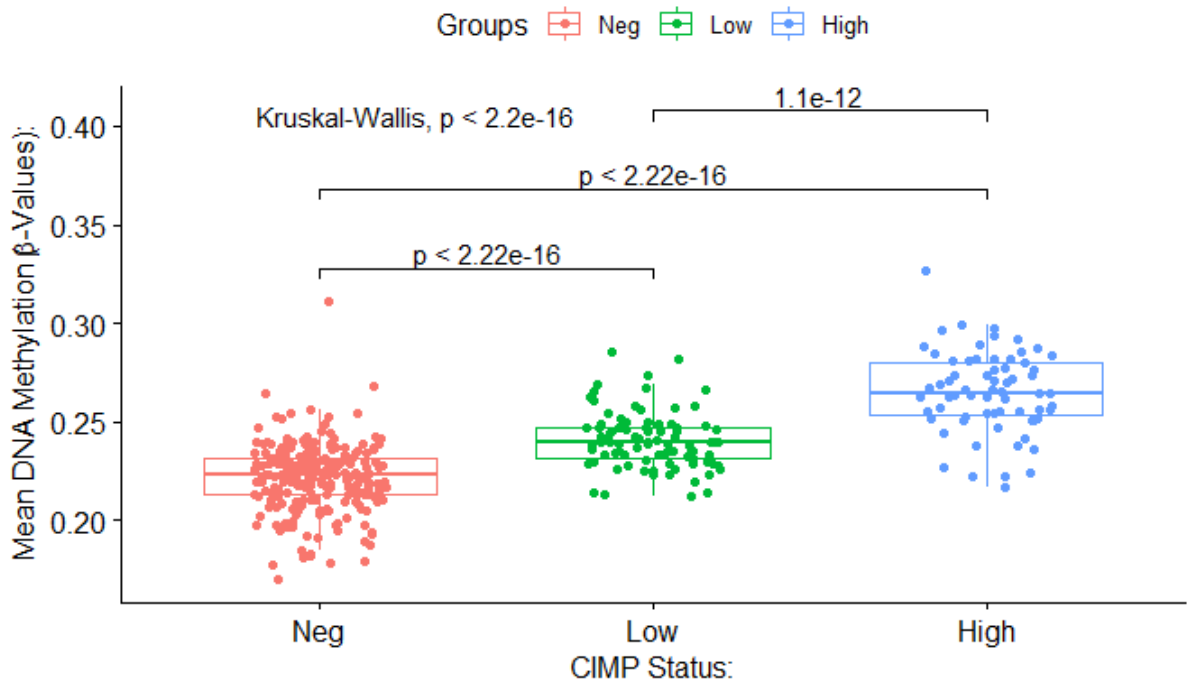


Figure 3.9 – The Average DNA Methylation of $CIMP^{+}$ Cancers: The average DNA methylation β-value of colorectal cancers from the TCGA-COAD and TCGA-READ domain. Shown are the average DNA methylation values of $CIMP^{-}$ (red), $CIMP^{Low}$ (green) and $CIMP^{High}$ (blue) cancers. Comparisons made between each of the groups were made using a Wilcox test and between all groups was calculated using a Kruskal-Wallis test.

Of the seven samples with truncating mutations in *TET2* described in section 3.3.2, two were classified as $CIMP^{-}$, two were classified as $CIMP^{High}$ and three did not have the relevant CIMP status data. Of the three cancers with pathogenic *IDH1* mutations, two were classified as $CIMP^{Low}$ and one as $CIMP^{High}$. Furthermore, of the two cancers with pathogenic mutations in *IDH2,* one was $CIMP^{Low}$ and the other described as $CIMP^{High}$. These classifications are summarised in Table 3.2. When compared to their WT counterparts, the proportion of *TET2*-mutant cancers classified as $CIMP^{High}$ was much higher (12.9% in *TET2*-WT cancers vs 28.6% of *TET2*-mutant cancers). The same trend could be seen in IDH-mutant cancers, where 100% of IDH-mutant cancers are classified as either $CIMP^{Low}$ or $CIMP^{High}$, compared to only 28.6% of IDH-WT cancers (see Table 3.2).

Cancer Type:	CIMP^{-}:	CIMP^{Low}:	CIMP^{High}:	No CIMP Data:
All TCGA-COAD / TCGA-READ (n = 526)	234 (44.5%)	86 (16.3%)	68 (12.9%)	138 (26.3%)
TET2-Mutant (n = 7)	2 (28.6%)	0 (0.0%)	2 (28.6%)	3 (42.8%)
TET2-WT (n = 519)	232 (44.7%)	86 (16.6%)	66 (12.7%)	135 (26%)
IDH1-Mutant (n = 3)	0 (0%)	2 (66.7%)	1 (33.3%)	0 (0%)
IDH1-WT (n = 523)	234 (44.7%)	84 (16.1%)	67 (12.8%)	138 (26.4%)
IDH2-Mutant (n = 2)	0 (0%)	1 (50%)	1 (50%)	0 (0%)
IDH2-WT (n = 524)	234 (44.7%)	85 (16.2%)	67 (12.8%)	138 (26.3%)
All IDH-Mutant (n = 5)	0 (0%)	3 (60%)	2 (40%)	0 (0%)
All IDH-WT (n = 521)	234 (44.9%)	83 (15.9%)	66 (12.7%)	138 (26.5%)

Table 3.2 – The CIMP Statuses of *TET2*-Mutant and IDH-Mutant Colorectal Cancers: The number and proportion of cancers from the TCGA-COAD and TCGA-READ domains described as either CIMP^{-}, CIMP^{Low}, CIMP^{High}. Also included are the number and proportion of cancers with no CIMP status assigned to them. Details on the CIMP statuses of cancers with pathogenic mutations in *TET2*, *IDH1* or *IDH2*, alongside their wild-type (WT) counterparts is also presented.

While these apparent correlations with CIMP are encouraging, it is possible that these observations are influenced by the relatively small sample size of cancers with pathogenic mutations in *TET2* or IDH (1.33% and 0.95% of all cancers respectively). Therefore, in order to more thoroughly investigate the correlation of pathogenic *TET2* and IDH mutations with CIMP^{+} cancer, additional CRC datasets with CIMP status data were obtained from the S:CORT consortium and DFCI. As discussed in Chapter I of this thesis, CIMP has been previously associated with MSI^{+} cancers, mutations in *BRAF* or *KRAS*, female patients, an older age of onset and localisation to the proximal colon (181,184). Therefore, in order to further strengthen the analysis, only cancers with all of this clinical information, plus tumour stage and the mutation status of *TET2, IDH1* and *IDH2* were included in the analysis. In total, 1,256 cancers had all the relevant information across the three datasets, 338 from TCGA-COAD and TCGA-READ, 472 from the S:CORT consortium and 446 from DFCI. From this data, a series of linear mixed-effects models were constructed using a binary CIMP status (0 for CIMP^{-} and 1 for CIMP^{+}) as the outcome variable with fixed effects of MSI status, patient diagnosis age, patient sex, tumour stage and tumour location (either proximal or distal colon). This allowed each candidate CIMP driver gene suggested to be correlated with CIMP to be studied in turn as another fixed effect of the model. To correct for any variation between studies, the study associated with each patient (either TCGA-COAD / TCGA-READ, S:CORT or DFCI) was input into the model as a random effect.

Firstly, *BRAF* and *KRAS* were used as positive controls given their previously established correlation with CIMP. Unsurprisingly, pathogenic mutations in *BRAF* (n = 161) were significantly correlated with CIMP^{+} disease (odds ratio = 1.522, 95% confidence intervals = 1.409 – 1.644, $p < 2.2 \times 10^{-16}$). Similarly, cancers with pathogenic *KRAS* mutations (n = 444) were also significantly correlated with CIMP^{+} disease (odds ratio = 1.086, 95% confidence intervals = 1.033 – 1.141, p = 0.00114). Interestingly, pathogenic mutations in *IDH1* (n = 7) were also significantly correlated with CIMP^{+} cancers (odds ratio = 1.931, 95% confidence intervals = 1.424 – 2.619, $p = 2.44 \times 10^{-5}$), suggesting that the associations with CIMP reported in the TCGA data alone may be accurate. For cancers with mutations in *TET2* and *IDH2*, only data from TCGA-COAD, TCGA-READ and DFCI were included in the model since neither of these genes were a part of the S:CORT sequencing panel. For cancers with truncations in *TET2* (n = 3) there was no significant correlation with CIMP^{+} disease, possibly

due to the low number of samples that were able to be included in the model due to a lack of the necessary clinical data for many *TET2*-mutant cancers. There was also a non-significant correlation between *IDH2*-mutant cancers (n = 2) and CIMP. Similarly to *TET2,* this was possibly a consequence of the low number of *IDH2*-mutant cancers. However, when *IDH1* and *IDH2* mutations were combined (n = 9), a more significant correlation with $CIMP^+$ cancer was identified than for either IDH gene alone (odds ratio = 1.826, 95% confidence intervals = 1.396 – 2.39, $p = 1.23 \times 10^{-5}$). The data for each model is presented in Table 3.3.

In summary, these data suggest that pathogenic mutations in *IDH1* and *IDH2* are significantly correlated with CIMP in CRC according to analysis of three separate datasets. There is also data supporting a correlation between *TET2* mutations and CIMP from TCGA-COAD and TCGA-READ data alone. However, the small number of CRCs with *TET2* mutations and a lack of necessary clinical co-variate data have made this correlation difficult to confirm across multiple datasets.

3.3.4 – $CIMP^+$ CRCs Present with Hyper-Methylation of CpG Islands & Bivalent Promoter Regions

The above data suggest that IDH-mutant and, to a lesser extent, *TET2*-mutant CRCs are correlated with $CIMP^+$ disease. As discussed in Chapter I of this thesis, $CIMP^+$ cancers are characterised by the hyper-methylation of CpG islands and gene promoters, often including those of key tumour suppressor genes (99). In order to compare the regions of hyper-methylation in $CIMP^+$ cancers compared to their $CIMP^-$ counterparts, probes which are significantly differentially methylated in $CIMP^+$ cancers compared to $CIMP^-$ cancers first needed to be identified. These probes could be identified by converting the β-values of each of the 20,618 probes to M-values and linear models were fitted for the M-values of each probe, comparing $CIMP^+$ cancers to $CIMP^-$ cancers. As seen in Figure 3.10a, $CIMP^+$ cancers presented with an increase in the number of probes with an M-value > 0, which corresponds to a β-value > 0.5. This indicates that a number of probes were hyper-methylated in $CIMP^+$ cancers compared to their WT counterparts. In addition to this, of the 9,020 probes with significantly altered methylation in $CIMP^+$ cancers ($p_{(BHC)} < 0.05$), 8.432 (93.48%) were hyper-methylated whereas only 588 (6.52%) were significantly hypo-methylated (see Figure 3.10b).

These differentially methylated probes in $CIMP^+$ cancers could be mapped to CpG islands, as defined by probe annotation data from Illumina. Of the 20,618 methylation probes used in previous analyses, 15,665 (75.98%) mapped to CpG island regions. Of the 8,432 significantly hyper-methylated probes ($p_{(BHC)} < 0.05$) in $CIMP^+$ CRCs compared to $CIMP^-$ cancers, 6,597 (78.24%) mapped to CpG island regions, which represented a significantly greater proportion than would be expected ($p_{(\chi^2)} < 0.00001$, Table 3.4). However, of the 588 hypo-methylated probes, only 183 (31.12%) mapped to CpG islands, which was significantly fewer than expected ($p_{\chi^2)} < 0.00001$, Table 3.4). This data indicates that hyper-methylated probes in $CIMP^+$ cancers localise to CpG islands, as has been previously reported in the literature – whereas hypo-methylated probes appear to not localise to these regions.

BRAF			
Variable:	**Odds Ratio:**	**95% Confidence Intervals:**	**$P_{(Odds\ Ratio)}$:**
BRAF Mutation	1.522	1.409 – 1.644	$< 2 \times 10^{-16}$
MSI^+	1.275	1.179 – 1.38	1.62×10^{-9}
Male Sex	1.00241	0.958 – 1.0491	0.917
Tumour Stage IV	0.989	0.936 – 1.044	0.681
Diagnosis Age	1.00454	1.0024 – 1.00668	3.35×10^{-5}
Proximal Colon	1.193	1.136 – 1.252	1.81×10^{-12}
KRAS			
Variable:	**Odds Ratio:**	**95% Confidence Intervals:**	**$P_{(Odds\ Ratio)}$:**
KRAS Mutation	1.086	1.033 – 1.141	0.00114
MSI^+	1.559	1.446 – 1.68	$< 2 \times 10^{-16}$
Male Sex	0.984	0.939 – 1.032	0.51312
Tumour Stage IV	1.011	0.955 – 1.07	0.71474
Diagnosis Age	1.00495	1.00273 – 1.00718	1.37×10^{-5}
Proximal Colon	1.239	1.179 – 1.303	$< 2 \times 10^{-16}$
TET2			
Variable:	**Odds Ratio:**	**95% Confidence Intervals:**	**$P_{(Odds\ Ratio)}$:**
TET2 Mutation	0.796	0.574 – 1.105	0.17266
MSI^+	1.638	1.523 – 1.763	$< 2 \times 10^{-16}$
Male Sex	0.971	0.92 – 1.024	0.27554
Tumour Stage IV	1.019	0.945 – 1.098	0.62517
Diagnosis Age	1.0033	1.00082 – 1.0058	0.009
Proximal Colon	1.211	1.146 – 1.28	2.54×10^{-11}
IDH1			
Variable:	**Odds Ratio:**	**95% Confidence Intervals:**	**$P_{(Odds\ Ratio)}$:**
IDH1 Mutation	1.931	1.424 – 2.619	2.44×10^{-5}
MSI^+	1.534	1.131 – 2.08	$< 2 \times 10^{-16}$
Male Sex	0.99	0.951 – 1.031	0.685
Tumour Stage IV	1.016	0.96 – 1.075	0.582
Diagnosis Age	1.0051	1.0028 – 1.0073	8.44×10^{-6}
Proximal Colon	1.253	1.193 – 1.317	$< 2 \times 10^{-16}$
IDH2			
Variable:	**Odds Ratio:**	**95% Confidence Intervals:**	**$P_{(Odds\ Ratio)}$:**
IDH2 Mutation	1.501	0.9 – 2.502	0.1202
MSI^+	1.624	1.51 – 1.746	$< 2 \times 10^{-16}$
Male Sex	0.969	0.919 – 1.022	0.24952
Tumour Stage IV	1.021	0.947 – 1.1	0.58705
Diagnosis Age	1.0034	1.00087 – 1.0059	0.00832
Proximal Colon	1.212	1.147 – 1.281	2.07×10^{-11}
IDH1 & IDH2			
Variable:	**Odds Ratio:**	**95% Confidence Intervals:**	**$P_{(Odds\ Ratio)}$:**
IDH1/2 Mutation	1.826	1.396 – 2.39	1.23×10^{-5}
MSI^+	1.529	1.42 – 1.646	$< 2 \times 10^{-16}$
Male Sex	0.99	0.945 – 1.038	0.682
Tumour Stage IV	1.016	0.96 – 1.075	0.576
Diagnosis Age	1.0049	1.0027 – 1.0071	1.57×10^{-5}
Proximal Colon	1.255	1.194 – 1.398	$< 2 \times 10^{-16}$

Table 3.3 – Linear Mixed-Effects Model of *TET2*-Mutant and IDH-Mutant Colorectal Cancers: Linear mixed-effects modelling using $CIMP^+$ as the outcome variable and MSI^+, male sex, tumour stage IV, diagnosis age and proximal colon localisation as fixed effects that can have an effect on the outcome variable. Also assessed are several candidate genes potentially correlated with $CIMP^+$ cancers that have been assessed in turn with each of the above co-variates, including *BRAF*, *KRAS*, *TET2*, *IDH1* and *IDH2*. Presented are the odds ratio of each co-variate in the model, as well as the 95% confidence intervals and p-value ($P_{(Odds\ Ratio)}$) associated with this odds ratio.

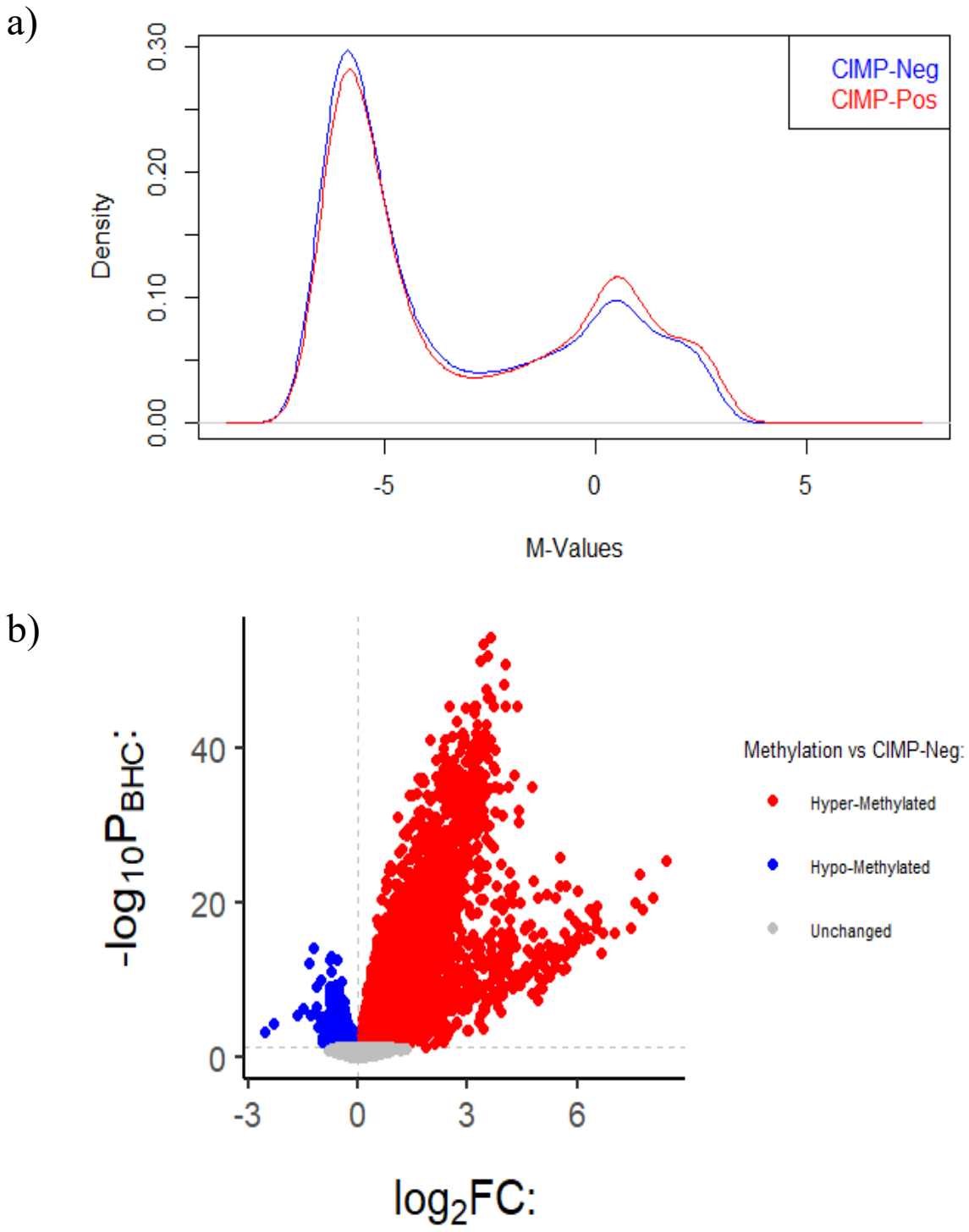


Figure 3.10 – CIMP$^+$ Colorectal Cancers Show Probe Hyper-Methylation Compared to CIMP$^-$ Cancers: (a) A density plot showing the distributions of M-values of the 20,618 methylation probes included the batch-corrected analysis of colorectal cancers from The Cancer Genome Atlas datasets. Shown are the M-value distributions of CIMP$^-$ (blue) and CIMP$^+$ (red) colorectal cancers. (b) A volcano plot comparing the $\log_2$ fold-change in the DNA methylation of each of the 20,618 probes and the associated $-\log_{10}$ Benjamini-Hochberg corrected p-value of this change ($p_{(BHC)}$). Shown on the plot are probes that are significantly hyper-methylated (red, $p_{(BHC)} < 0.05$), hypo-methylated (blue, $p_{(BHC)} < 0.05$) and unchanged (grey, $p_{(BHC)} \geq 0.05$).

In addition to CpG island regions, differentially methylated probes in CIMP$^+$ cancers could also be mapped to promoter regions. The locations of promoter regions in large intestine tissue was obtained from Ensembl (268) and, of the 20,618 methylation probes described above, 16,754 (81.26%) mapped to these promoter regions. Of the 8,432 hyper-methylated probes in CIMP$^+$ cancers, 6,741 (79.95%) mapped to promoter regions. Interestingly, this represented a significantly smaller proportion than would have been expected ($p_{(\chi^2)}$ = 0.00195, Table 3.4). Of the 588 significantly hypo-methylated probes in CIMP$^+$ cancers, 302 (51.36%) mapped to these bulk promoter regions, representing significantly less than would be expected ($p_{(\chi^2)} < 0.00001$, Table 3.4). This data indicates that differentially methylated probes in CIMP$^+$ cancers appear to not preferentially localise to bulk promoter regions.

This data goes against what has previously been reported in the literature, where CIMP$^+$ cancers are thought to show preferential hyper-methylation of CpG islands and the promoter regions of tumour suppressor genes. The study by Court & Arnaud suggested that bivalent promoter regions were the true regions of preferential hyper-methylation in CIMP$^+$ cancer (371). From this study, a map of bivalent promoter regions was obtained and it was found that 3,922 (19.02%) of the 20,618 methylation probes mapped to these bivalent promoter regions. Of the 8,432 hyper-methylated probes in CIMP$^+$ cancers, 2,715 (32.2%) probes mapped to these regions – which is significantly more than would be expected ($p_{(\chi^2)} <$ 0.00001, Table 3.4). Of the 588 significantly hypo-methylated probes, 39 (6.63%) mapped to bivalent promoter regions, which was significantly less than would be expected ($p_{(\chi^2)} <$ 0.00001, Table 3.4). This data indicates that hyper-methylated probes in CIMP$^+$ cancers map to bivalent promoter regions, as previously suggested by Court & Arnaud (371).

In order to further investigate this, a total of 2,046 of the significantly hyper-methylated probes in CIMP$^+$ cancers (24.26%) also presented with a $\log_2$(Fold Change) in methylation of > 1, indicating extensive hyper-methylation of these probes. Of these 2,046 extensively hyper-methylated probes, a total of 1,270 (62.07%) mapped to bivalent promoter regions, perhaps further indicating that bivalent promoter regions were the target of aberrant hyper-methylation in CIMP$^+$ cancer. As seen in Figure 3.11, the mean β-value of the extensively hyper-methylated probes inside bivalent promoter regions (n = 1,270) was significantly greater than the mean β-value of probes outside bivalent promoters (n = 776) in CIMP$^+$ cancers, according to a paired Wilcoxon test ($p < 2.2 \times 10^{-16}$).

Overall, it appears that hyper-methylated probes in CIMP$^+$ cancers preferentially map to CpG islands and bivalent promoter regions of the genome. Therefore, this may represent a mechanism by which hyper-methylation may drive tumorigenesis via the transcriptional silencing of key tumour suppressor genes. The data previously presented in this chapter suggest that *TET2*-mutant and IDH-mutant CRCs may be correlated with CIMP$^+$ disease. Given the role of *TET2* in the regulation of DNA methylation and its hypothesised inhibition in IDH-mutant cancers, it would be interesting to assess any differences between *TET2*-mutant and IDH-mutant CIMP$^+$ cancers and their WT CIMP$^+$ counterparts to see if they may represent a sub-class of CIMP$^+$ CRC with unique characteristics.

Hyper-Methylated Probes (n = 8,432):					
Genomic Feature:	**Total # Overlapping Probes (%):**	**Expected # Overlapping Hyper-Methylated Probes (%):**	**Observed # Overlapping Hyper-Methylated Probes (%):**	**χ^2:**	**$P_{(\chi^2)}$:**
CpG Islands	15,665 (75.98)	6,407 (75.98)	6,597 (78.24)	23.462	< 0.00001
All Promoters	16,754 (81.26)	6,852 (81.26)	6,741 (79.95)	9.596	0.00195
Bivalent Promoters	3,922 (19.02)	1,604 (19.02)	2,715 (32.2)	950.3	< 0.00001
Hypo-Methylated Probes (n = 588)					
Genomic Feature:	**Total # Overlapping Probes (%):**	**Expected # Overlapping Hypo-Methylated Probes (%):**	**Observed # Overlapping Hypo-Methylated Probes (%):**	**χ^2:**	**$P_{(\chi^2)}$:**
CpG Islands	15,665 (75.98)	447 (75.98)	183 (31.12)	650.217	< 0.00001
All Promoters	16,754 (81.26)	478 (81.26)	302 (51.36)	346.403	< 0.00001
Bivalent Promoters	3,922 (19.02)	112 (19.02)	39 (6.63)	58.776	< 0.00001

Table 3.4 – Genomic Features Overlapping with Differentially Methylated Probes in CIMP$^+$ Colorectal Cancers: A summary of the genomic regions that differentially methylated probes in CIMP$^+$ cancers compared to CIMP$^-$ cancers are suggested to preferentially localise to. Included are the number and proportion of the 20,618 methylation probes mapping to each genomic feature, the expected number of the 8,432 hyper-methylated or 588 hypo-methylated probes in CIMP$^+$ cancers that should map to the same genomic feature based on this proportion, the observed number of differentially methylated probes mapping to the feature, the chi-squared test statistic (χ^2) for the feature and its associated p-value ($p_{(\chi^2)}$). The genomic features investigated are CpG islands, promoter regions obtained from large intestine (233) and bivalent promoter regions obtained from Court & Arnaud (329).

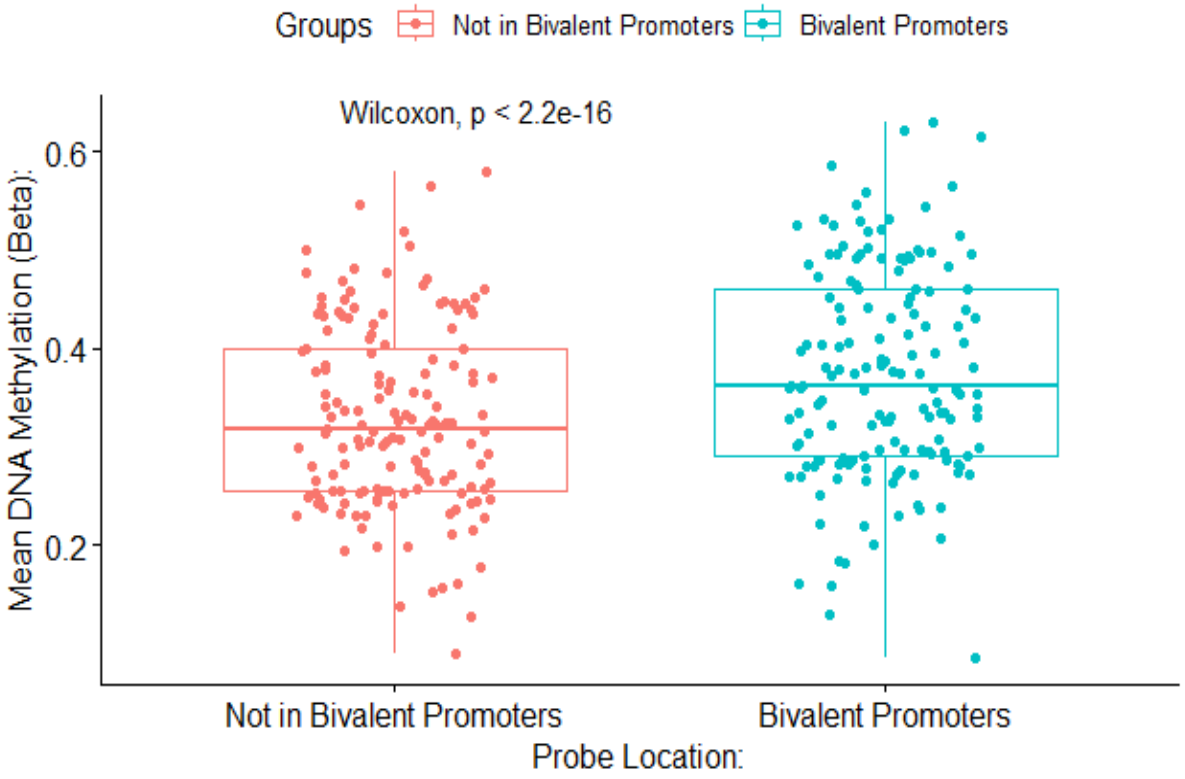


Figure 3.11 – CIMP$^+$ Colorectal Cancers Show Preferential Hyper-Methylation of Bivalent Promoter Regions: Boxplots indicating the mean DNA methylation β-value of extensively hyper-methylated probes (log_2(Fold Change) > 1, $p_{(BHC)}$ < 0.05) inside and outside of bivalent promoter regions in CIMP$^+$ colorectal cancers.

3.3.5 – *TET2*-Mutant & IDH-Mutant CIMP$^+$ Cancers Show Small Increases in DNA Methylation at Bivalent Promoters Compared to Other CIMP$^+$ Cancers

Sections 3.3.2 and 3.3.3 suggest that *TET2*-mutant and IDH-mutant CRCs may be associated with DNA hyper-methylation and CIMP$^+$ disease. Section 3.3.4 presented data suggesting that hyper-methylated probes in CIMP$^+$ cancers compared to CIMP$^-$ cancers preferentially mapped to CpG islands and bivalent promoter regions. However, so far little has been done to compare the characteristics of *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs to their WT CIMP$^+$ counterparts. To this end, the average DNA methylation of the 20,618 probes used above from *TET2*-mutant or IDH-mutant CIMP$^+$ CRCs were compared to the average DNA methylation of *TET2*-WT CIMP$^+$ or IDH-WT CIMP$^+$ CRCs respectively. As seen in Figure 3.12 *TET2*-mutant CIMP$^+$ CRCs appeared to present with higher average DNA methylation than *TET2*-WT CIMP$^+$ CRCs, however this increase was not significant (p = 0.25, Figure 3.12a). The same, albeit non-significant, increase in DNA methylation was also seen in

IDH1-mutant CIMP$^+$ CRCs ($p = 0.28$, Figure 3.12b). Interestingly, *IDH2*-mutant CIMP$^+$ CRCs presented with a near-significant increase in average DNA methylation compared to their WT CIMP$^+$ counterparts ($p = 0.099$, Figure 3.12c), which was also seen when *IDH1*-mutant and *IDH2*-mutant CIMP$^+$ CRCs were combined ($p = 0.058$, Figure 3.12d). This suggests that there may be some differences between *TET2*-mutant or IDH-mutant CIMP$^+$ CRCs and their WT CIMP$^+$ counterparts.

In order to investigate this in more detail, the mean probe β-values for the 2,046 extensively hyper-methylated probes were calculated for the *TET2*-mutant or IDH-mutant CIMP$^+$ cancers and compared to the mean probe β-value in *TET2*-WT or IDH-WT CIMP$^+$ cancers using a paired Wilcoxon test. As seen in Figure 3.13, the mean β-value of these extensively hyper-methylated probes in *TET2*-mutant and IDH-mutant cancers was significantly higher than in their WT CIMP$^+$ counterparts. As seen in Figure 3.13a, The mean β-value of these probes was significantly greater in *TET2*-mutant CIMP$^+$ cancers compared to *TET2*-WT CIMP$^+$ cancers ($p < 2.2 \times 10^{-16}$). The same trend was also apparent in the *IDH1*-mutant CIMP$^+$ CRCs, which also presented with significantly greater mean probe β-values than their WT CIMP$^+$ counterparts ($p = 1.6 \times 10^{-6}$, Figure 3.13b). *IDH2*-mutant CIMP$^+$ cancers (Figure 3.13c) also presented with a significantly increased mean β-value of these extensively hyper-methylated probes compared to their WT CIMP$^+$ counterparts ($p < 2.2 \times 10^{-16}$). When the *IDH1*-mutant and *IDH2*-mutant CIMP$^+$ cancers were combined (Figure 3.13d), the average β-value of these extensively hyper-methylated probes was significantly greater than that of IDH-WT CIMP$^+$ CRCs ($p < 2.2 \times 10^{-16}$). Overall, the data presented in Figure 3.12 and Figure 3.13 suggest that *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs may present with an even greater degree of hyper-methylation than their WT CIMP$^+$ counterparts.

In order to determine if this increase in methylation was a consequence of the clinical characteristics of the *TET2*-mutant and IDH-mutant CIMP$^+$ cancers, the diagnosis ages, MSI status and tumour mutation burden of *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs were compared to WT CIMP$^+$ CRCs, as both age and MSI have previously been shown to correlate with CIMP$^+$ disease (see Table 3.3). As seen in Figure 3.14, there were non-significant differences in the diagnosis ages of the *TET2*-mutant or IDH-mutant CIMP$^+$ CRCs compared to their WT CIMP$^+$ counterparts. As seen in Figure 3.14a, *TET2*-mutant CIMP$^+$ CRCs appear, on average, to be slightly older than *TET2*-WT CIMP$^+$ CRCs, although the difference was not significant ($p = 0.15$, Figure 3.14a). Conversely, *IDH1*-mutant cancers appeared to be slightly younger than *IDH1*-WT CIMP$^+$ cancers, however this again was not significant ($p = 0.43$, Figure 3.14b). CIMP$^+$ CRCs with pathogenic *IDH2* mutations were slightly older than their *IDH2*-WT counterparts, a difference that was nearly significant ($p = 0.082$, Figure 3.14c). When *IDH1*-mutant and *IDH2*-mutant cancers were combined (Figure 3.14d), the IDH-mutant cancers were older than the IDH-WT cancers, but this difference was non-significant ($p = 0.62$).

In order to assess if there was a significant difference in the MSI status of *TET2*-mutant or IDH-mutant CIMP$^+$ CRCs compared to WT CIMP$^+$ CRCs, a Fisher's Exact Test (FET) was performed, comparing the number of MSS and MSI$^+$ cancers of each genotype. Although the proportion of *TET2*-mutant CIMP$^+$ CRCs that were also MSI$^+$ ($n = 2/2$) was higher than *TET2*-WT CIMP$^+$ cancers ($n = 43/147$), this difference was not significant ($p_{(FET)} = 0.0898$). There was also no significant difference in the proportion of *IDH1*-mutant CIMP$^+$ CRCs that were MSI$^+$ ($n = 0/3$) compared to their *IDH1*-WT counterparts ($n = 45/146$, $p_{(FET)} = 0.554$). In

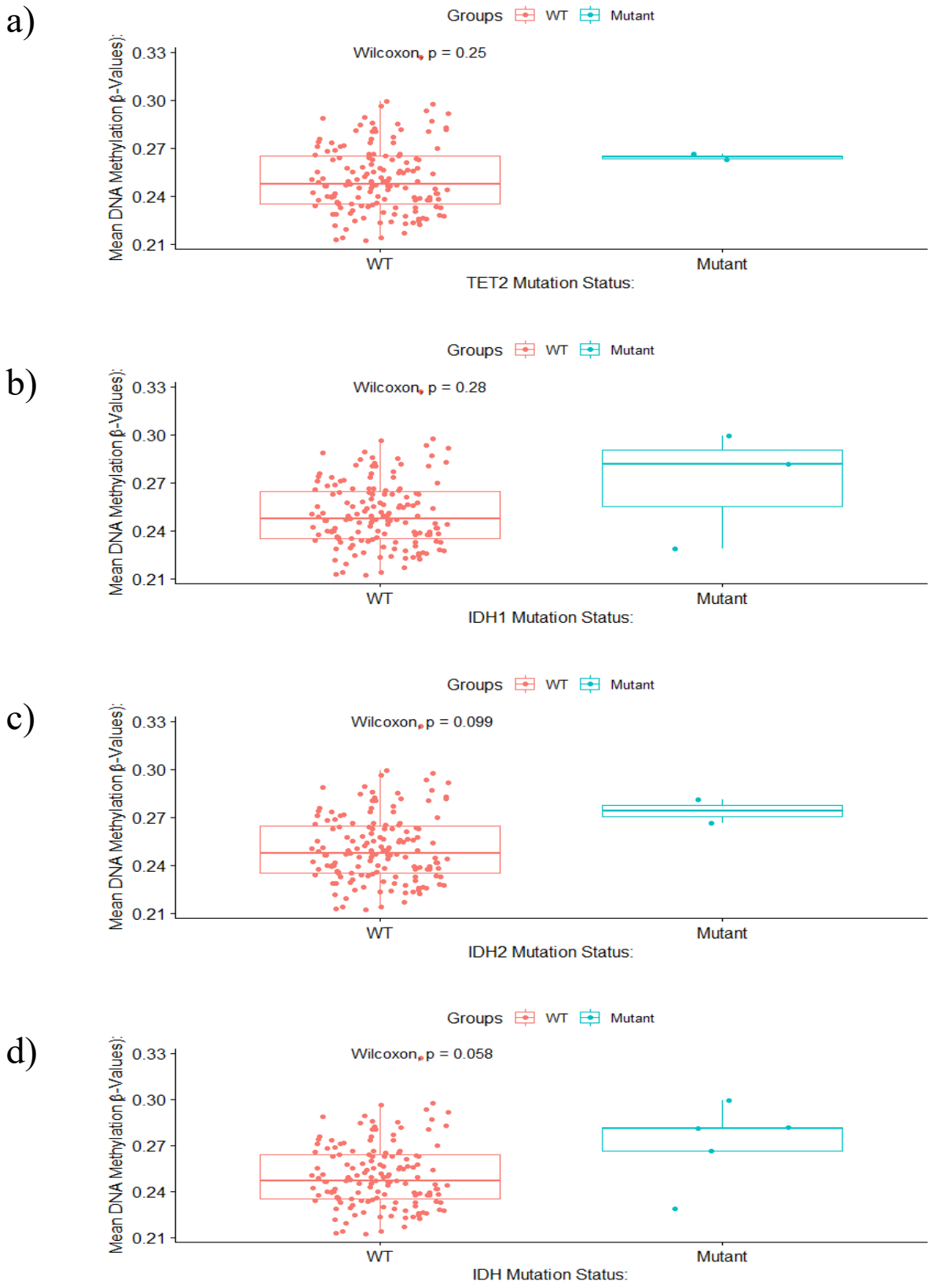


Figure 3.12 – Average DNA Methylation of *TET2*-Mutant or IDH-Mutant CIMP$^+$ Colorectal Cancers: The average DNA methylation β-value of cancers taken from the TCGA-COAD and TCGA-READ datasets. Shown are the average DNA methylation β-values of either wild-type (WT – red) CIMP$^+$ cancers or CIMP$^+$ cancers with pathogenic mutations (blue) in *TET2* (a), *IDH1* (b), *IDH2* (c) or either *IDH1* or *IDH2* (d).

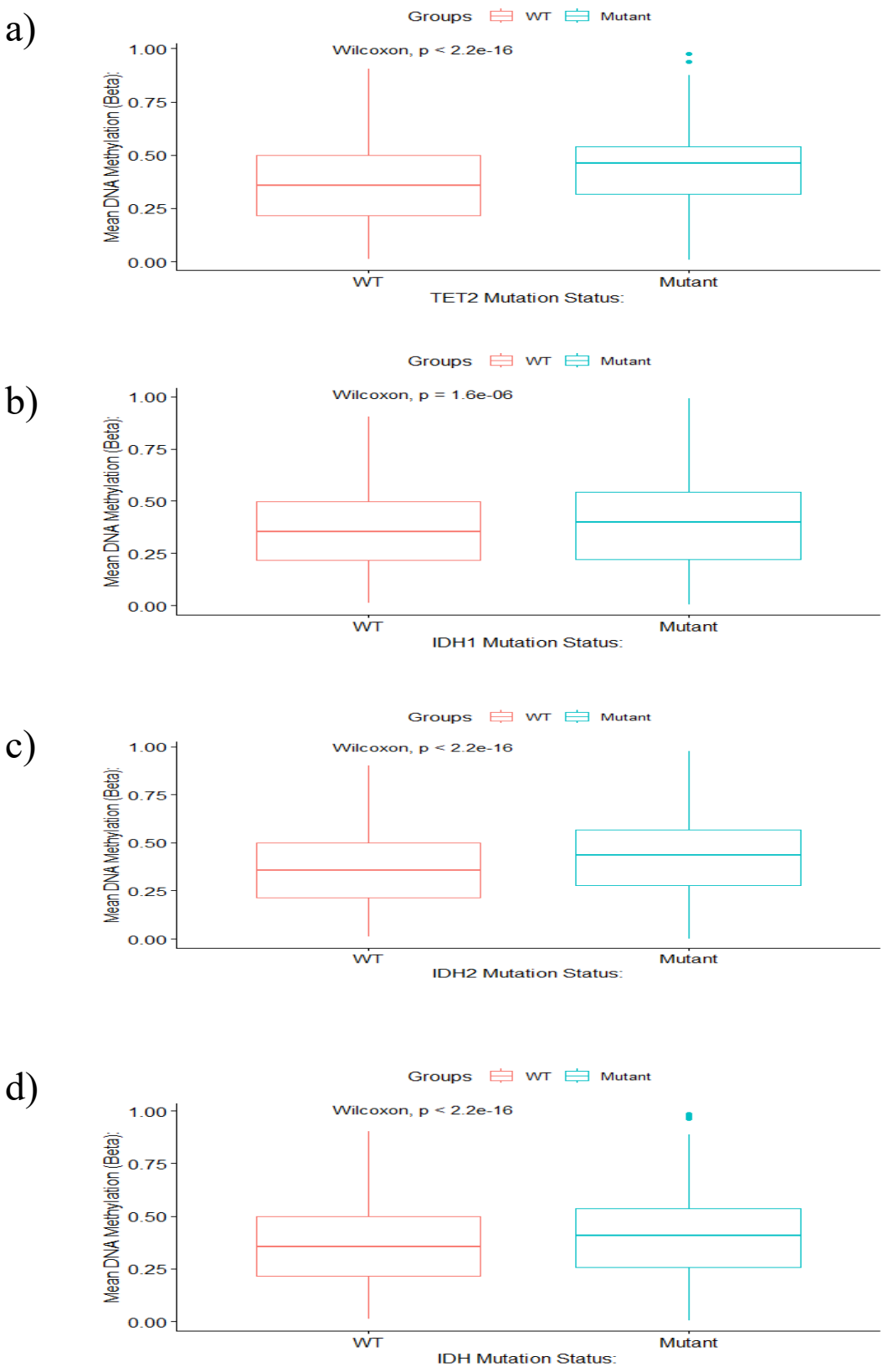


Figure 3.13 – Average DNA Methylation of Hyper-Methylated Probes in *TET2*-Mutant or IDH-Mutant CIMP$^+$ Colorectal Cancers: The average DNA methylation β-value of extensively hyper-methylated probes ($\log_2$(Fold Change >1, $p_{(BHC)} < 0.05$) in wild-type (WT – red) CIMP$^+$ colorectal cancers or CIMP$^+$ cancers with pathogenic mutations (mutant – blue) in *TET2* (a), *IDH1* (b), *IDH2* (c) or either *IDH1* or *IDH2* (d).

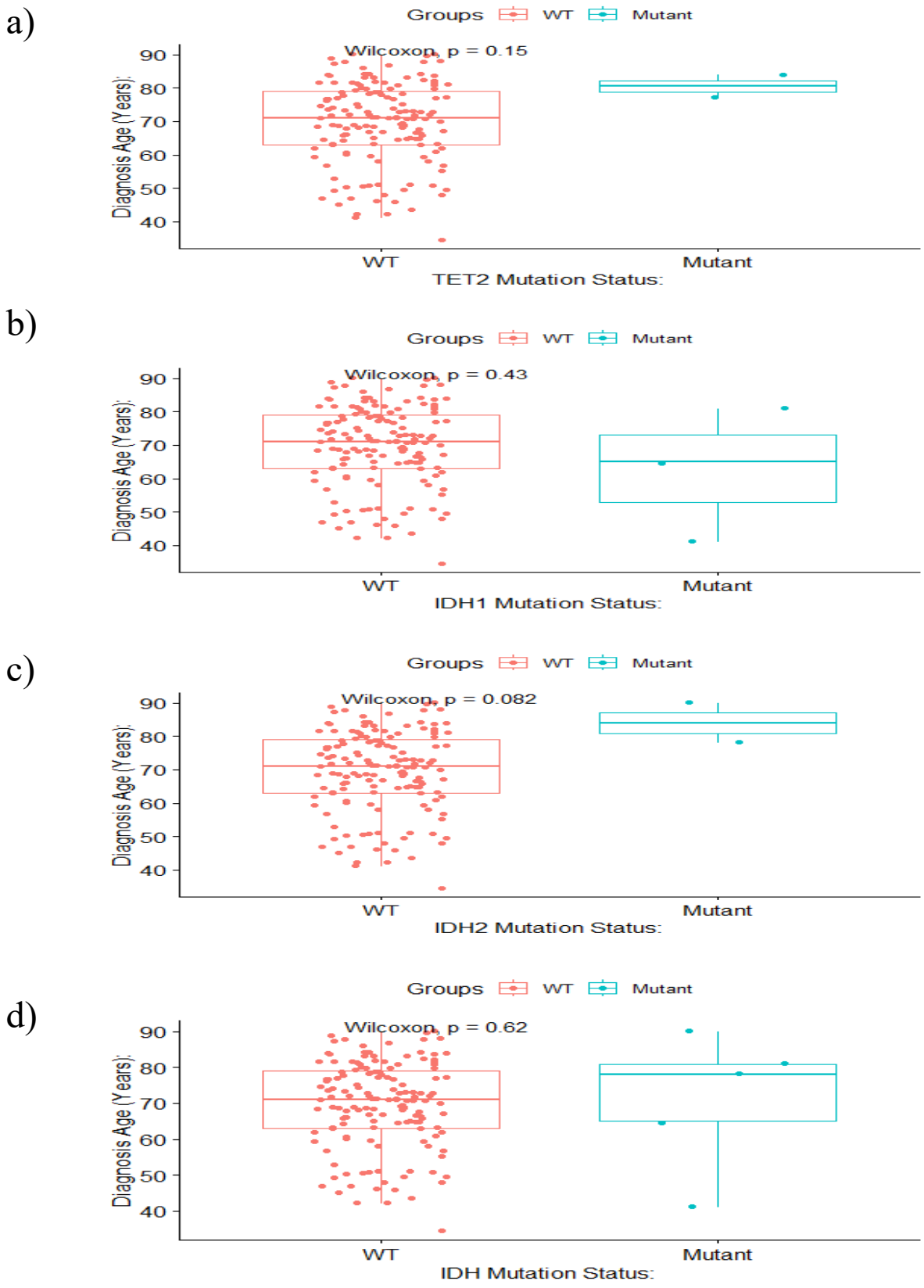


Figure 3.14 – Diagnosis Age of *TET2*-Mutant or IDH-Mutant CIMP$^+$ Colorectal Cancers: The diagnosis age of cancers taken from the TCGA-COAD and TCGA-READ datasets. Shown are the average diagnosis ages of either wild-type (WT – red) CIMP$^+$ cancers or CIMP$^+$ cancers with pathogenic mutations (blue) in *TET2* (a), *IDH1* (b), *IDH2* (c) or either *IDH1* or *IDH2* (d).

IDH2-mutant CIMP$^+$ CRCs, the proportion of MSI$^+$ cancers (n = 1/2) was not significantly different from *IDH2*-WT CIMP$^+$ CRCs (n = 44/147, $p_{(FET)} = 0.514$). When *IDH1* and *IDH2* mutations were combined, the proportion of IDH-mutant CIMP$^+$ CRCs (n = 1/5) was not significantly different from IDH-WT CIMP$^+$ cancers (n = 44/144, $p_{(FET)} = 1$).

When the total mutation burdens of *TET2*-mutant and IDH-mutant CRCs were compared to their WT CIMP$^+$ counterparts (Figure 3.15), no significant differences were identified. While *TET2*-mutant CIMP$^+$ cancers presented with a slightly higher average mutation burden than *TET2*-WT CIMP$^+$ cancers (Figure 3.15a), this difference was not significant (p = 0.095). The same trend was identified in *IDH1*-mutant CIMP$^+$ cancers (p = 0.66, Figure 3.15b), *IDH2*-mutant CIMP$^+$ cancers (p = 0.8, Figure 3.15c) and CIMP$^+$ CRCs with pathogenic mutations in either *IDH1* or *IDH2* (p = 0.61, Figure 3.15d). Overall, there appear to be no differences in the ages, MSI statuses or mutation burdens of these *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs compared to WT CIMP$^+$ cancers, therefore not explaining the slight increase in average DNA methylation identified in these cancers in Figure 3.12 and Figure 3.13.

As shown in Table 3.4 and Figure 3.11, CIMP$^+$ cancers appear to show preferential hyper-methylation of bivalent promoter regions. Therefore, the increase in DNA methylation in *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs compared to their WT CIMP$^+$ counterparts may be a consequence of a further increase in DNA methylation at these bivalent promoter regions in *TET2*-mutant and IDH-mutant CIMP$^+$ cancers. Therefore, the 2,046 extensively hyper-methylated probes were categorised as either within bivalent promoter regions (n = 1,270) or not within these regions (n = 776). The mean probe β-value for these types of probe could then be calculated for both *TET2*-mutant and IDH-mutant CIMP$^+$ cancers – as well as in their WT CIMP$^+$ counterparts. As seen in Figure 3.16, there is evidence to suggest that the mean β-value of these extensively hyper-methylated probes was highest in bivalent promoter regions of *TET2*-mutant or IDH-mutant CIMP$^+$ cancers. As seen in Figure 3.16a, the mean β-value of probes in bivalent promoter regions was significantly greater than the mean β-value of probes outside bivalent promoter regions in *TET2*-WT CIMP$^+$ cancers ($p = 8.6 \times 10^{-5}$). The mean β-value of *TET2*-mutant CIMP$^+$ cancers in these bivalent promoter regions was even higher than that if the *TET2*-WT CIMP$^+$ cancers. However, possibly due to the low number of *TET2*-mutant cancers, this difference was not significant (p = 0.35, Figure 3.16a). A similar trend was also observed in *IDH1*-mutant CIMP$^+$ cancers, where the mean β-value of probes in bivalent promoters was significantly higher than the mean β-value of probes outside these regions in *IDH1*-WT CIMP$^+$ CRCs (p = 0.0001, Figure 3.16b). The mean β-value of probes in bivalent promoters was greater still in *IDH1*-mutant CIMP$^+$ cancers, but this difference was not significant (p = 0.64, Figure 3.16b). Upon examination of *IDH2*-mutant CIMP$^+$ cancers, the mean β-value of probes inside bivalent promoter regions was significantly greater than that of probes outside of bivalent promoters in *IDH2*-WT CIMP$^+$ cancers ($p = 8.6 \times 10^{-5}$, Figure 3.16c). The mean β-value of probes in bivalent promoters in *IDH2*-mutant CIMP$^+$ cancers was even higher still than the mean β-value of its WT CIMP$^+$ counterparts – however this difference was not significant (p = 0.38, Figure 3.16c). When *IDH1*-mutant and *IDH2*-mutant CIMP$^+$ cancers were combined (Figure 3.16d), the mean β-value of probes in bivalent promoters was significantly greater than that of probes outside bivalent promoters in IDH-WT CIMP$^+$ cancers (p = 0.00011, Figure 3.16d). Interestingly, the mean β-value of probes in bivalent promoters of IDH-mutant CIMP$^+$ cancers was also significantly greater than that of probes outside bivalent promoters in IDH-WT CIMP$^+$ cancers (p = 0.04, Figure

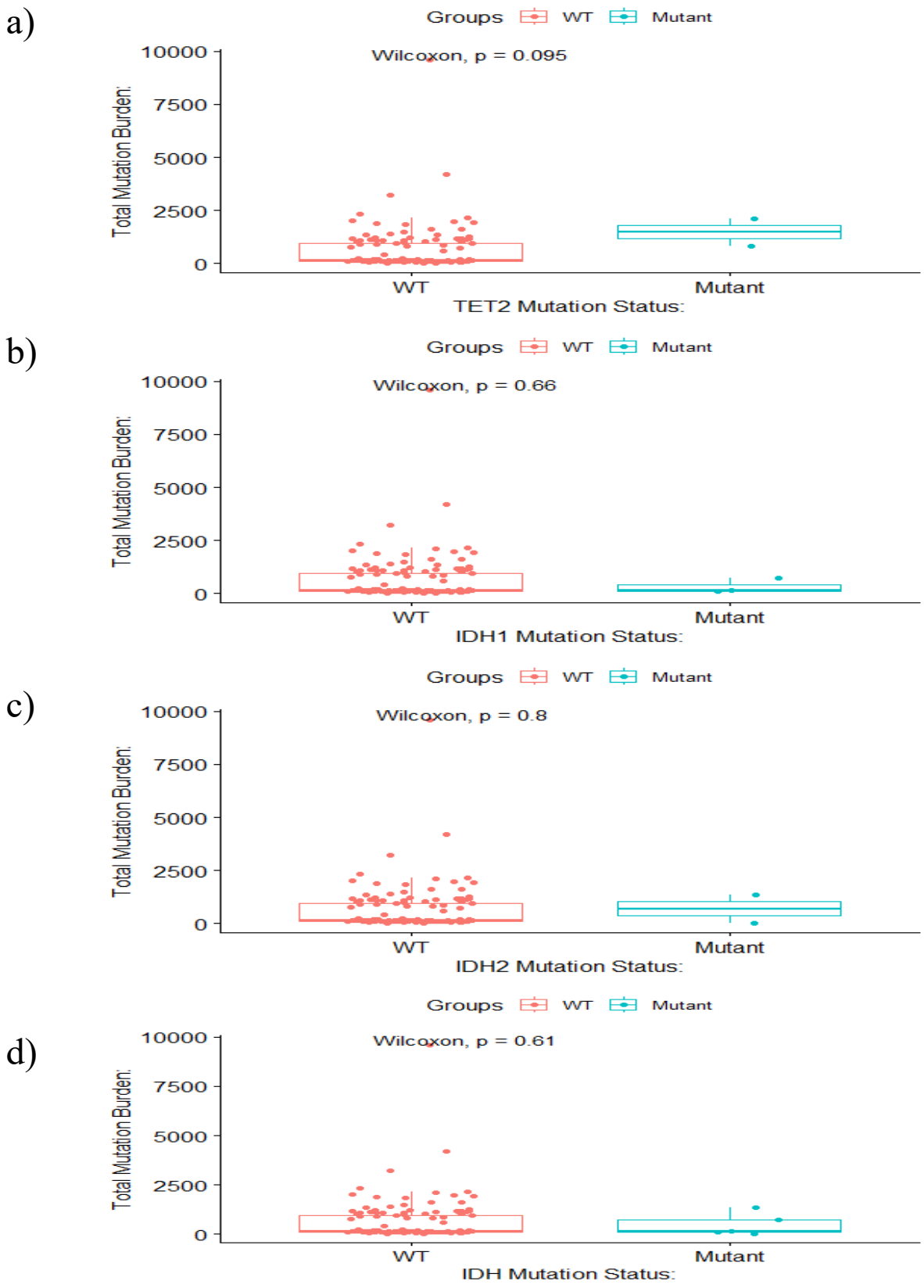
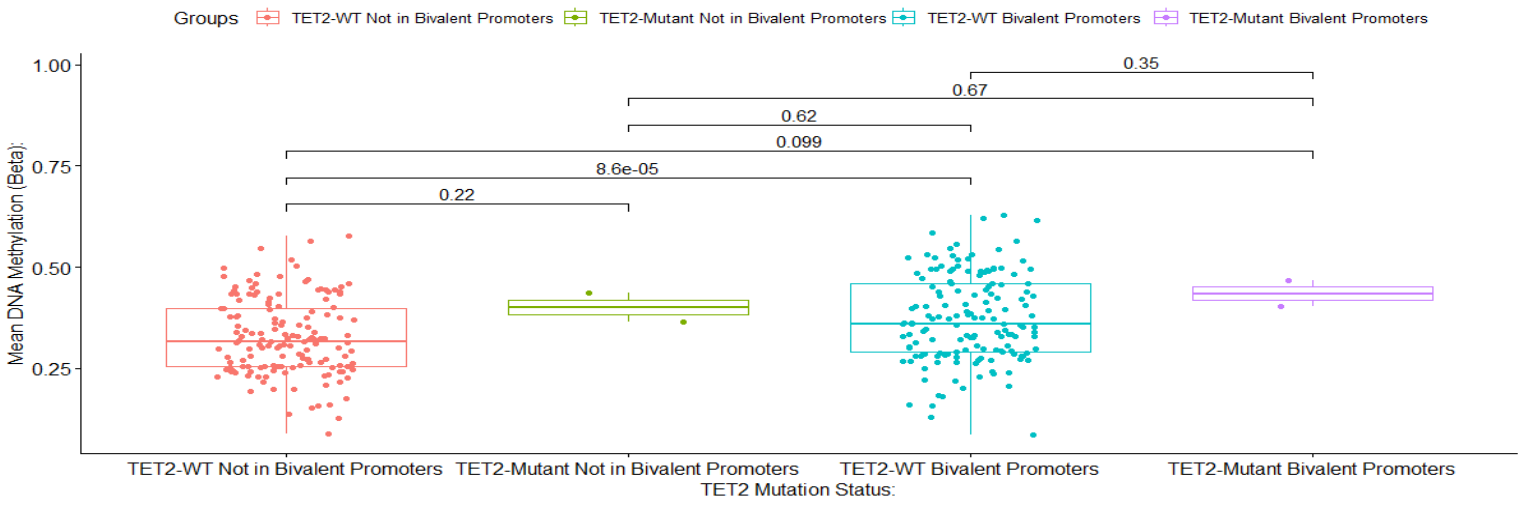
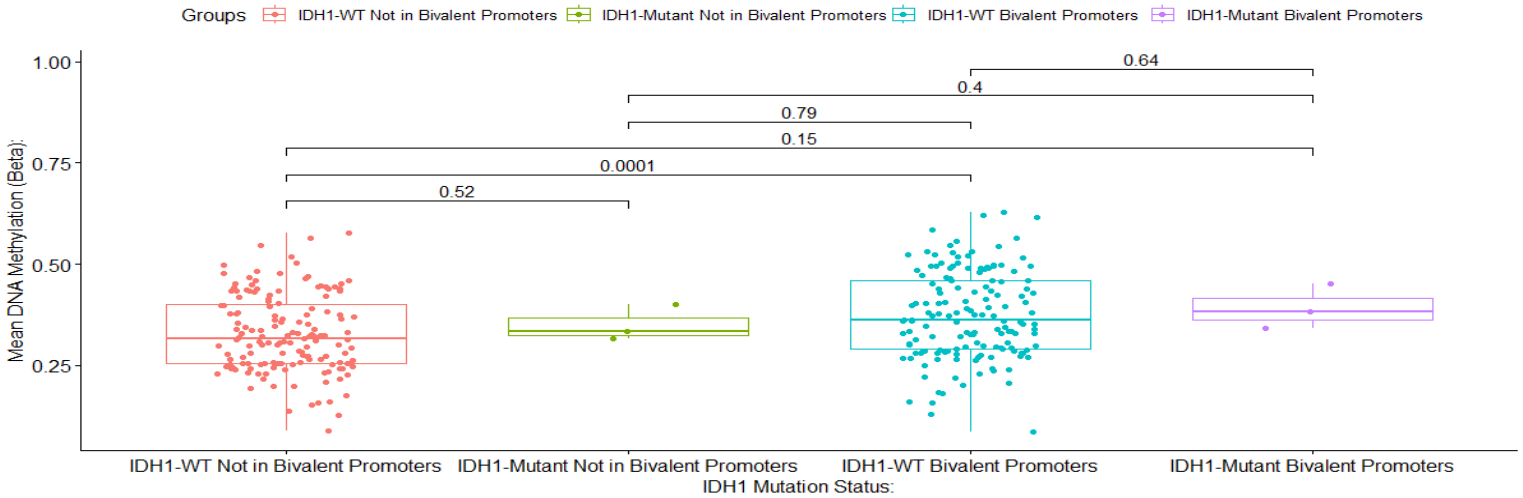


Figure 3.14 – Mutation Burden of *TET2*-Mutant or IDH-Mutant CIMP$^+$ Colorectal Cancers: The total mutation burden of cancers taken from the TCGA-COAD and TCGA-READ datasets. Shown are the mutation burdens of either wild-type (WT – red) CIMP$^+$ cancers or CIMP$^+$ cancers with pathogenic mutations (blue) in *TET2* (a), *IDH1* (b), *IDH2* (c) or either *IDH1* or *IDH2* (d).

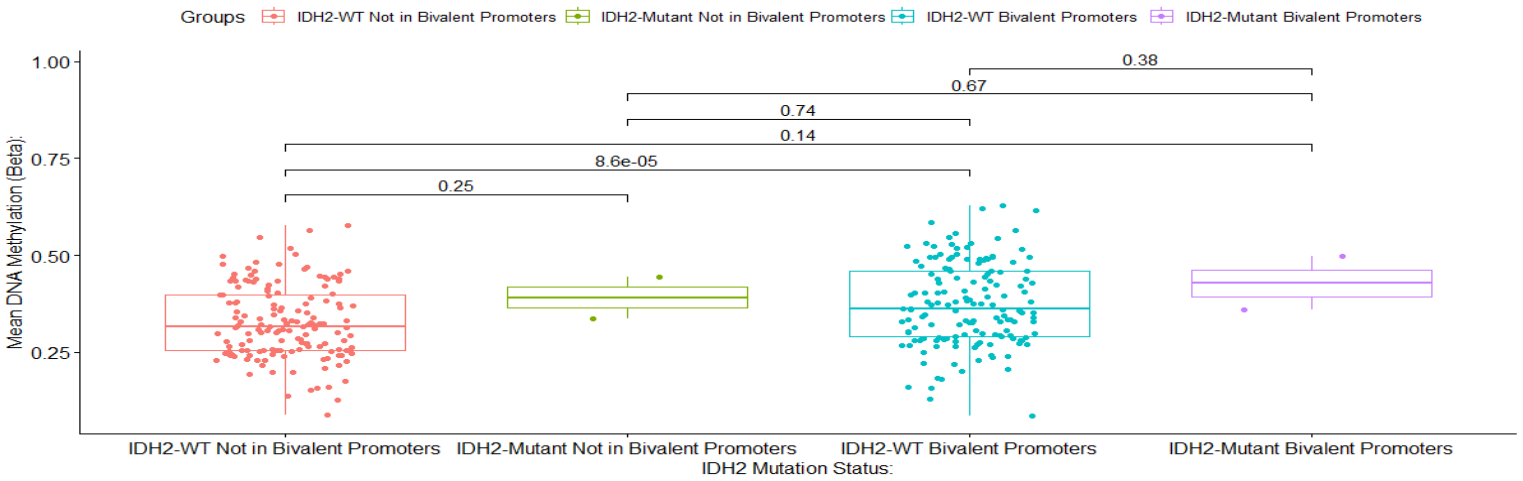
a)


Groups
TET2-WT Not in Bivalent Promoters
TET2-Mutant Not in Bivalent Promoters
TET2-WT Bivalent Promoters
TET2-Mutant Bivalent Promoters
Mean DNA Methylation (Beta):
1.00
0.75
0.50
0.25
0.35
0.67
0.62
0.099
8.6e-05
0.22
TET2-WT Not in Bivalent Promoters
TET2-Mutant Not in Bivalent Promoters
TET2-WT Bivalent Promoters
TET2-Mutant Bivalent Promoters
TET2 Mutation Status:

b)


Groups
IDH1-WT Not in Bivalent Promoters
IDH1-Mutant Not in Bivalent Promoters
IDH1-WT Bivalent Promoters
IDH1-Mutant Bivalent Promoters
Mean DNA Methylation (Beta):
1.00
0.75
0.50
0.25
0.64
0.4
0.79
0.15
0.0001
0.52
IDH1-WT Not in Bivalent Promoters
IDH1-Mutant Not in Bivalent Promoters
IDH1-WT Bivalent Promoters
IDH1-Mutant Bivalent Promoters
IDH1 Mutation Status:

c)



d)

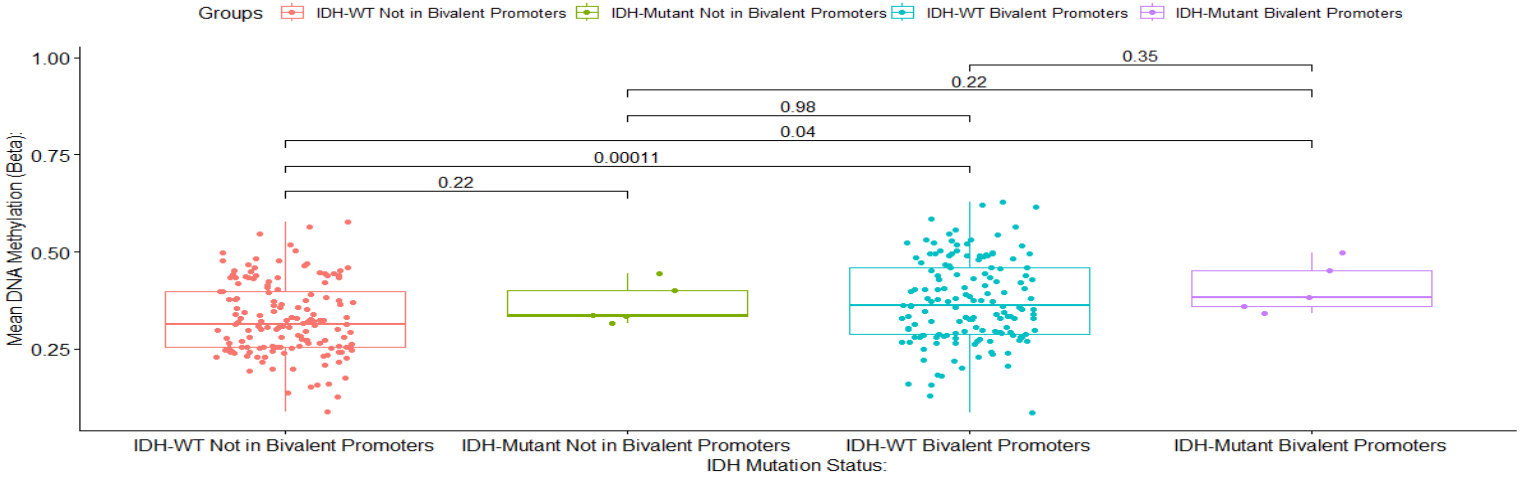


Figure 3.16 – Hyper-Methylated Probes within Bivalent Promoters in *TET2*-Mutant and IDH-Mutant CIMP$^+$ Colorectal Cancers: The mean probe β-values of the 2,046 extensively hyper-methylated ($\log_2$(Fold Change) > 1, $p_{(BHC)} < 0.05$) probes in *TET2*-mutant or IDH-mutant CIMP$^+$ colorectal cancer and their wild-type (WT) CIMP$^+$ counterparts. Shown are the mean probe β-values of probes located within (n = 1,270) and outside (n = 776) bivalent promoter regions in cancers with pathogenic mutations in *TET2* (a), *IDH1* (b), *IDH2* (c) or either *IDH1* or *IDH2* (d).

3.16d). The mean β-value of probes inside bivalent promoters was higher in IDH-mutant CIMP$^+$ cancers than their WT CIMP$^+$ counterparts, although this difference was not significant ($p = 0.35$, Figure 3.16d). Overall, this data indicates that probe hyper-methylation at bivalent promoters may be further increased in *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs compared to their WT CIMP$^+$ counterparts. However, the limited number of *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs present within the TCGA-COAD and TCGA-READ datasets meant that these increases were not significant.

3.3.6 – Significantly Hyper-Methylated Probes in *TET2*-Mutant & IDH-Mutant CIMP$^+$ Cancers Map to Candidate Tumour Suppressor Genes

Following the suggestion that *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs show greater enrichment of hyper-methylated probes at bivalent promoter regions than their WT CIMP$^+$ counterparts, the final stage of analysis was to characterise the similarities between of *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs. When *TET2*-mutant CIMP$^+$ CRCs ($n = 2$) were compared to *TET2*-WT CIMP$^-$ cancers ($n = 232$), a total of 613 significantly hyper-methylated probes were identified ($p_{(BHC)} < 0.05$). The five IDH-mutant CIMP$^+$ cancers presented with 2,756 significantly hyper-methylated probes compared to IDH-WT CIMP$^-$ cancers ($n = 234$) ($p_{(BHC)} < 0.05$). Of the 613 hyper-methylated probes in *TET2*-mutant cancers, 317 (51.71%) were also significantly hyper-methylated in IDH-mutant cancers.

These 317 significantly hyper-methylated probes in *TET2*-mutant and IDH-mutant CIMP$^+$ cancers were then mapped to their closest gene based on the associated probe annotation data provided by Illumina. As seen in Table 3.5, thirty-seven significantly hyper-methylated probes were located near the transcription start site of thirty-two genes which have been previously suggested to be down-regulated in cancer. Hyper-methylation of the region around the transcription start site of these genes would indicate epigenetic silencing in these *TET2*-mutant and IDH-mutant CIMP$^+$ cancers, perhaps driven by aberrant CIMP-mediated DNA hyper-methylation (373). Of the thirty-seven significantly hyper-methylated probes around the transcription start sites of these candidate tumour suppressor genes, twenty-two (59.46%) mapped to bivalent promoters. This suggests that hyper-methylation of these regions is not only a characteristic of CIMP$^+$ cancers, but may also represent a mechanism by which CIMP may drive tumorigenesis. Genes of particular interest in Table 3.5 include members of the Ras-associated domain family (RASSF), which have been suggested to be involved in cell cycle arrest and apoptosis (374). Aberrant promoter hyper-methylation, resulting in transcriptional silencing, of *RASSF5* and *RASSF1* have been previously reported in CRC, indicating that these are tumour suppressor genes and their epigenetic silencing in CIMP$^+$ cancers may have a role in driving tumorigenesis (374). Other candidate tumour suppressor genes with previous evidence of promoter hyper-methylation in CRC include deleted in lung and oesophageal cancer 1 (*DLEC1* – see Table 3.5). A number of studies have identified reduced *DLEC1* expression in CRC via promoter hyper-methylation, whereas over-expression of the *DLEC1* protein has been shown to reduce tumour clonogenicity in a number of cancer types (375,376).

Probe #:	Probe ID:	Chromosome #:	Position (hg19):	Within Bivalent Promoter?	Associated Gene:	Reference:
1	cg00282347	1	6,241,040	Yes	*CHD5*	(377)
2	cg06154570	1	40,105,764	Yes	*HEYL*	(378)
3	cg23698058	1	84,544,097	No	*PRKACB*	(379)
4	cg19452316	1	206,680,966	Yes	*RASSF5*	(374)
5	cg26491213	2	45,168,819	Yes	*SIX3*	(380)
6	cg05098471	2	66,662,163	Yes	*MEIS1*	(381)
7	cg00234616	2	74,740,572	Yes	*TLX2*	(382)
8	cg25361106	2	74,740,746	Yes	*TLX2*	(382)
9	cg23881725	3	38,080,642	Yes	*DLEC1*	(375)
10	cg21554552	3	50,378,425	Yes	*RASSF1*	(374)
11	cg20881888	3	50,383,079	Yes	*ZMYND10*	(383)
12	cg06156376	3	157,823,814	Yes	*SHOX2*	(384)
13	cg26359204	4	85,419,877	Yes	*NKX6-1*	(385)
14	cg02519806	4	119,810,036	No	*SYNPO2*	(386)
15	cg07786760	4	155,412,677	Yes	*DCHS2*	(387)
16	cg21504918	5	150,400,001	No	*GPX3*	(388)
17	cg21516478	5	150,400,328	No	*GPX3*	(388)
18	cg17820459	5	150,400,531	No	*GPX3*	(388)
19	cg09038885	5	159,343,148	Yes	*ADRA1B*	(389)
20	cg05874450	6	105,627,246	No	*POPDC3*	(390)
21	cg22030890	6	146,865,233	No	*RAB32*	(391)
22	cg04113075	6	146,865,487	No	*RAB32*	(391)
23	cg04528819	7	130,418,315	Yes	*KLF14*	(392)
24	cg09835543	9	93,405,134	Yes	*DIRAS2*	(393)
25	cg02676865	10	99,259,738	Yes	*UBTD1*	(394)
26	cg13055001	11	67,169,813	No	*PPP1CA*	(395)
27	cg22946876	12	26,111,213	No	*RASSF8*	(396)
28	cg25917510	12	54,402,888	No	*HOXC8*	(397)
29	cg05022306	12	54,403,254	No	*HOXC8*	(397)
30	cg02007463	12	56,101,543	No	*ITGA7*	(398)

Probe #:	Probe ID:	Chromosome #:	Position (hg19):	Within Bivalent Promoter?	Associated Gene:	Reference:
31	cg25995916	12	108,155,205	No	*PRDM4*	(399)
32	cg24777065	14	92,413,917	Yes	*FBLN5*	(400)
33	cg02831604	15	76,628,725	Yes	*ISL2*	(401)
34	cg06220235	16	11,349,023	No	*SOCS1*	(402)
35	cg16303453	16	66,879,016	Yes	*CA7*	(403)
36	cg20199629	17	48,207,508	Yes	*SAMD14*	(404)
37	cg09441152	18	77,712,293	Yes	*SLC66A2*	(405)

Table 3.5 – Cancer-Associated Genes Mapping to Hyper-Methylated Probes in Both *TET2*-Mutant & IDH-Mutant CIMP$^+$ Colorectal Cancers: Details of the candidate tumour suppressor genes associated with significantly hyper-methylated probes in *TET2*-mutant and IDH-mutant CIMP$^+$ colorectal cancers compared to their wild-type CIMP$^-$ counterparts. Included are the ID of the probe, the chromosome and position (hg19) of the probe, the gene associated with the probe and the reference number of the study implicating the down-regulation of this gene in cancer.

In summary, there was a large degree of overlap between significantly hyper-methylated probes in *TET2*-mutant and IDH-mutant CIMP^{+} CRCs compared to their respective WT CIMP^{-} counterparts. Of these common probes, approximately 10% were located around the transcription start site of candidate tumour suppressor genes which have been reported in a number of different cancers – including CRC (see Table 3.5). More than half of these probes were found within bivalent promoters, further exemplifying the functional significance of these regions in tumorigenesis in CIMP^{+} cancers.

3.4 – Discussion

Alterations to the DNA methylation profile of cells have been associated with tumorigenesis in several types of cancer, including CRC (see Figure 1.1). Cancers are frequently characterised by global DNA hypo-methylation accompanied by localised regions of DNA hyper-methylation – often acting to epigenetically silence tumour suppressor genes (406). Further characterisation of the role of DNA methylation in CRC identified a subset of cancers with extensive CpG site hyper-methylation (CIMP^{+} cancers) (211,215). As discussed in Chapter I of this thesis, both the underlying causes and potential clinical implications of CIMP^{+} cancer remains incompletely understood (215,217). Chapter II of this thesis suggested that reduced expression of *TET2* is associated with CRC disease predisposition. The role of this gene in driving active DNA de-methylation suggests that reduced expression of the gene may be associated with DNA hyper-methylation – which may consequently drive CRC development. Contrary to the hypothesis set out in section 3.1.5, knockout of *Tet2* in the intestinal tissues of mice did not result in tumorigenesis, with *Tet2*-knockout animals presenting with no obvious abnormalities in these tissues. Furthermore, IHC analysis revealed no significant reductions in 5-hmC or gain of 5-mC in *Tet2*-knockout animals compared to controls. However, IHC has limited utility in quantitatively assessing the changes in the DNA methylation profile of these animals, especially given the scarcity of 5-hmC in DNA. The study by Verma *et al.* identified localised, but not global, hyper-methylation in triple TET-knockout human embryonic stem cells – perhaps further indicating that methylation analysis via IHC may not provide an accurate assessment of the consequences of *Tet2* loss on DNA methylation (407). Alternatively, it could be that *Tet1* and *Tet3* activity is increased in *Tet2*-deficient animals to compensate for this loss, which may explain the lack of methylation changes in *Tet2*-knockout animals compared to controls.

The lack of a malignant phenotype in these *Tet2*-knockout animals, while disappointing, does not necessarily mean that *TET2* does not represent a novel cancer predisposition gene. As discussed in Chapter I of this thesis, the previous study by Kim *et al.* crossed the same *Tet2$^{fl/fl}$* animals used in this chapter with an *MMTV-cre*, producing a *Tet2* deficiency in the mammary tissue of these animals (298). While the *Tet2*-knockout animals presented with impaired differentiation of luminal cells and other breast tissue abnormalities, a further cross with *PyMT*-mutant animals was required for enhanced tumorigenesis to be observed. Therefore, it may be prudent in future analyses to abolish *Tet2* expression in animals predisposed to CRC development. This could be done by crossing *Tet2;Vil1-cre* animals with *APC$^{Min/+}$* animals, a line commonly used in the study of CRC development – as these animals often present with

intestinal tumours at three or four months of age (408,409). It is possible that, similarly to the study by Kim *et al., Tet2*-knockout in animals predisposed to CRC development may present with accelerated tumorigenesis compared to *Tet2*-proficient controls.

While the *Tet2*-knockout animals presented in this chapter presented with no malignant phenotype, there may have been some abnormalities within the intestinal compartment that could not be reported by the methods used in this chapter. Previously, knock-in of the $Idh1^{R132H}$ mutation into the intestinal compartment using *Vil1-cre* also resulted in no gross abnormalities or tumorigenesis – but did result in a modest but significant increase in both DNA methylation and 2-HG (unpublished data). Therefore, it would be worthwhile to perform more in-depth quantitative analysis of DNA methylation in *Tet2;Vil1-cre* animals in addition to the IHC performed for 5-hmC and 5-mC. Methylation array analysis may reveal more quantifiable changes in DNA methylation in *Tet2*-knockout animals compared to their *Tet2*-WT counterparts and also allow any changes in DNA methylation to be mapped to genomic features, for example bivalent promoter regions and the promoters of candidate tumour suppressor genes. The previous study by Kamdar *et al.* identified promoter hyper-methylation in seven candidate tumour suppressor genes of *TET2*-knockout prostate cancer cells, suggesting that *TET2* may drive disease progression via this mechanism (410). These (potentially more subtle) abnormalities were not investigated further as the aims of this chapter were to assess the role of *Tet2*-knockout in driving colorectal tumorigenesis – therefore representing a model of human disease. This meant that this chapter aimed to only assess gross changes in the intestinal tissues of *Tet2*-knockout animals.

It was also presented in this chapter that pathogenic mutations in *TET2* or IDH in human CRCs were correlated with $CIMP^+$ disease and presented with DNA hyper-methylation compared to their WT counterparts. This was perhaps not unexpected, given the well-described hyper-methylation of *TET2*-mutant and IDH-mutant AML (330). In the context of AML, mutations in *TET2* and IDH have been reported to be mutually exclusive, indicating that the common mechanism underlying pathogenesis in both *TET2*-mutant and IDH-mutant AML is the abolition of *TET2* activity (296,297,330). This mutual exclusivity was also apparent in the context of CRC, indicating that this same mechanism is the likely driving force underpinning the DNA hyper-methylation in these CRCs. However, the limited number *TET2*-mutant and IDH-mutant cancers makes this phenomenon less striking than in AML, where mutations in *TET2* or IDH are much more common. It was also shown in this chapter that *TET2*-mutant and IDH-mutant $CIMP^+$ CRCs show a slight, albeit non-significant increase in DNA methylation compared to WT $CIMP^+$ cancers, indicating that these cancers may even represent a distinct sub-class of $CIMP^+$ cancer, characterised by substantial DNA hyper-methylation. These *TET2*-mutant and IDH-mutant $CIMP^+$ CRCs presented with significantly increased DNA methylation of probes identified as extensively hyper-methylated in $CIMP^+$ cancers compared to their WT $CIMP^+$ counterparts – further indicating that these *TET2*-mutant and IDH-mutant cancers may be characterised by a greater degree of hyper-methylation than WT $CIMP^+$ cancers. These *TET2*-mutant and IDH-mutant $CIMP^+$ cancers showed no significant differences in age, MSI status or total mutation burden compared to their WT $CIMP^+$ counterparts, indicating that the difference in methylation was not a consequence of differing clinical characteristics.

As discussed in sections 3.1.1 and 3.1.4, *TET2* plays a critical role in catalysing active DNA de-methylation and 2-HG produced by mutant IDH is thought to inhibit *TET2* activity (296,297,309,330). Therefore, in both *TET2*-mutant and IDH-mutant cancers, DNA hyper-methylation should also be accompanied by a reduction in global 5-hmC as a consequence of this reduction in *TET2* activity (411). The study by Haffner *et al.* identified a reduction of 5-hmC in human colon cancer, despite the rarity of *TET2* or IDH mutations in this tumour type (333,411). Furthermore, the study by Uribe-Lewis *et al.* performed *TET2* knock-down in the CRC cell line HCT116 and observed a substantial decrease in 5-hmC (333). The same study also identified that up to 30% of promoters marked by 5-hmC are bivalent promoter regions and, similarly to the results presented in this chapter, suggested that bivalent promoter regions are vulnerable to hyper-methylation in cancer (333). However, the study suggested that 5-hmC may play a role in protecting these bivalent promoter regions from aberrant DNA hyper-methylation (333). The study by Verma *et al.* showed that triple knockout of TET genes from human embryonic stem cells resulted in hyper-methylation of bivalent promoters of developmental genes and suggested that the TET proteins were responsible for protecting bivalent promoter regions from aberrant hyper-methylation (407). Therefore, it could be that loss of 5-hmC at bivalent promoter regions – driven by either mutations in *TET2* or its inhibition by 2-HG – means these regions are no longer protected from aberrant DNA hyper-methylation, which may explain the (non-significant) increase in DNA methylation of the *TET2*-mutant and IDH-mutant cancers compared to their WT CIMP$^+$ counterparts (333,407).

A limitation of the results presented in this chapter relates to the number of *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs present within the TCGA-COAD and TCGA-READ domains. The limited number of cancers meant that the potential associations with CIMP and enhanced hyper-methylation at bivalent promoters were not always significant. As discussed in section 3.1, mutations in *TET2* and IDH are rare in CRC, meaning that in order to obtain methylation data from additional CRCs with *TET2* or IDH mutations a vastly expanded dataset would be required. This was addressed in part by using additional data from the S:CORT consortium and DFCI, which provided additional *TET2*-mutant and IDH-mutant cancers. However, *TET2* and *IDH2* were not part of the S:CORT panel of driver genes, meaning that the same analysis of methylation array data in these cancers was not possible. Similarly, methylation array data from DFCI was not available, meaning DNA methylation analysis of cancers from this cohort could not be performed.

The profiles of hyper-methylated probes in *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs showed substantial overlap, indicating similarities between the two types of CIMP$^+$ cancer. The study by Wilson *et al.* also observed a sizable overlap between the differentially methylated regions of *TET2*-mutant and IDH-mutant AMLs (412). These common hyper-methylated probes shared between *TET2*-mutant and IDH-mutant CRC included probes mapping to regions around the transcription start site of candidate tumour suppressor genes. Examples of these genes include glutathione peroxidase 3 (*GPX3*), which plays an important role in the protection of cells from oxidative DNA damage (413). Interestingly, promoter hyper-methylation of *GPX3* has been identified in a number of cancer types, including colon, gastric and ovarian cancer (388,413). The study by Barrett *et al.* studied *Gpx3*$^{-/-}$ mice and found that these animals presented with inflammation of the colonic tissue and an increased number of tumours compared to control animals (388). Other candidate tumour suppressor

genes included sterile alpha motif domain containing 14 (*SAMD14*), which has been reported to undergo promoter hyper-methylation and be associated with a poor prognosis in gastric cancer (404). In addition to their role in potentially driving colorectal tumorigenesis, the promoter hyper-methylation of some of these candidate tumour suppressor genes may have clinical utility as novel CRC biomarkers (384). The study by Bergheim *et al.* suggested that circulating cell-free DNA (ccfDNA) may be used to detect promoter hyper-methylation of candidate tumour suppressor genes, for example short stature homeobox 2 (*SHOX2*) (384). The study found a correlation between *SHOX2* promoter methylation in ccfDNA and tumour stage, nodal infiltration status and metastasis – indicating that this gene may have potential as a biomarker of disease progression (384). Similarly, the study by Naumov *et al.* suggested that hyper-methylation of CpG sites around the transcription start site of T-cell leukaemia homeobox 2 (*TLX2*) may also act as a biomarker for CRC (414). Therefore, it is possible that promoter hyper-methylation of other candidate tumour suppressor genes may also have the same potential as biomarkers for CRC pathogenesis. As presented in Chapter II of this thesis, inherited DNA variants may reduce the expression of *TET2* and predispose an individual to CRC. Given the potential of these candidate tumour suppressor genes as CRC biomarkers and the promoter hyper-methylation of these genes in *TET2*-mutant CRC, there may be clinical benefits in screening ccfDNA of at risk individuals with inherited variants affecting *TET2* expression. Earlier detection of promoter hyper-methylation of these candidate tumour suppressors may improve prognosis of affected individuals and serve as an effective method of monitoring disease progression and response to anti-cancer therapies.

In addition to the hyper-methylation of promoters and the region around the transcription start site driving gene silencing, aberrant DNA hyper-methylation may have alternative mechanisms of altering gene expression in cancer (415). The study by Arechederra *et al.* suggested that hyper-methylation of CpG islands within either the 5' untranslated region or gene bodies up-regulates gene expression, potentially driving the aberrant over-expression of oncogenes in hepatocellular carcinoma (415). While the analysis in this chapter has been restricted to assessing gene silencing via hyper-methylation of the promoter or first exon (373), it is possible that CRC pathogenesis may also be driven in *TET2*-mutant and IDH-mutant cancers via hyper-methylation within the gene bodies of oncogenes associated with CRC tumorigenesis. However, more than 80% of the DNA methylation probes used in this chapter mapped to promoters, making assessment of gene body hyper-methylation difficult – thus representing one of the limitations of this analysis. In subsequent analyses, DNA methylation profiling of *TET2*-mutant and IDH-mutant CRCs should be expanded to include more non-promoter elements of the genome, thus allowing an assessment of gene body hyper-methylation of CRC-associated oncogenes. Furthermore, this data could be combined with RNA-sequencing data from these cancers in order to observe the effect of promoter or gene body hyper-methylation on the expression of the associated tumour suppressor genes and oncogenes respectively.

DNA hyper-methylation has frequently been observed at bivalent promoter domains (416). The study by Court & Arnaud, similarly to the data presented in this chapter, suggest that hyper-methylation in $CIMP^+$ cancers map to bivalent promoter domains (371). Furthermore, nearly two-thirds of the extensively hyper-methylated probes in $CIMP^+$ cancers mapped to these bivalent promoter regions, further indicating their potential vulnerably to CIMP-

mediated hyper-methylation. Perhaps bivalent promoters represent regions of characteristic hyper-methylation in CIMP$^+$ cancers, potentially adding additional criteria that define CIMP$^+$ disease in addition to the small number of panel genes used to date (213,217). If bivalent promoters represent a preferred target for hyper-methylation in CIMP$^+$ cancers, the number of candidate genes driving tumour development in these cancers may be substantially reduced, potentially providing a clearer illustration of the mechanisms of tumorigenesis in CIMP$^+$ cancers. In addition to this, the characteristic DNA hyper-methylation of CIMP$^+$ cancers may also represent unique opportunities for therapeutic intervention. Novel approaches which target the epigenetic profiles of cancer cells have shown *in vitro* potential as future anti-cancer therapies (417). The study by Lee *et al.* suggests that the development of a novel histone de-acetylase inhibitor was able to reduce the proliferation of CRC cells and induce apoptosis *in vitro*, suggesting that epigenetic therapies may offer substantial clinical benefits (417). As discussed in Chapter I of this thesis, DNA de-methylating agents – such as 5-aza – show promising efficacy in reversing DNA hyper-methylation *in vitro*, but are hampered by toxicity which prevents them from being translated into the clinic (206,207). In recent years, a number of drug candidates that inhibit mutant IDH have been developed, with the goal of inhibiting the production of 2-HG (418–420). Examples of these include ivosidenib, which inhibits mutant *IDH1* (421). Ivosidenib was able to reduce serum 2-HG and induce differentiation in AML and can be safely combined with other anti-cancer therapies (421). In addition to this, mutant *IDH2* inhibitors have also been developed, including enasidenib – which has been shown to reduce 2-HG by up to 90%, reverse DNA methylation alterations and induce differentiation in AML (421). Therefore, it may be that these IDH inhibitors may also be of use in patients with IDH-mutant CRC, where reductions in 2-HG and the induction of differentiation may improve survival of these patients.

As discussed in Chapter I of this thesis, there is currently no universally-accepted model for the development of CIMP in cancer. Some publications have implicated mutations in *BRAF* as the causal factor driving a CIMP$^+$ tumour (422). However, the data presented in this chapter indicates a role for *TET2* and IDH mutations in CIMP development. While *TET2* and IDH mutations remain rare in CRCs and thus cannot explain the hyper-methylation observed in a substantial proportion of CIMP$^+$ cancers, it is possible that *TET2* and IDH mutations in CRC represent one potential mechanism for the development of CIMP$^+$ disease. In conclusion, the data presented in this chapter suggest that DNA hyper-methylation is potentially driven by mutations in the *TET2* or IDH genes, therefore abolishing active DNA de-methylation pathways. These *TET2*-mutant and IDH-mutant CIMP$^+$ CRCs may represent a sub-cluster of CIMP$^+$ cancers with unique features compared to their WT CIMP$^+$ counterparts and therefore may represent a plausible model for the origins of CIMP in a subset of CRCs.

Chapter IV – The Role of *MBD4* in Colorectal Tumorigenesis

4.1 – Background

4.1.1 – Mutation Signatures Underpin Cancer Development

Chapter I of this thesis highlighted the importance of both germline and somatic mutations in driving tumorigenesis in several cancer types, with mutations in key cancer-specific driver genes facilitating tumour development (4,5). There are many potential mechanisms underlying these mutations, both endogenous and exogenous – resulting in a single cancer potentially presenting with thousands of mutations (423). In recent years, the development of cost-effective next-generation sequencing technologies has deepened our understanding of the underlying mechanisms that underpin these driver gene mutations (424). As a result of this, the Catalogue of Somatic Mutations in Cancer (COSMIC) database has catalogued more than fifty somatic "mutation signatures" that describe patterns in single-base substitutions (SBS) and suggest the potential mechanisms by which these SBS occur (423,425,426). Examples of these mutation signatures include SBS3, which shows an extensive number of single-nucleotide variations (SNVs) across many trinucleotide contexts (see Figure 4.1a). It has been proposed that this mutation signature is driven by defective homologous recombination DNA repair mechanisms (425). Another common somatic mutation signature is SBS4, characterised by C → A mutations, particularly in the context C**C**A or C**C**C, with the middle base of the three being subject to mutation (see Figure 4.1b). This mutation signature has been attributed to tobacco smoking and is therefore often identified in lung cancer (425,427). The study by van den Heuvel *et al.* identified SBS4 in nearly a third of non-small cell lung cancer tumours and was found to be correlated with the earlier stages of disease development (427). While there has been progress in recent years in describing the underlying mechanisms of several somatic mutation signatures, the potential mechanism behind many of these signatures is yet to be described. In addition to SBS, double-base mutation signatures (DBS), insertion-deletion signatures (ID) and copy-number signatures (CN) have also been described by the COSMIC database (423,425).

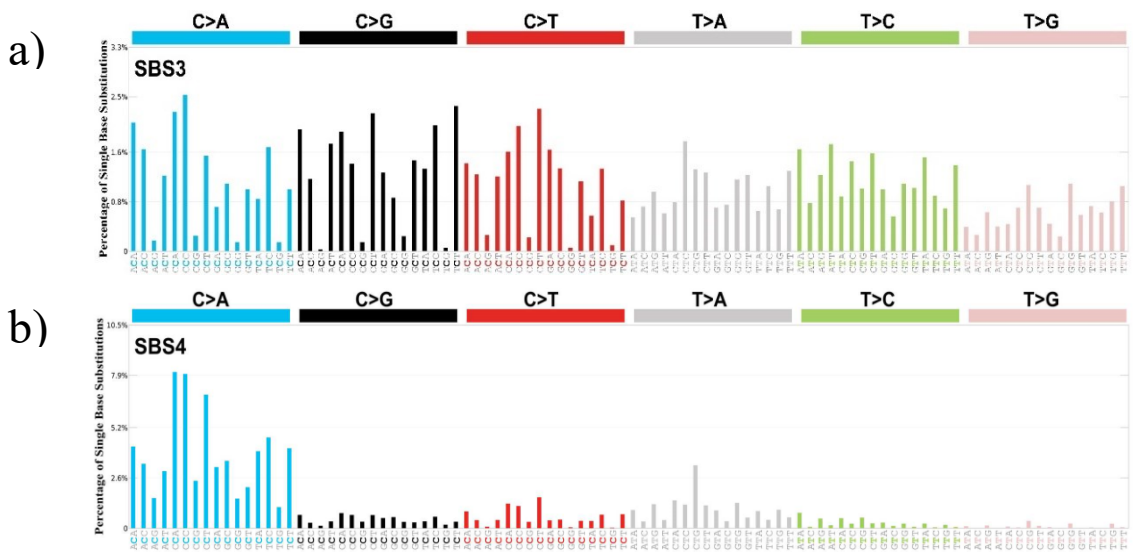


Figure 4.1 – Mutation Signatures in Cancer: Examples of two of the somatic mutation signatures identified in the COSMIC database. The mutational spectra, as well as the trinucleotide context for each signature is shown. Examples include SBS3 (a) – a signature associated with defective homologous recombination-based DNA repair and SBS4 (b) – associated with tobacco smoking. All signature plots obtained from the COSMIC database (https://cancer.sanger.ac.uk/signatures/sbs/).

There is evidence that mutation signatures can be detected in normal human tissues, including the colon, lung, endometrium and bladder, suggesting that mutation signatures may be functionally important in driving tumorigenesis and are not simply a reflection of passenger mutations that occur post-tumour development (423). However, the utility of mutation signatures in the clinic remains limited. While there are obvious cancer-associated risk factors associated with mutation signatures, for example tobacco smoking (SBS4) and ultra-violet light exposure (SBS7), the mutation signatures present within a cancer may also be of use in predicting response to specific chemotherapeutic agents (428). The meta-analysis by Litchfield *et al.* of over one thousand cancers treated with immune checkpoint inhibitor therapy identified a significant association between APOBEC-associated mutation signatures (e.g. SBS2) and response to immune checkpoint inhibition (429). The study by Chopra *et al.* identified forty-three cancers with defective homologous recombination (therefore presenting with the mutation signature SBS3) and reported improved prognosis after receiving treatment with poly-ADP ribose polymerase (*PARP*) inhibitors for two weeks pre-surgery in addition to subsequent chemotherapy (430). While mutation signatures may provide some insight into treatment response, there are also a number of mutation signatures associated with specific chemotherapeutics. For example, SBS11 is associated with temozolomide and SBS24 is associated with aflatoxin B1 – both DNA alkylating agents (431). These mutation signatures associated with chemotherapeutic agents may have clinical utility in predicting the long-term effects of treatment (432).

4.1.2 – SBS1

While many mutation signatures, including the examples provided above, have been associated with specific biological processes, exogenous DNA damage or anti-cancer therapies, two mutation signatures have been associated with patient age. These signatures are suggested to be the result of mutation accumulation throughout the life of an individual. Therefore, older individuals are thought to accumulate more mutations over their lifetimes in what is referred to as a "clock-like" pattern. These clock-like signatures are SBS1 and SBS5 (see Figure 4.2a and Figure 4.2b respectively) (426). SBS5 is characterised by a largely even distribution of SNVs across a number of trinucleotide contexts, with a particular enrichment for C → T and T → C mutations with a mechanism of action that is yet to be identified (426,433). However, unlike SBS5, SBS1 shows significant SNV enrichment in a restricted number of trinucleotide contexts (426). These mutations are almost exclusively C → T mutations in the context A**C**G, C**C**G, G**C**G and T**C**G – therefore representing C → T mutations at CpG sites (425,426). It has been suggested that the underlying mechanism associated with SBS1 is the failure to repair the spontaneous deamination of 5-mC to

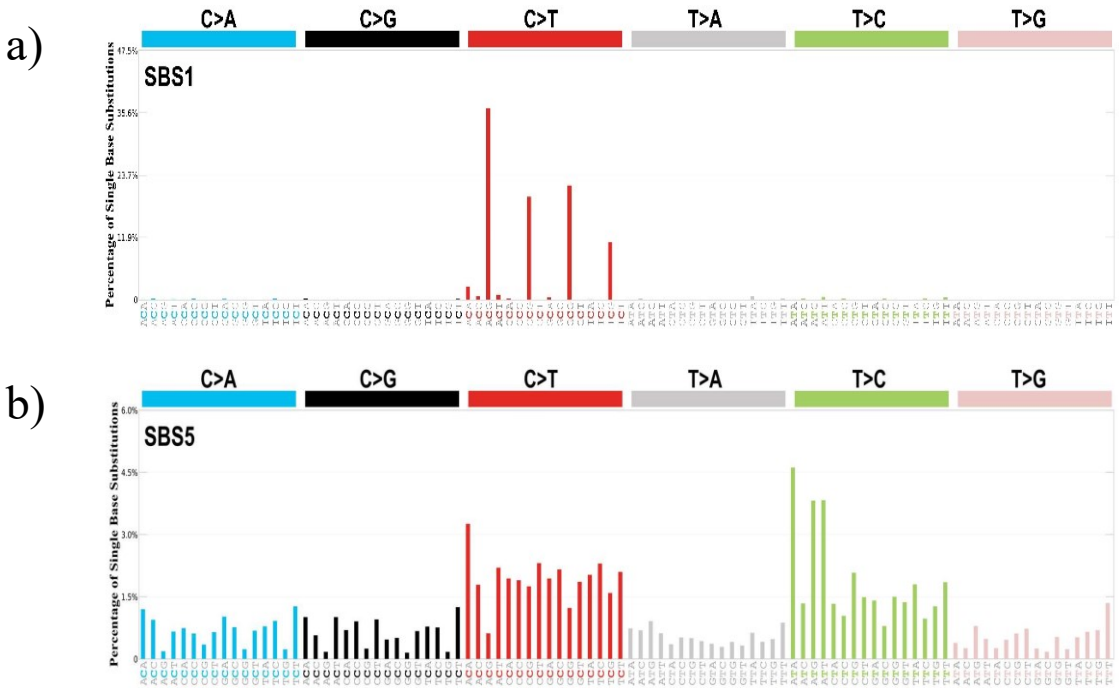


Figure 4.2 – Clock-Like Mutation Signatures: The two age-dependent, or "clock-like", mutation signatures described by COSMIC in cancer. Shown are the mutation spectra, trinucleotide context and frequency of mutation associated with SBS1 (a) and SBS5 (b). All signature plots obtained from the COSMIC database (https://cancer.sanger.ac.uk/signatures/sbs/).

thymine, thus resulting in a T:G DNA mismatch which is propagated into a C → T mutation following DNA replication (434,435).

Spontaneous deamination of both methylated and non-methylated cytosine is a common occurrence, with the deamination of 5-mC and non-methylated cytosine resulting in thymine and uracil respectively (435). The deamination of 5-mC is far more common than the deamination of non-methylated cytosine, occurring approximately five times more frequently, equating to approximately twelve deamination events per genome per day (434,436). Furthermore, single-stranded DNA is thought to be at 100-fold greater risk of spontaneous deamination than double-stranded DNA (437,438). The deamination of 5-mC to thymine results in a T:G mispairing in the DNA which is normally repaired by thymine DNA glycosylase (*TDG*) or methyl-CpG binding domain 4 (*MBD4*) acting in concert with BER proteins (312,436). However, failure to repair this T:G mismatch results in a C → T mutation at CpG sites, which represent the most common SNV in tumorigenesis (312,425,436). Interestingly, the number of SBS1-associated C → T mutations are not consistent across human tissues but instead are correlated with mitosis (425,426). Therefore, the number of C → T mutations are highest in tissues with rapid cellular turnover, possibly as a consequence of the increased rate of DNA replication exposing single-stranded DNA more frequently – increasing the likelihood of spontaneous deamination of 5-mC (425,439,440). Intestinal tissues have the fastest rate of cellular turnover of any tissue, indicating that these tissues may be at greatest risk of C → T mutagenesis at CpG sites (425,439,440).

The study of Lee-Six *et al.* identified SBS1 in more than 85% of investigated intestinal crypts and was concluded to be a ubiquitous mutation signature of the normal colon (433). The rate of SBS1-associated mutation accumulation was found to vary across the colon, ranging from 12.7 mutations per year in the descending/sigmoid colon to 16.8 mutations per year in the ascending colon and caecum (433). In the study by Alexandrov *et al.,* SBS1 could be identified in several types of cancer, including CRC, ovarian cancer, pancreatic cancer and lung adenocarcinoma, further highlighting its relevance in human disease (425).

4.1.3 – *MBD4* Mutations in Human Cancer

As discussed above, spontaneous deaminations of 5-mC are repaired by *TDG* or *MBD4* working in unison with the BER pathway and the failure to repair these deaminations is associated with SBS1 (312,435,441). The *MBD4*, also referred to as *MED1*, protein consists of an N-terminal methylcytosine-binding domain and a C-terminal glycosylase domain which displays homology with bacterial endonucleases involved in BER (442). Following deamination of 5-mC to thymine, *MBD4* excises the mispaired thymine from DNA via the cleavage of an N-glycosidic bond – therefore catalysing the first step in the BER pathway (312,443,444). It has also been demonstrated that *MBD4* is also able to excise mispaired uracil, produced by the deamination of non-methylated cytosine, albeit with a lower affinity than for mispaired thymine (312). This excision generates an abasic site in the DNA which is subsequently repaired by BER proteins, via DNA polymerase-β mediated filling of the abasic site and subsequent re-ligation of the DNA (312,445). A summary of *MBD4*-mediated repair of spontaneous deamination events is presented in Figure 4.3.

Germline mutations in *MBD4* have been identified in multiple cancers, including AML, CRC and uveal melanoma (446–448). Somatic mutations in *MBD4* are more common in MSI^+ cancers due to a number of highly repetitive regions in the protein-coding sequence of the gene – including an $A_{(10)}$ tract between codons 310-313 and three $A_{(6)}$ sequences between codons 247-248, 280-282 and 327-329 (449). An illustration of the structure of the *MBD4* protein is provided in Figure 4.4. In CRC, it is estimated that between 20-43% of primary MSI^+ carcinomas harbour inactivating mutations in *MBD4*, whereas the study by Riccio *et al.* reported no mutations in *MBD4* at these repetitive regions in MSS cancers, indicating that these regions may be mutation hotspots in MSI^+ cancers only – which is plausible given the reported increase in frameshift mutations at repetitive regions in MSI^+ cancers (96,449). Unsurprisingly, patients with germline mutations in *MBD4* presented with an increased number of C → T mutations at CpG sites (221,446,447). The study by Sanders *et al.* identified three AML patients with germline biallelic inactivation of *MBD4,* one with a single homozygous in-frame deletion ($MBD4^{H567del}$) and two siblings with compound heterozygous mutations, including a $MBD4^{V314Rfs*13}$ frameshift mutation and a splice acceptor mutation which disrupted exon seven (446). These cancers had a remarkably high mutation burden which was more than thirty-fold higher than other AMLs from the TCGA cohort (446). Strikingly, more than 95% of mutations in these three AMLs were C → T mutations at CpG sites, which were not attributed to old age as all three patients were less than thirty-five years old (446). Two of the three patients also presented with a number of colorectal polyps, suggesting that *MBD4* mutations may also drive colorectal polyposis (446).

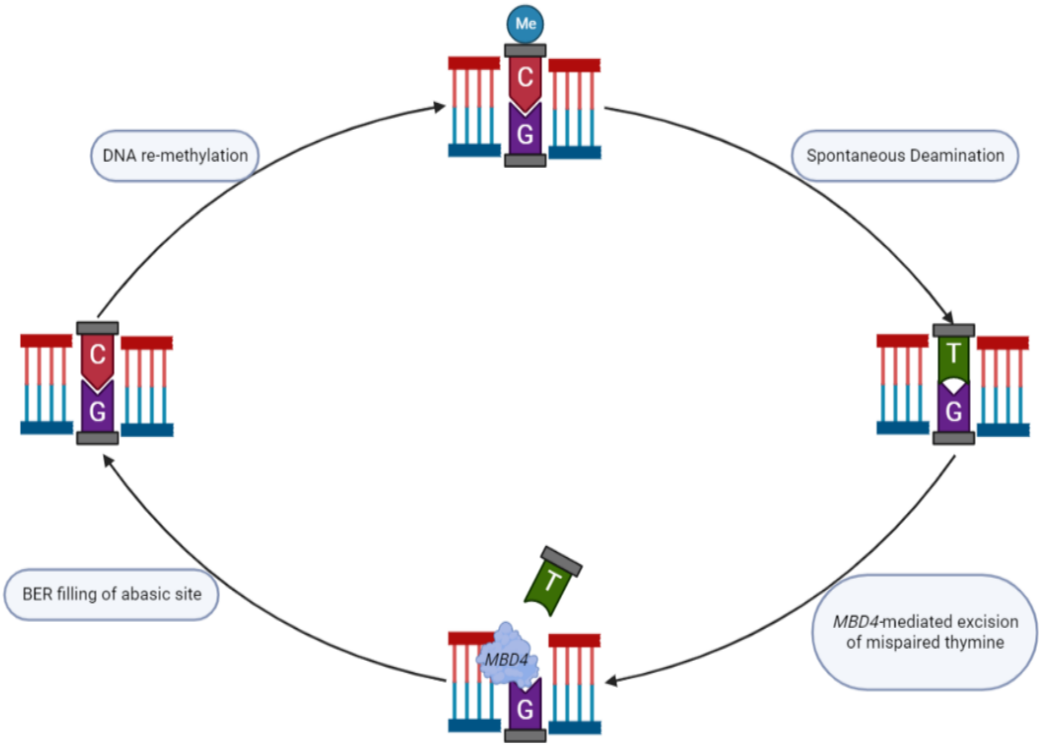


Figure 4.3 – *MBD4*-Mediated Repair of Spontaneous Deamination: A schematic illustration of the mechanism by which spontaneous deamination of 5-methylcytosine (5-mC) is repaired. Spontaneous deamination of 5-mC (top) to thymine results in a T:G mismatch in the DNA (right) that is recognised by the DNA glycosylase methyl-CpG binding domain 4 (*MBD4*). *MBD4* then catalyses the cleavage of an N-glycosidic bind in the DNA, excising the mispaired thymine and subsequently creating an abasic site (bottom). This abasic site is then filled with the correct base via components of the base excision repair (BER) pathway, including DNA polymerase-β and DNA ligase (left). The repaired T:G mismatch is then re-methylated. Created with BioRender.com (https://app.biorender.com/).

Figure 4.4 – The Structure of the *MBD4* Gene: A diagrammatic illustration of the structure of the methyl-CpG binding domain 4 (*MBD4*) gene. Shown are the N-terminal methyl-CpG binding domain (MBD) and the C-terminal glycosylase domain (GD), as well as the two $A_{(6)}$ and one $A_{(10)}$ tract within the protein-coding sequence. Created with BioRender.com (https://app.biorender.com/).

The report by Tanakaya *et al.* identified a forty-two year old woman with a germline heterozygous *MBD4*^{Q73*} mutation and subsequent somatic LoH (447). This patient presented with approximately thirty colonic polyps despite having no previous family history of CRC – and no germline mutations in any previously identified CRC predisposition gene (447). Similarly to the AML patients reported by Sanders *et al.,* the patient presented with an increased number of C → T mutations at CpG sites (447). Furthermore, the study by Palles *et al.* performed whole-genome or whole-exome sequencing on 309 individuals from 198 likely unrelated families with multiple cases of colorectal adenomas or familial CRC and a further 1,611 individuals in replication cohorts with family history of colorectal adenomas, familial CRC or CRC with other tumours (221). Three individuals presented with homozygous germline truncations in *MBD4,* one deletion resulting in *MBD4*^{S205Tfs*9} and two duplications resulting in *MBD4*^{E314Rfs*13} (221). One individual presented with approximately sixty colorectal adenomas at thirty-six years of age, increasing to seventy following panproctocolectomy at age forty-seven (221). The same patient was subsequently diagnosed with MDS seven months post-surgery, which progressed to AML three months later (221). The second of these patients presented with thirty-three colorectal adenomas following panproctocolectomy at age fifty-three and was later diagnosed with uveal melanoma – and the third subject presented with twenty colorectal adenomas following panproctocolectomy at age thirty-nine (221). Interestingly, the brother of this third subject also carried the same homozygous *MBD4*^{E314Rfs*13} duplication and, when investigated via colonoscopy, presented with twenty colorectal polyps later determined to be colorectal adenomas with low-grade dysplasia (221).

Whole-exome sequencing was then performed on some of the colorectal adenomas from the first of these three subjects, as well as some of the colorectal adenomas presented by one of the patients from the previously discussed study by Sanders *et al.,* revealing a greatly increased mutation burden compared to sporadic colorectal adenomas (221). The vast majority of these mutations (>95%) were C → T mutations at CpG sites (221,446). Therefore, the mutation signatures of these adenomas almost totally recapitulated the existing COSMIC SBS1 signature (see Figure 4.2a), which is unsurprising given the previously established role of *MBD4* in repairing the spontaneous deamination of 5-mC (221,312). Interestingly, these adenomas presented with a significant enrichment of the *APC*^{R1450*} mutation compared to sporadic colorectal adenomas, a mutation which is driven by a C → T mutation at a CpG site (221). Similarly, the three *MBD4*-mutant AMLs in the study by Sanders *et al.* presented with C → T mutations at CpG sites in key AML driver genes – including *DNMT3A*^{R882C}, *IDH1*^{R132C} and *IDH2*^{R140Q} (446). Therefore, it is possible that germline loss of *MBD4* may be able to drive the accumulation of C → T mutations at CpG sites in key driver genes, presenting a potential mechanism for driving tumorigenesis.

Further study in the context of CRC by Wong *et al.* in *Mbd4*-knockout mouse models found no evidence of enhanced tumorigenesis or inferior survival in *Mbd4*$^{-/-}$ or *Mbd4*$^{+/-}$ animals compared to *Mbd4*-WT controls (450). Interestingly, cells of the small intestine of *Mbd4*$^{-/-}$ animals presented with a near two-fold increase in mutation burdens compared to WT controls (450). A greater proportion of these mutations were C → T mutations at CpG sites in *Mbd4*$^{-/-}$ animals (62%) compared to WT controls (28%) (450). Subsequent crosses with *Apc*$^{1638N/+}$ animals revealed a near-significant increase in tumour multiplicity of *Mbd4*$^{-/-}$*;Apc*$^{1638N/+}$compared to *Mbd4*$^{+/+}$*;Apc*$^{1638N/+}$ animals (450). These double-mutant animals also

presented with a significant increase in both intestinal micro-adenomas and tumours of the jejunum or ileum compared to single-mutant controls (450). Interestingly, these double-mutant animals presented with a significant increase in mutations in the wild-type allele of *Apc,* with 79% of *Mbd4*$^{-/-}$;*Apc*$^{1638N/+}$ animals harbouring a truncation in the wild-type *Apc* allele vs only 32% of *Mbd4*$^{+/+}$;*Apc*$^{1638N/+}$ animals (450). Consistent with the results presented by Palles *et al.,* the majority (82%) of these *Apc* mutations in double-mutant animals were C → T mutations at CpG sites, indicating that *Mbd4* loss may drive *Apc* mutations via unrepaired spontaneous deaminations of 5-mC within the *Apc* protein-coding sequence (221,450).

4.1.4 – Chapter Aims

The COSMIC mutation signature SBS1 has been characterised by a mutation spectrum almost exclusively comprised of C → T mutations at CpG sites (425). This mutation signature has been associated with an age-related accumulation of mutations resulting from unrepaired spontaneous deaminations of 5-mC to thymine (425). Germline mutations in *MBD4* have been demonstrated to increase the number of C → T mutations at CpG sites in a number of cancer types, which may result in mutations in key cancer driver genes (221,446,450). However, the role of *MBD4* as a novel cancer predisposition gene has yet to be fully explored. Recently, Degasperi *et al.* reported an additional six patients with germline mutations in *MBD4* with subsequent somatic LoH in breast cancer, sarcoma and uveal melanoma – however the molecular features of these cancers are yet to be reported (451). In addition to this, there are a number of CpG site-specific factors that may influence the likelihood of spontaneous deaminations occurring. These factors include the DNA methylation status, replication timing and transcription strand of the CpG site – the roles of which are yet to be fully explored in C → T mutagenesis at CpG sites. Previous studies by Fang *et al.* and Poulos *et al.* have suggested that more highly-methylated CpG sites are at higher risk of C → T mutation than lowly-methylated sites (452,453). Poulos *et al.* also suggested that CpG sites located in later-replicating regions of the genome are more likely to undergo C → T mutagenesis than those in earlier-replicating regions (453). This may be a result of the reduced activity of DNA MMR in these later-replicating regions (454). Previous studies in bacteria have suggested that spontaneous deaminations are more common on the coding strand of DNA compared to the template strand (455). This may be a result of single-stranded DNA being more vulnerable to deamination than double-stranded DNA – a state in which the coding strand of DNA is exposed to during transcription, while the template strand is shielded by a complex of transcription-associated proteins (434,437,455,456). Therefore, the aims of this chapter are as follows:

- Assess the role of germline and somatic *MBD4* mutations in cancer pathogenesis by assessing effects on mutation burden, mutation signature composition and driver gene profiles.
- Assess the role of DNA methylation status, replication timing and transcription strand on C → T mutagenesis at CpG sites of the genome.

The hypotheses associated with these chapter aims are as follows:

- Germline and somatic mutations in *MBD4* will drive an increase in C → T mutations at CpG sites and an increase in the prevalence of SBS1 compared to *MBD4*-WT controls.
- *MBD4*-mutant cancers will present with pathogenic C → T mutations at CpG sites within the protein-coding sequence of cancer-specific driver genes.
- C → T mutations at CpG sites will be more prevalent in highly-methylated regions of DNA compared to lowly-methylated regions.
- C → T mutations at CpG sites will be more prevalent in later-replicating regions of DNA compared to earlier-replicating regions.
- C → T mutations at CpG sites will be more prevalent on the coding strand of DNA than the template strand.

4.2 – Materials & Methods

4.2.1 – 100,000 Genomes Project Data

Whole-genome sequencing of colorectal adenocarcinomas (hg38-aligned) from V14 of the 100KGP CRC domain (n = 2,988) were filtered to remove non-CRCs (n = 354), cancers sequenced using a PCR-based library preparation methodology (n = 198), samples derived from metastases (n = 148) and samples with prior DNA-damaging chemotherapy or radiotherapy pre-dating sampling (n = 381). Whole-genome sequencing data available from this dataset included 100x coverage from the tumour sample and an average coverage of 33x from the accompanying normal blood (295). This left a total of 1,907 primary, treatment-naïve CRCs with PCR-free library preparation to be carried forward for further analysis. A further eighteen hyper-mutated samples with pathogenic mutations in the exonuclease domain of DNA polymerase-ε (*POL-ε* – n = 17) or DNA polymerase-δ (*POL-δ* – n = 1) were excluded from the analysis, leaving a total of 1,889 cancers to be taken forward. Cancers were categorised as either MSI^+ or MSS by Dr Juan Fernández-Tajes using the Detecting Microsatellite Instability by Next-Generation Sequencing (mSINGS) tool (457). This analysis indicated there were a total of 357 MSI^+ cancers and 1,532 MSS cancers. Participant ages were calculated using the accompanying clinical data of each cancer by Dr Steve Thorn.

Germline and somatic VCFs generated using the Strelka2 variant caller were then searched for truncations in the *MBD4* gene. Cancers with germline truncations in *MBD4* were assessed for LoH by analysing the associated Sequenza copy number data for the cancer (458). Samples that harboured somatic mutations were analysed further via chi-squared (χ^2) testing with one degree of freedom to determine if the mutation was homozygous or heterozygous. In addition to this, somatic VCFs from cancers with germline *MBD4* mutations and somatic LoH described by Degasperi *et al.* (n = 6) were also obtained for downstream analysis (451).

For each cancer, only somatic variants with the "PASS" flag in the VCF filter field were used for downstream analyses.

4.2.2 – *MBD4*$^{-/-}$ Polyp Whole-Genome Sequencing

Four adenomatous colorectal polyps were extracted from one of the patients reported in the study by Palles *et al.,* who presented with a germline biallelic *MBD4*^{S205Tfs*9} mutation. Three polyps were collected from the descending colon and one from the sigmoid colon to be used in whole-genome sequencing analysis (221). Polyp samples were prepared for whole-genome sequencing analysis by Dr Sara Galavotti (University of Birmingham). Briefly, DNA was extracted from the polyps and accompanying normal blood using the Qiagen AllPrep DNA and RNA extraction kit (Qiagen – catalogue #80284) according to the manufacturer's instructions. The resultant DNA was quantified and sent to Novogene for whole-genome sequencing. The resulting sequencing files were assessed for sequencing quality and lack of adaptor sequences using FastQC (Babraham Bioinformatics) (459). Sequencing reads from both polyp samples and the control blood were aligned to the hg38 reference genome according to the BCBio pipeline. Germline and somatic variant calling was performed using the VarDict and Strelka2 variant callers respectively (460,461).

4.2.3 – Mutation Spectrum & Signature Extraction

The SBS mutation spectrum of the *MBD4*-mutant cancers identified by Degasperi *et al.,* including total SNV burden and mutation counts for each of the ninety-six trinucleotide contexts, was obtained from the supplementary materials of the study (451). The range of cancers presenting with a germline truncation in *MBD4* and subsequent somatic LoH included ductal breast cancer (n = 2), lobular breast cancer (n = 1), myxofibrosarcoma (n = 1), sarcoma of unspecified sub-type (n = 1) and uveal melanoma (n = 1) (451). In addition to this, the same mutation spectra data from *MBD4*-WT cancers of each of the above tumour types was also available in the supplementary data (451). The ages of each participant at tumour sampling were calculated in the same way as described above.

In order to extract mutation signatures from these cancers, as well as the primary CRCs and *MBD4*-mutant polyps described in sections 4.2.1 and 4.2.2 respectively, SigProfilerExtractor was used (462). Briefly, somatic mutations from each cancer that were flagged as "PASS" for all quality filters were used as the input for mutation signature extraction. The chromosome number, genomic position (hg38), relevant filter information, reference allele and alternate allele were used to identify the spectrum of SBS, doublet mutations (DBS) and insertions/deletions (ID) for each cancer. For each of these mutation classes, decomposed mutation signatures were generated, describing the prevalence of previously reported mutation signatures (COSMIC V3.2) in the mutation spectrum of each cancer.

4.2.4 – Characterisation of Driver Gene Mutations

Following the identification of samples with germline inactivation of *MBD4,* somatic VCFs of each sample could be searched for mutations in cancer-specific driver genes. Following a literature search for commonly-mutated driver genes in CRC, breast cancer, sarcoma and uveal melanoma, each *MBD4*-mutant sample was searched for pathogenic mutations in these driver genes. Following this, the genomic co-ordinates (hg38) of C → T or G → A mutations in these driver genes could be searched in the UCSC genome browser (hg38) to identify if the driver gene mutation was localised to a CpG site (372).

4.2.5 – DNA Methylation, Replication Timing & Transcription Strand Mutation Mapping

Fractional methylation data from whole-genome bisulphite sequencing of normal sigmoid colon was obtained from the RoadMap Epigenomics Consortium and converted from hg19 genomic co-ordinates to hg38 using the UCSC LiftOver tool (372,463). This fractional methylation data contained a quantitative estimate of DNA methylation for 27,134,409 CpG sites in the genome ranging from 0 – 1, where 0 indicates an unmethylated CpG site and 1 indicates a highly-methylated CpG site. From this data, CpGs could be assigned to one of twelve DNA methylation bins (0, 0.01 – 0.1, 0.11 – 0.2 etc.) based on their quantitative DNA methylation estimate (see Table 4.1).

In addition to this, DNA replication timing data for the CRC cell line HCT116 was obtained from Dr Duncan Sproul (University of Edinburgh) and Dr Ioannis Kafetzopoulos (Babraham Institute). Quantitative replication timing estimates for 10,000 base-pair windows of the genome were obtained via the Repli-Seq method described by Marchal *et al.* (464). Briefly, two replicates of HCT116 cells were treated with the thymidine analogue bromodeoxyuridine (BrdU), fixed in ethanol and stained with propidium iodide – allowing cells to be separated according to cell cycle stage by fluorescence-activated cell sorting (FACS) (464). Cell populations in early, mid and late S-phase were isolated for lysis and DNA extraction. Genomic DNA was then sonicated and prepared for BrdU immunoprecipitation before finally being sequenced. The replication timing (T) value for each 10,000 base-pair region was defined as the ratio of reads in early S-phase compared to late S-phase:

$$Replication\ Timing\ (T) = \ln(\frac{Early}{Late})$$

Replication timing values across the genome were then smoothed using quantile normalisation, in order to reduce noise. From this quantification of replication timing, the genome could be divided into four equally-sized replication timing bins – Earliest ($T > 2.33304025$), Early ($-0.461036 \leq T \leq 2.3304025$), Late ($-1.8222075 \leq T < -0.461036$) and Latest ($T < -1.8222075$). This replication data was then combined with the above fractional methylation data using BEDTools (v2.30) to provide a T-value for 27,099,859 CpG sites in addition to their DNA methylation estimate. The number of CpG sites in each DNA methylation and replication timing bin is presented in Table 4.1. From this data, somatic C → T mutations at CpG sites in each cancer, represented as either C → T or G → A mutations, could be binned according to both DNA methylation and replication timing using BEDTools (v2.30). In order to correct for differences in the number of CpG sites in each bin, the number of mutations in a bin was normalised against the number of CpGs in the bin according to the following formula, similar to previous work described by Sanders *et al.* (446):

$$Mutation\ Rate = M_x/(\frac{N_x}{1{,}000{,}000})$$

Where M_x represents the number of C → T mutations at CpG sites in bin x and N_x is the number of CpG sites in bin x (see Table 4.1). This provides the number of mutations per million CpGs for each bin, allowing comparisons to be made between methylation and replication timing bins.

Transcription strand data was obtained from Gencode for hg38, assigning a gene as either being transcribed on the sense (trx^+) or antisense (trx^-) strand of the DNA (465). In total, the transcription strand of 15,316,904 CpGs could be inferred from this data. The number of CpGs on each transcription strand is presented in Table 4.2. Mutations could then be assigned to either the coding or template strand of a gene according to the convention set out by Vöhringer *et al.* (466). Briefly, it can be assumed that the coding strand of trx^+ genes is the 5' → 3' DNA strand, whereas the template strand is the 3' → 5' strand. Conversely, the coding strand of trx^- genes can be defined as the 3' → 5' DNA strand and the template strand as the 5' → 3' (466). Therefore, coding strand C → T mutations at CpG sites can be represented as either C → T mutations in trx^+ genes or G → A mutations in trx^- genes. Alternatively, template strand mutations are represented by C → T mutations in trx^- genes or G → A mutations in trx^+ genes. The C → T mutations at CpG sites in each cancer were binned as above using BEDTools (v2.30). The rate of C → T mutagenesis at CpG sites was calculated

# CpG Sites:	0	0.01 – 0.1	0.11 – 0.2	0.21 – 0.3	0.31 – 0.4	0.41 – 0.5	0.51 – 0.6	0.61 – 0.7	0.71 – 0.8	0.81 – 0.9	0.91 – 0.99	1	Total:
Earliest	694,339	602,568	147,037	155,811	204,836	281,637	316,732	422,848	716,854	1,716,719	5,014,966	338,282	10,612,629
Early	233,566	231,985	71,449	87,908	120,828	179,115	213,025	303,816	535,601	1,282,430	3,383,876	221,772	6,865,371
Late	68,720	82,367	37,493	57,952	86,330	135,056	172,162	262,622	476,485	1,106,427	2,123,362	104,684	4,713,660
Latest	66,505	110,886	53,230	71,304	102,776	157,427	204,738	312,647	546,398	1,212,966	1,987,113	82,209	4,908,199
Total:	1,063,130	1,027,806	309,209	372,975	514,770	753,235	906,657	1,301,933	2,275,338	5,318,542	12,509,317	746,947	27,099,859

Table 4.1 – CpG Site DNA Methylation & Replication Timing Distribution: The number of CpG sites found within each DNA methylation and replication timing bin of the genome, according to fractional methylation data of the normal sigmoid colon and Repli-Seq data from the colorectal cancer cell line HCT116. Each column represents one of the twelve possible DNA methylation bins of the genome (ranging from 0 – 1) and each row represents one of the four replication timing bins of the genome (Earliest, Early, Late and Latest).

# CpG Sites:	0	0.01 – 0.1	0.11 – 0.2	0.21 – 0.3	0.31 – 0.4	0.41 – 0.5	0.51 – 0.6	0.61 – 0.7	0.71 – 0.8	0.81 – 0.9	0.91 – 0.99	1	Total:
trx^+	358,959	332,317	86,842	99,421	135,810	199,943	234,881	331,394	582,202	1,399,919	3,807,020	241,583	7,810,291
trx^-	351,222	328,064	86,486	97,211	131,091	190,992	224,866	318,261	558,274	1,340,868	3,647,016	232,262	7,506,613
Total:	710,181	660,381	173,328	196,632	266,901	390,935	459,747	649,655	1,140,476	2,740,787	7,545,036	473,845	15,316,904

Table 4.2 – CpG Site DNA Methylation & Transcription Strand Distribution: The number of CpG sites found within each DNA methylation bin and transcription strand, according to fractional methylation data of the normal sigmoid colon and Gencode data on the direction of transcription of genes. Each column represents one of the twelve possible DNA methylation bins of the genome (ranging from 0 – 1) and each row represents whether a gene is transcribed on the sense (trx^+) or antisense (trx^-) strand.

the same way as above, to correct for the differences in the number of CpG sites in each DNA methylation bin (see Table 4.2).

4.2.6 – Statistical Analysis

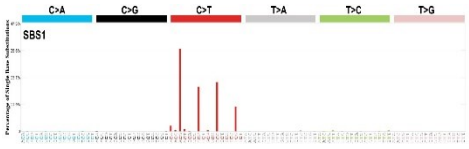
Comparisons between the total SNV burdens, ages, number C → T mutations at CpG sites and proportion of the total SNV burden between *MBD4*-mutant and *MBD4*-WT cancers was performed using a Wilcoxon test via the R package *ggpubr* (365). In the instances where there were multiple groups being compared, a Kruskal-Wallis test was also performed in addition to pairwise Wilcoxon tests between each group. In order to compare the slopes of the relationship between C → T mutation rate at CpGs sites and fractional methylation, linear regressions and interaction terms between equations were calculated. When comparing the total number of mutations on the coding and template transcription strands in each cancer, a $\log_2$(Coding/Template) ratio of mutations was calculated.

4.3 - Results

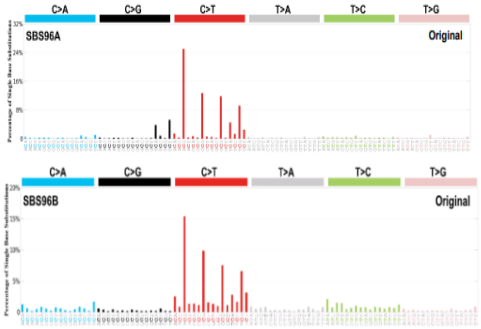
4.3.1 – Germline Mutations in *MBD4* Drive Elevated C → T Mutagenesis at CpG Sites

In addition to the four colorectal polyps from a patient with a germline biallelic truncation in *MBD4* and the cancers described by Degasperi *et al.,* the CRC domain of the 100KGP was searched for additional cancers with *MBD4* truncations. Of the 1,889 primary, treatment-naïve cancers with a PCR-free library preparation, a total of 129 cancers with truncations in *MBD4* were identified, 122 somatic and 7 germline. In agreement with previous findings, the vast majority (121 – 93.8%) of these mutations were identified in MSI^+ cancers. Sequenza copy number data from each of the seven CRCs with germline truncations in *MBD4* indicated no somatic LoH and inspection of the binary alignment map (BAM) files of each of these cancers indicated that these germline truncations were heterozygous. Therefore, the only samples with germline biallelic loss of *MBD4* included in this analysis were the six cancers described by Degasperi *et al.* with germline heterozygous *MBD4* truncations and subsequent somatic LoH and the four colorectal polyps extracted from a patient with a germline biallelic $MBD4^{S205Tfs*9}$ mutation.

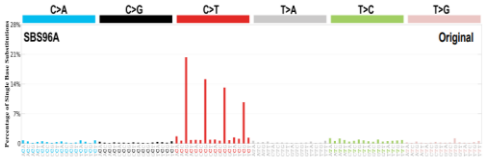
In order to characterise these cancers and colorectal polyps in more detail, the mutation signatures of each sample were extracted using SigProfilerExtractor. According to the study by Degasperi *et al.,* cancers with germline *MBD4* mutations and somatic LoH presented with the novel mutation signature SBS96 (451). Similarly to SBS1, this mutation signature is characterised almost exclusively by C → T mutations at CpG sites, but presents with a greater proportion of C → T mutations in the context C**C**G and fewer in the context G**C**G than SBS1 (see Figure 4.5a) (410). As seen in Figure 4.5, the mutation signatures of each

a) 

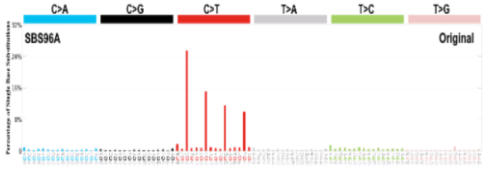
SBS1

b) 

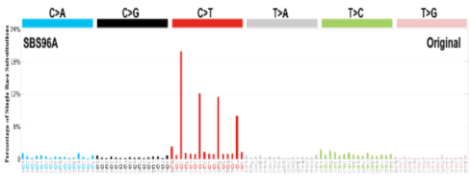
Ductal Breast Cancer

c) 

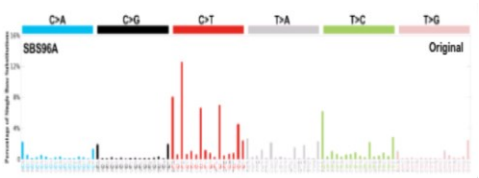
Lobular Breast Cancer

d) 

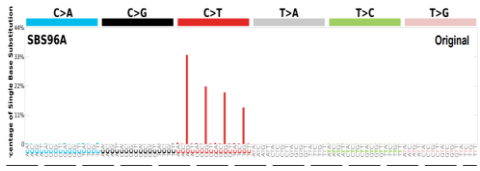
Myxofibrosarcoma

e) 

Sarcoma of Unspecified Sub-type

f) 

Uveal Melanoma

g) 

MBD4$^{-/-}$ Colorectal Polyps

Figure 4.5 – Mutation Signatures of Cancers and Colorectal Polyps with Germline *MBD4* Mutations: The mutation signatures of several cancer types with germline heterozygous truncations in *MBD4* subsequent loss of the wild-type allele. Presented is the mutation signature SBS1 (a, obtained from https://cancer.sanger.ac.uk/signatures/sbs/). Also presented are the mutation signatures, as determined by SigProfilerExtractor, of ductal breast cancers (b), lobular breast cancer (c), myxofibrosarcoma (d), sarcoma of unspecified sub-type (e) and uveal melanoma (f) recently proposed to present with SBS96 as a result of *MBD4* mutation. Mutation signatures are also presented for colorectal polyps isolated from a patient with a germline biallelic truncation in *MBD4* (g).

MBD4-mutant cancer reported by Degasperi *et al.* more closely resembled this novel signature than SBS1 (Figure 4.5b – Figure 4.5f). In addition to this, the four *MBD4*-mutant colorectal polyps described above also presented with a mutation signature that strongly resembled SBS96 (Figure 4.5g), suggesting that this mutation signal is truly associated with loss-of-function *MBD4* mutations – characterised almost exlusively by C → T mutations at CpG sites.

The unique mutation signatures of these cancers with germline mutations in *MBD4* suggest that C → T mutations at CpG sites represent the dominant mutation type. Therefore, mutation counts for each of the ninety-six potential trinucleotide contexts were extracted for each cancer, available in the supplementary material of Degasperi *et al.* (451). These mutation channels could then be compared to cancers that were presumed to be *MBD4*-WT. Interestingly, for each cancer type presented by Degasperi *et al.*, the *MBD4*-mutant cancers displayed an increase in total SNV burden, number of C → T mutations at CpG sites and the proportion of the total SNV burden comprised by these C → T mutations at CpG sites. However, in most cases there was only one *MBD4*-mutant cancer so these increases were not significant. In the *MBD4*-mutant ductal breast cancers (n = 2), there was an increase in the total SNV burden compared to *MBD4*-WT cancers (n = 1,633). However, this increase in SNV burden was not significant (p = 0.12, Figure 4.6a). These *MBD4*-mutant cancers were also slightly younger than their *MBD4*-WT counterparts, but again this difference was not significant (p = 0.46, Figure 4.6b). However, there was a significant increase in both the number (p = 0.015, Figure 4.6c) and proportion (p = 0.014, Figure 4.6d) of C → T mutations at CpG sites in the *MBD4*-mutant cancers compared to the *MBD4*-WT group. Since there was no significant difference between the ages of the *MBD4*-mutant and *MBD4*-WT cancers (Figure 4.6b), this increase in the C → T mutation burden at CpG sites could not be explained by the ages of the *MBD4*-mutant cancers.

Similarly, the single *MBD4*-mutant lobular breast cancer also presented with a higher total SNV burden than the *MBD4*-WT lobular breast cancer population (n = 313). However, again possibly due to the low number of *MBD4*-mutant cancers, this increase was not significant (p = 0.17, Figure 4.7a). There was also no significant difference in the ages of the *MBD4*-mutant and *MBD4*-WT cancers (p = 0.9, Figure 4.7b). Similarly to ductal breast cancer, the *MBD4*-mutant lobular breast cancer presented with a much higher number (Figure 4.7c) and proportion (Figure 4.7d) of C → T mutations at CpG sites which, despite the low number of *MBD4*-mutant cancers, was nearly significant (p = 0.085).

This trend was also apparent outside of breast cancer. In myxofibrosarcoma, the single *MBD4*-mutant cancer showed a non-significant increase in the total SNV burden compared to *MBD4*-WT cancers (n = 114) (p = 0.18, Figure 4.8a). Similarly to ductal breast cancer, this *MBD4*-mutant cancer was also younger than its *MBD4*-WT counterparts (Figure 4.8b), although this difference was not significant (p = 0.17). Furthermore, there was also a near-significant increase in both the number (p = 0.094, Figure 4.8c) and proportion (p = 0.089, Figure 4.8d) of C → T mutations at CpG sites in the *MBD4*-mutant cancer compared to the *MBD4*-WT cancers. In the single *MBD4*-mutant sarcoma of unspecified sub-type, there was a similar non-significant increase in the total SNV burden (p = 0.24, Figure 4.9a) compared to *MBD4*-WT cancers (n = 292). This was accompanied by no difference in the ages of the *MBD4*-mutant and *MBD4*-WT cancers (p = 0.81, Figure 4.9b). Furthermore, there was also a non-significant increase in both the number (p = 0.092, Figure 4.9c) and proportion (p =

Ductal Breast Cancer:

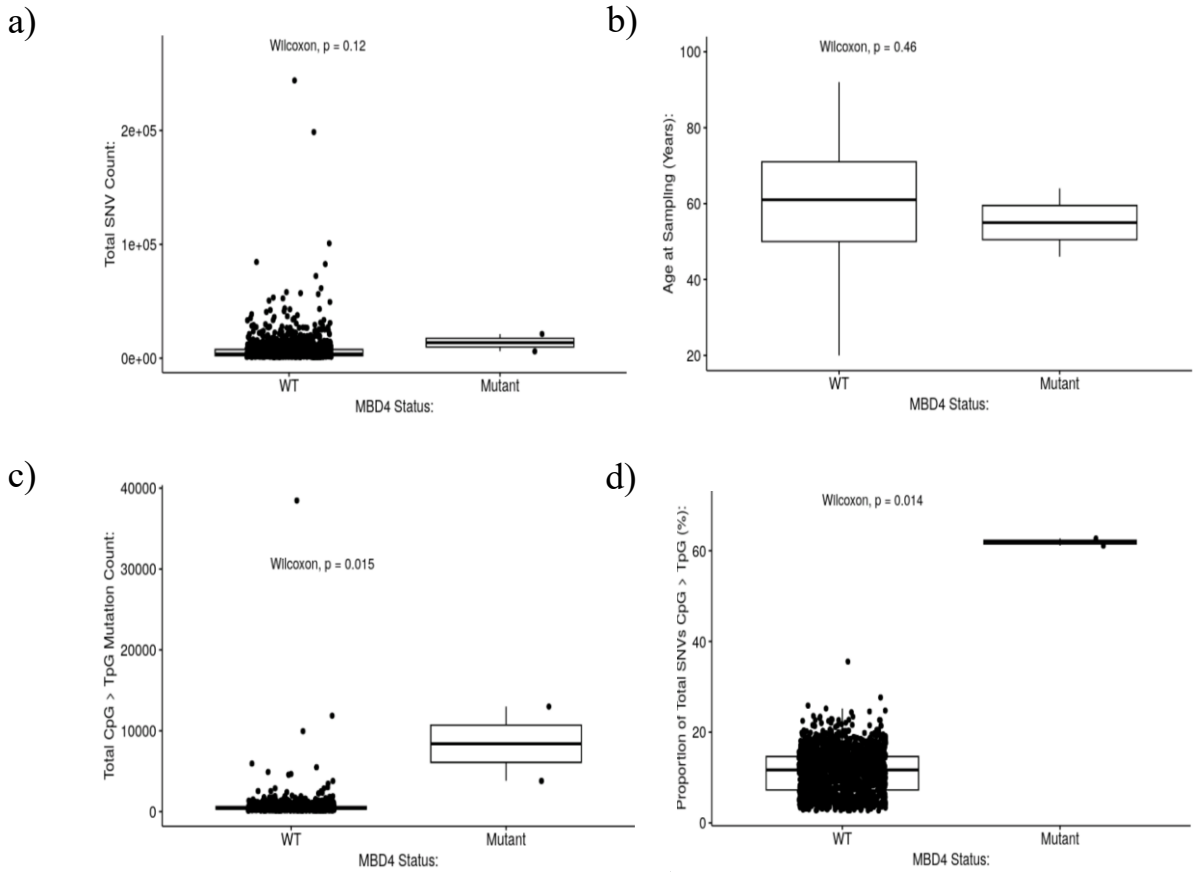


Figure 4.6 – The Mutation Profile of *MBD4*-Mutant Ductal Breast Cancer: Boxplots comparing the total single-nucleotide variation (SNV) burden (a), sampling age (b), number of C → T mutations at CpG sites (c) and the proportion of the total SNV burden that are C → T mutations at CpG sites (d) of *MBD4* wild-type ductal breast cancers (WT) and ductal breast cancers with a germline truncation in *MBD4* and somatic loss of heterozygosity (Mutant).

0.085, Figure 4.9d) of C → T mutations at CpG sites in the *MBD4*-mutant cancer compared to *MBD4*-WT cancers.

In addition to the *MBD4*-mutant breast cancers and sarcomas described above, the single *MBD4*-mutant uveal melanoma presented with a non-significant increase in the total SNV burden compared to its *MBD4*-WT (n = 7) counterparts (p = 0.25, Figure 4.10a). This *MBD4*-mutant cancer was also younger than the *MBD4*-WT cancers, although this difference was not significant (p = 0.5, Figure 4.10b). There were also non-significant increases in the number (Figure 4.10c) and proportion (Figure 4.10d) of C → T mutations at CpG sites in the *MBD4*-mutant cancer compared to the *MBD4*-WT group (p = 0.25). Overall, this indicates that the *MBD4*-mutant cancers presented in the study by Degasperi *et al.* are associated with increased C → T mutagenesis at CpG sites compared to their *MBD4*-WT counterparts.

Lobular Breast Cancer:

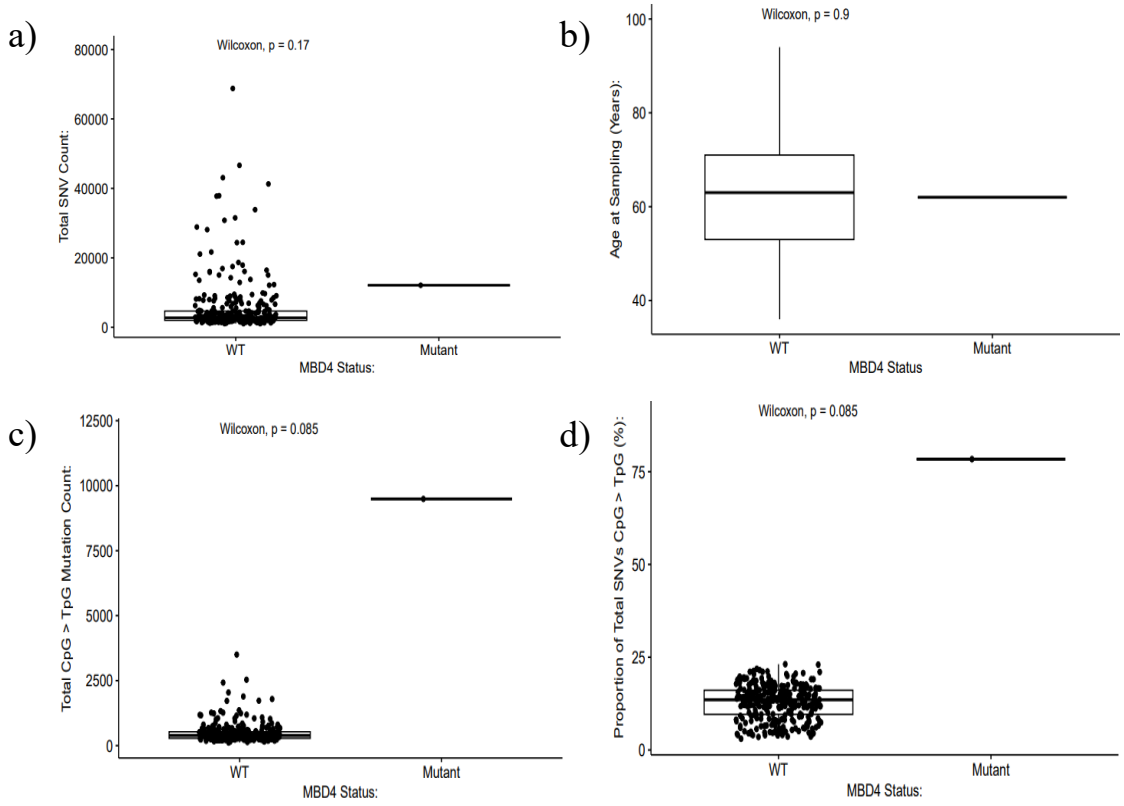


Figure 4.7 – The Mutation Profile of *MBD4*-Mutant Lobular Breast Cancer: Boxplots comparing the total single-nucleotide variation (SNV) burden (a), sampling age (b), number of C → T mutations at CpG sites (c) and the proportion of the total SNV burden that are C → T mutations at CpG sites (d) of *MBD4* wild-type lobular breast cancers (WT) and the single lobular breast cancer with a germline truncation in *MBD4* and somatic loss of heterozygosity (Mutant).

In contrast to this, the mutation spectrum and mutation signature profile of the CRCs with germline heterozygous truncations in *MBD4* was not significantly different to *MBD4*-WT cancers. As seen in Figure 4.11, MSS CRCs with germline heterozygous *MBD4* truncations (n = 5) had no significant differences in SNV burden (p = 0.65, Figure 4.11a) or age (p = 0.17, Figure 4.11b) compared to MSS *MBD4*-WT CRCs (n = 1,524). Furthermore, these cancers presented with no differences in both the number (p = 0.38, Figure 4.11c) or proportion (p = 0.36, Figure 4.11d) of C → T mutations at CpG sites compared to their *MBD4*-WT counterparts. Following mutation signature extraction from these cancers, there was no difference in the number (p = 0.35, Figure 4.11e) or proportion (p = 0.38, Figure 4.11f) of SNVs attributed to the mutation signature SBS1 in these CRCs with germline heterozygous *MBD4* truncations compared to *MBD4*-WT cancers. The same was apparent in MSI^+ CRCs with germline heterozygous *MBD4* truncations (n = 2). As seen in Figure 4.12, there were no difference in the SNV burden (p = 0.65, Figure 4.12a) or ages (p = 0.33, Figure 4.12b) of these cancers with germline heterozygous *MBD4* truncations compared to MSI^+

Myxofibrosarcoma:

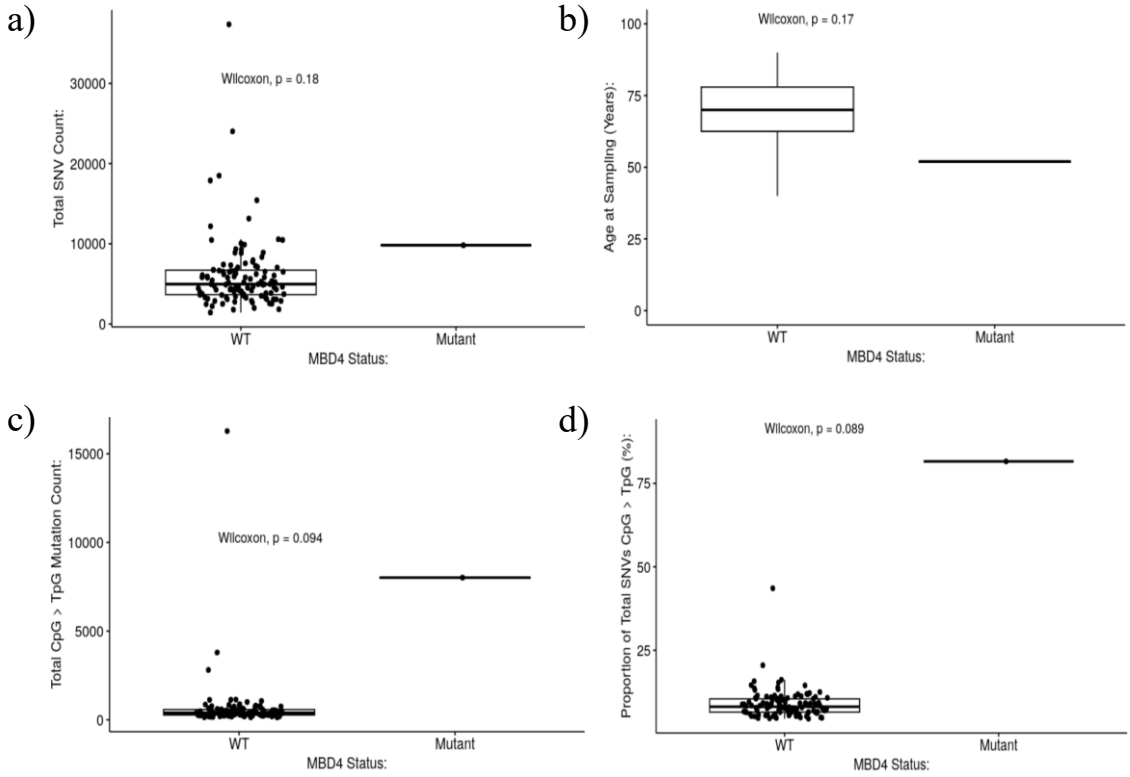


Figure 4.8 – The Mutation Profile of *MBD4*-Mutant Myxofibrosarcoma: Boxplots comparing the total single-nucleotide variation (SNV) burden (a), sampling age (b), number of C → T mutations at CpG sites (c) and the proportion of the total SNV burden that are C → T mutations at CpG sites (d) of *MBD4* wild-type myxofibrosarcomas (WT) and a myxofibrosarcoma with a germline truncation in *MBD4* and somatic loss of heterozygosity (Mutant).

MBD4-WT CRCs (n = 236). There was also no difference in the number (p = 0.31, Figure 4.12c) or proportion (p = 0.64, Figure 4.12d) of C → T mutations at CpG sites compared to *MBD4*-WT cancers. When the mutation signatures were extracted from these MSI^+ CRCs, there was no difference in the number (p = 0.35, Figure 4.12e) or proportion (p = 0.93, Figure 4.12f) of SNVs attributed to SBS1 compared to *MBD4*-WT cancers. Overall, this data indicates that germline heterozygous truncations in *MBD4* have no effect on C → T mutagenesis at CpG sites and therefore have no bearing on the prevalence of SBS1 within these cancers.

However, the *MBD4*-mutant colorectal polyps with a germline biallelic *MBD4* truncation presented with significant differences compared to MSS *MBD4*-WT CRCs. As seen in Figure 4.11, the *MBD4*-mutant polyps presented with an increased SNV burden compared to *MBD4*-WT CRCs (p = 0.0077, Figure 4.11a). The individual from whom the polyps were extracted was significantly younger than the *MBD4*-WT CRCs (p = 0.0018, Figure 4.11b). This was perhaps expected given that these were colorectal polyps being compared to CRCs. There was also a significant increase in both the number (p = 0.00055, Figure 4.11c) and proportion

Sarcoma of Unspecified Sub-Type:

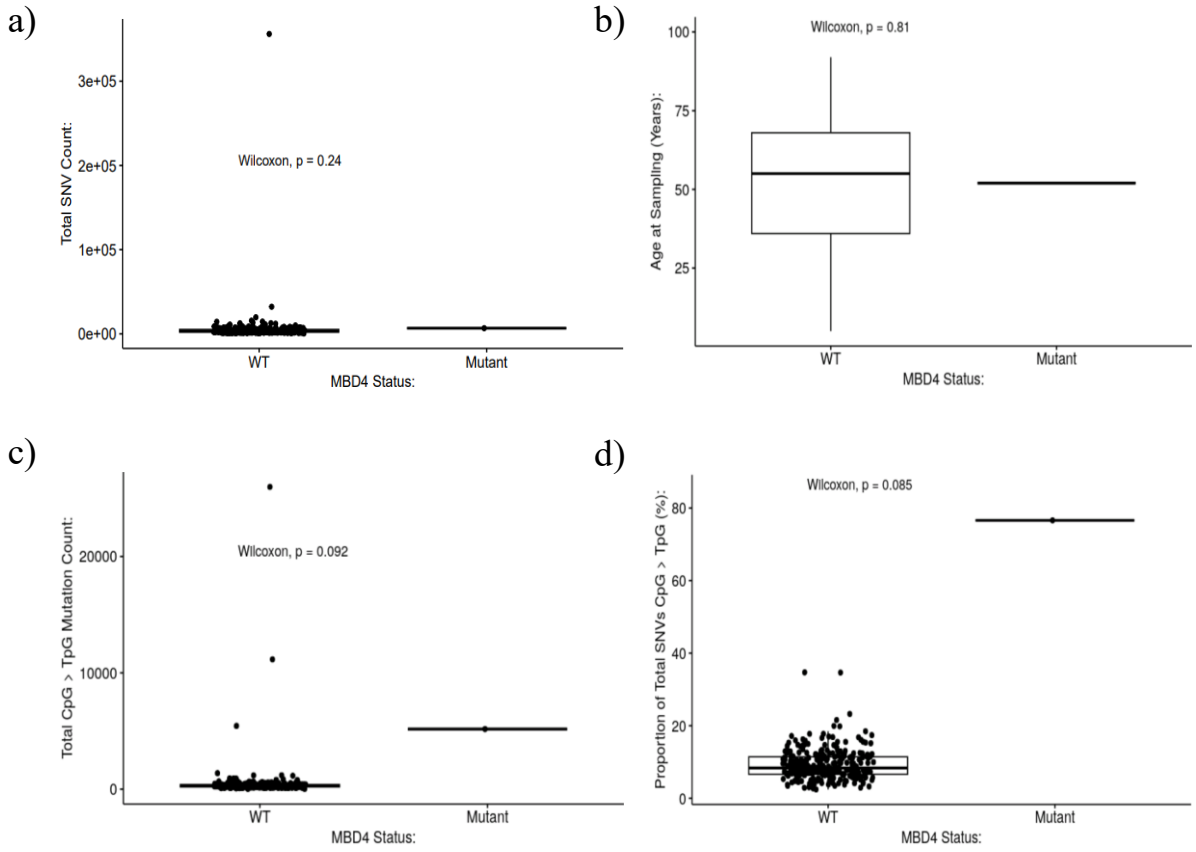


Figure 4.9 – The Mutation Profile of *MBD4*-Mutant Sarcoma of Unspecified Sub-type: Boxplots comparing the total single-nucleotide variation (SNV) burden (a), sampling age (b), number of C → T mutations at CpG sites (c) and the proportion of the total SNV burden that are C → T mutations at CpG sites (d) of *MBD4* wild-type sarcomas of unspecified sub-type (WT) and a sarcoma of unspecified sub-type with a germline truncation in *MBD4* and somatic loss of heterozygosity (Mutant).

(p = 0.00054, Figure 4.11d) of C → T mutations at CpG sites in these *MBD4*-mutant colorectal polyps compared to the *MBD4*-WT CRCs. The SNV burden and proportion of this burden comprised of C → T mutations at CpG sites was comparable to the whole-exome sequencing analysis of Palles *et al.* of colorectal polyps from *MBD4*-deficient individuals (221). Mutation signature extraction from these polyps revealed a significant increase in both the number (p = 0.00055, Figure 4.11e) and proportion (p = 0.00054, Figure 4.11f) of SNVs attributed to SBS1, indicating that these *MBD4*-mutant colorectal polyps present with increased C → T mutagenesis at CpG sites compared to *MBD4*-WT CRCs, which is reflected in the mutation signature composition of these groups.

While there was an apparent increase in C → T mutagenesis at CpG sites of cancers with germline heterozygous *MBD4* truncations and subsequent somatic LoH and in the *MBD4*-mutant colorectal polyps, the consequences of somatic *MBD4* truncations on C → T

Uveal Melanoma:

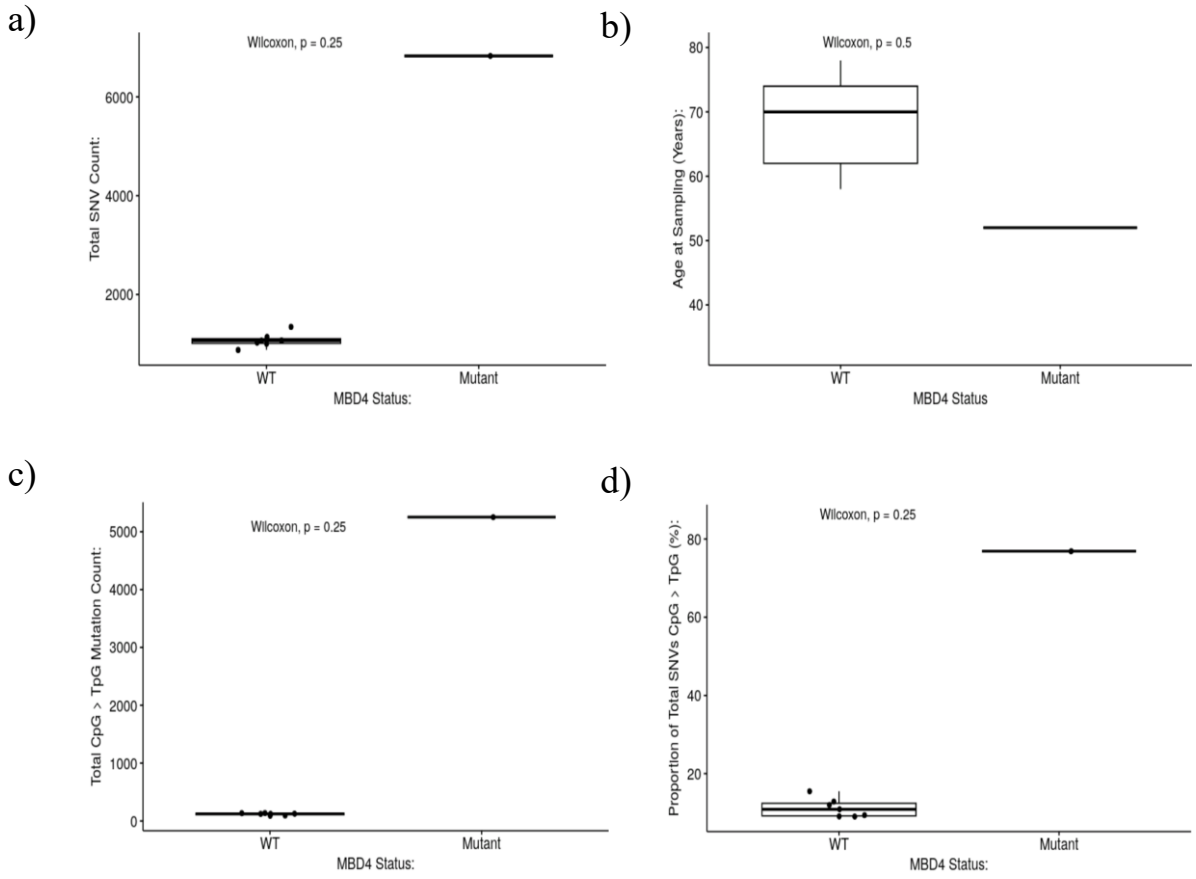
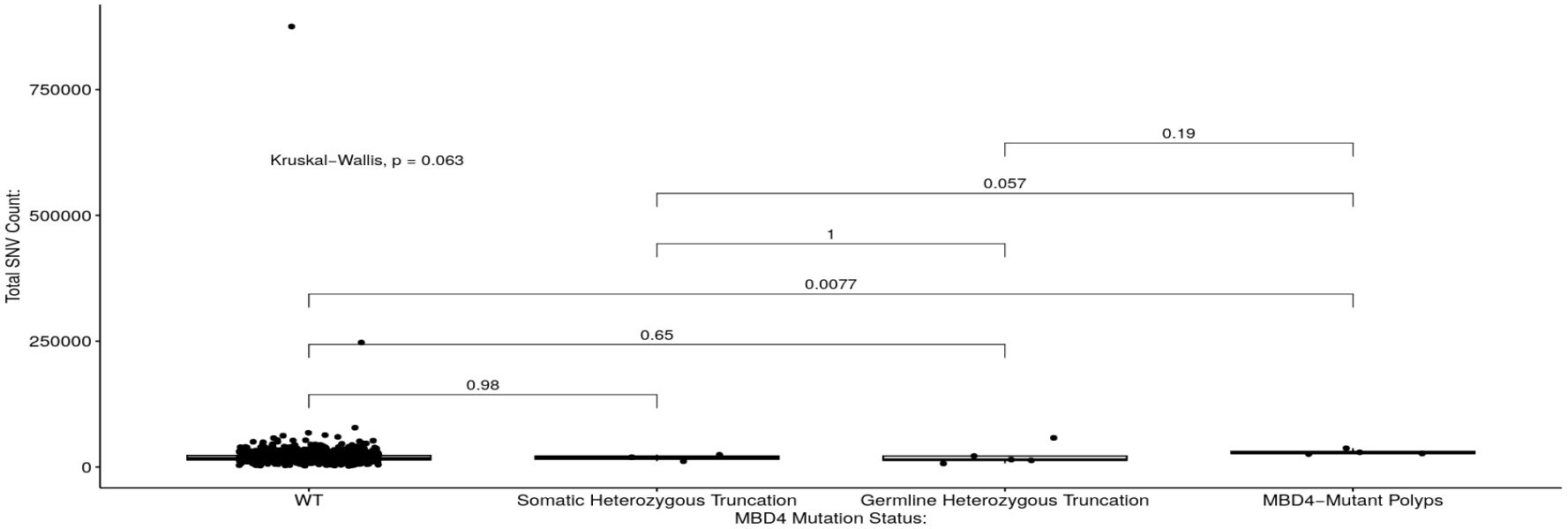


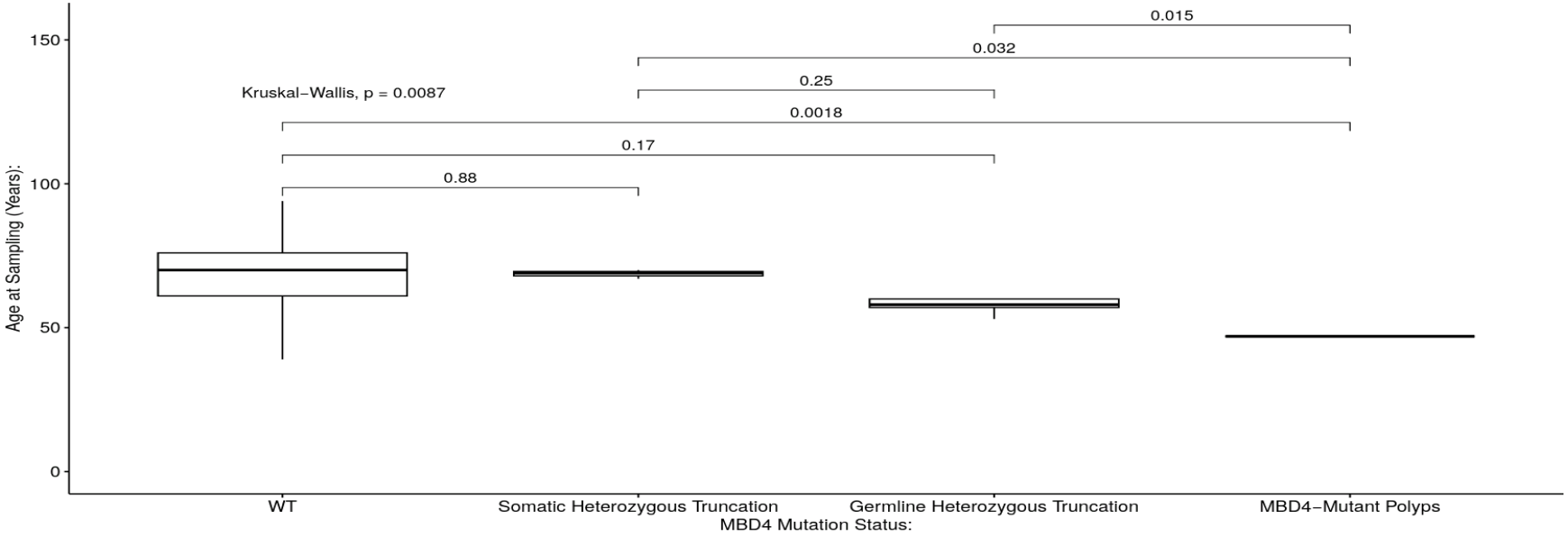
Figure 4.10 – The Mutation Profile of *MBD4*-Mutant Uveal Melanoma: Boxplots comparing the total single-nucleotide variation (SNV) burden (a), sampling age (b), number of C → T mutations at CpG sites (c) and the proportion of the total SNV burden that are C → T mutations at CpG sites (d) of *MBD4* wild-type uveal melanoma (WT) and the single uveal melanoma with a germline truncation in *MBD4* and somatic loss of heterozygosity (Mutant).

mutagenesis at CpG sites has yet to be fully explored in CRC. As previously described, a total of 122 CRCs presented with somatic truncations in *MBD4*. In order to determine if these truncations were likely heterozygous or homozygous, χ^2 analysis on the number of mutant reads in the BAM files of each cancer could be performed to identify if, after incorporating tumour purity estimates, the number of mutant reads was significantly greater than would be expected if the mutation was heterozygous. Similarly, the same analysis could be performed to identify if the number of mutant reads was not significantly less than the expected number if the mutation was homozygous. In order to correct for multiple testing, a Bonferroni-corrected $p_{(\chi^2)}$ threshold was set at 0.00041 (0.05/122). As seen in Table 4.3, a total of twelve CRCs presented with a $p_{(\chi^2)}$ below this threshold. These cancers also has a $p_{(\chi^2)} \geq 0.05$ when the null hypothesis assumed the mutation was homozygous, suggesting with high confidence that these cancers had somatic biallelic truncations in *MBD4.* All of these cancers were MSI^+

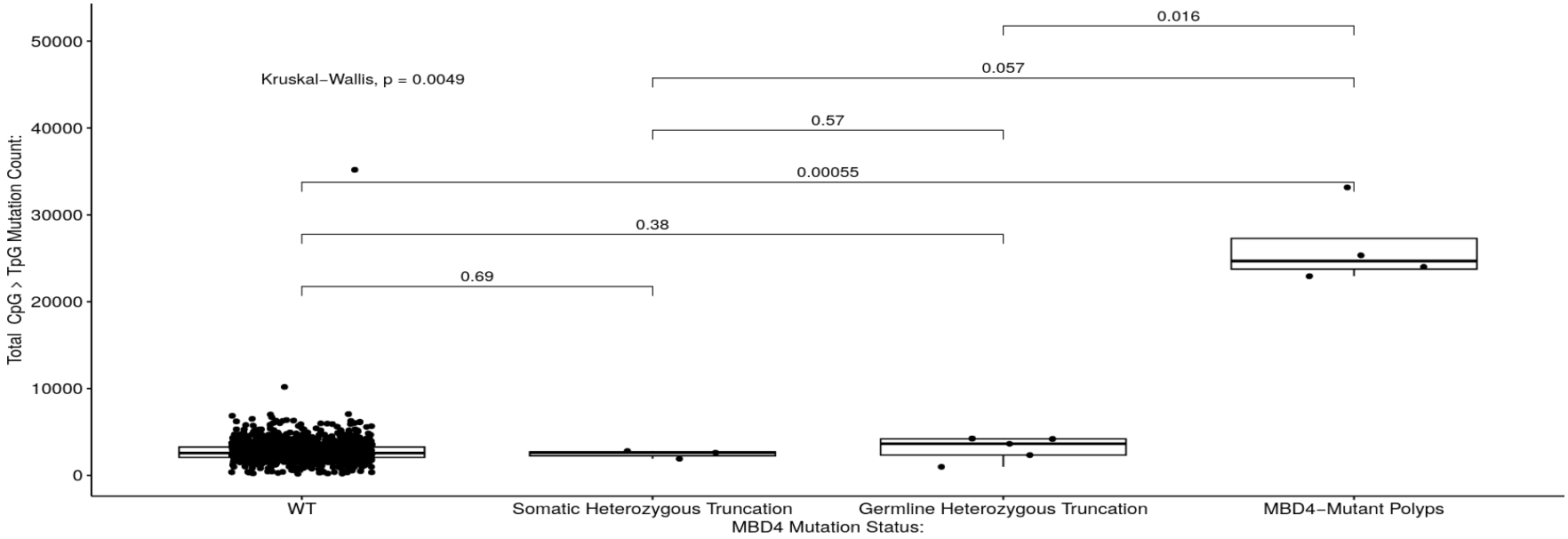
a)


750000
500000
250000
0
Total SNV Count:
Kruskal–Wallis, p = 0.063
0.19
0.057
1
0.0077
0.65
0.98
WT
Somatic Heterozygous Truncation
Germline Heterozygous Truncation
MBD4–Mutant Polyps
MBD4 Mutation Status:

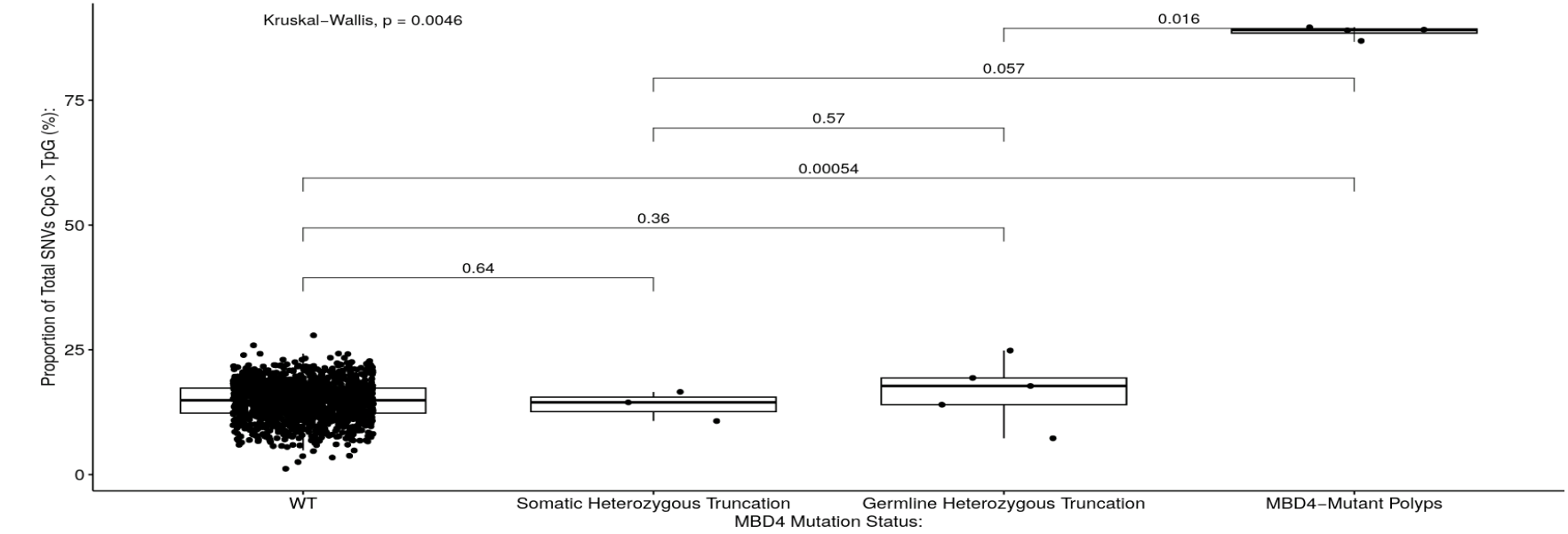
b)


150
100
50
0
Age at Sampling (Years):
Kruskal–Wallis, p = 0.0087
0.015
0.032
0.25
0.0018
0.17
0.88
WT
Somatic Heterozygous Truncation
Germline Heterozygous Truncation
MBD4–Mutant Polyps
MBD4 Mutation Status:

c)


Kruskal–Wallis, p = 0.0049
Total CpG > TpG Mutation Count:
50000
40000
30000
20000
10000
0
0.016
0.057
0.57
0.00055
0.38
0.69
WT
Somatic Heterozygous Truncation
Germline Heterozygous Truncation
MBD4–Mutant Polyps
MBD4 Mutation Status:

d)


Kruskal–Wallis, p = 0.0046
Proportion of Total SNVs CpG > TpG (%):
75
50
25
0
0.016
0.057
0.57
0.00054
0.36
0.64
WT
Somatic Heterozygous Truncation
Germline Heterozygous Truncation
MBD4–Mutant Polyps
MBD4 Mutation Status:

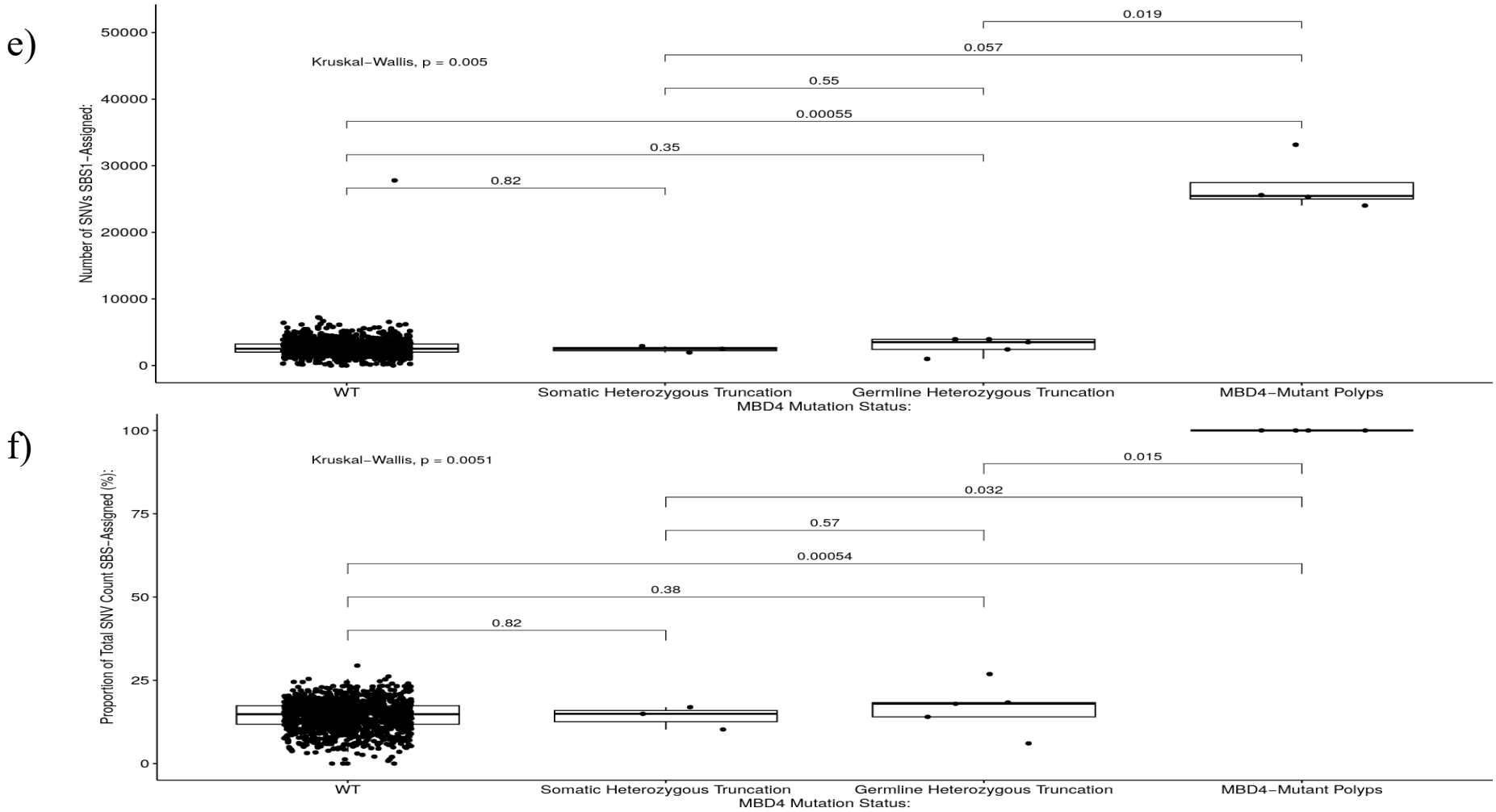
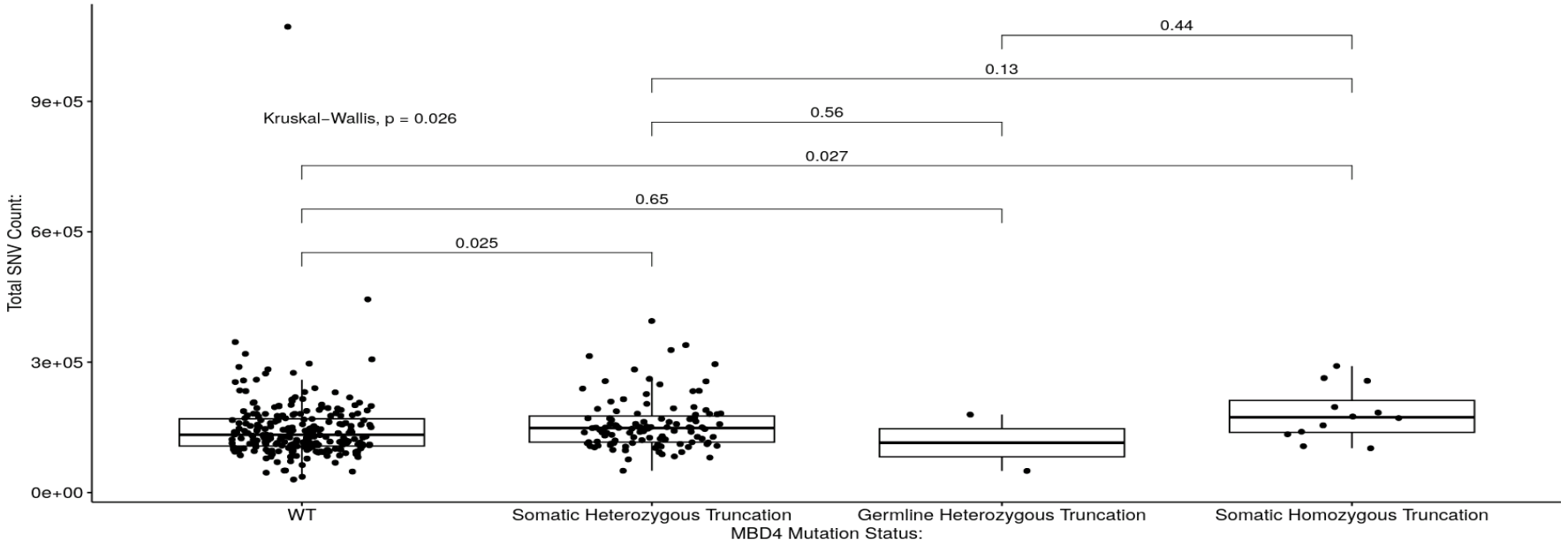
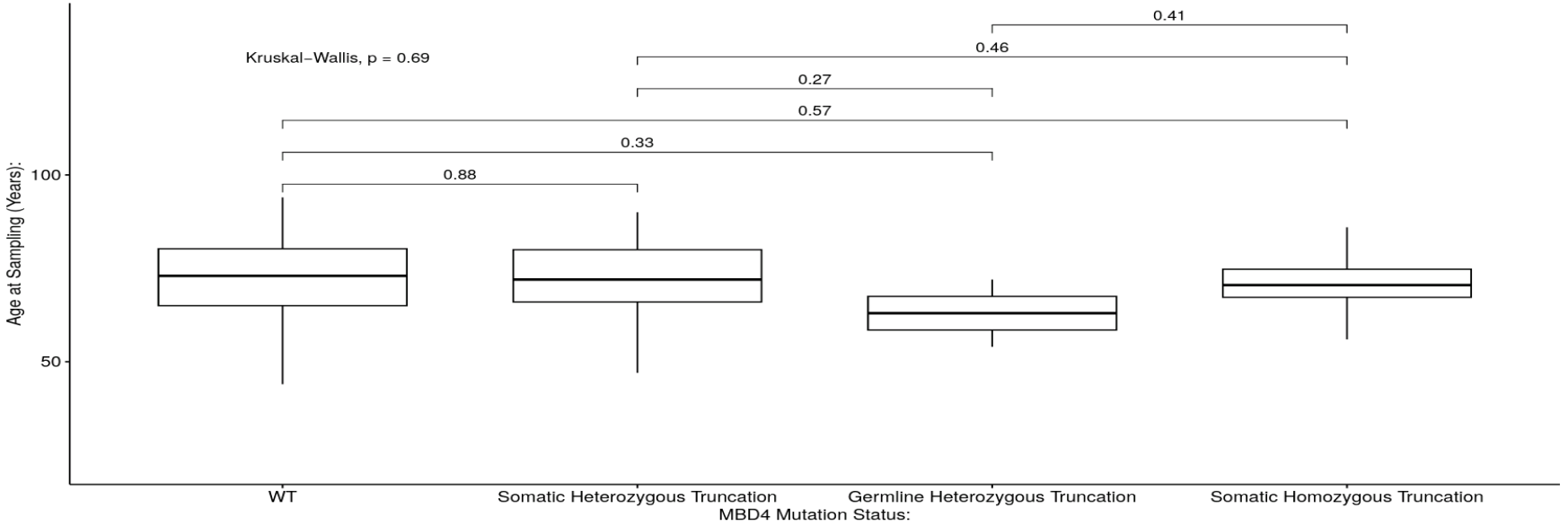


Figure 4.11 – The Mutation Profile of *MBD4*-Mutant Colorectal Polyps: Boxplots comparing the total single-nucleotide variation (SNV) burden (a), sampling age (b), number of C → T mutations at CpG sites (c), the proportion of the total SNV burden that are C → T mutations at CpG sites (d), the number of SNVs attributed to the COSMIC signature SBS1 (e) and the proportion of SNVs attributed to SBS1 (f) of microsatellite stable colorectal cancers that are *MBD4* wild-type (WT) or have either somatic or germline heterozygous truncations in *MBD4*. Also included is data from colorectal polyps extracted from a patient with a germline biallelic *MBD4* truncation.

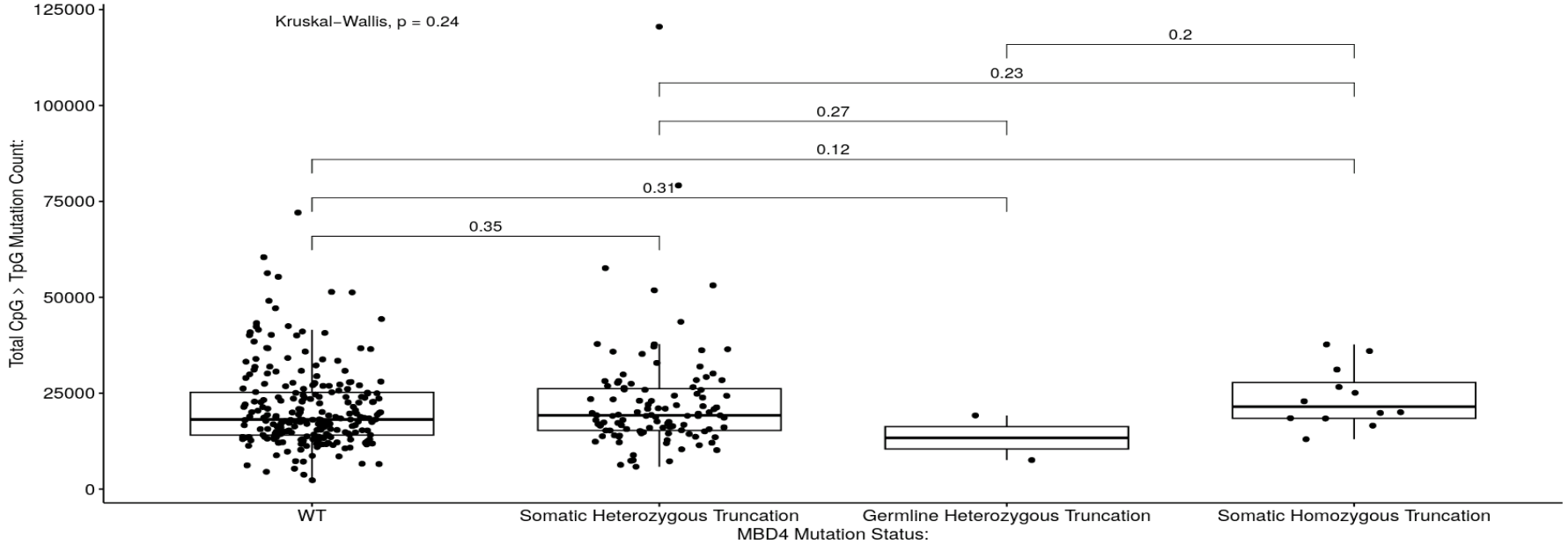
a)


Kruskal–Wallis, p = 0.026
0.44
0.13
0.56
0.027
0.65
0.025
Total SNV Count:
9e+05
6e+05
3e+05
0e+00
WT
Somatic Heterozygous Truncation
Germline Heterozygous Truncation
Somatic Homozygous Truncation
MBD4 Mutation Status:

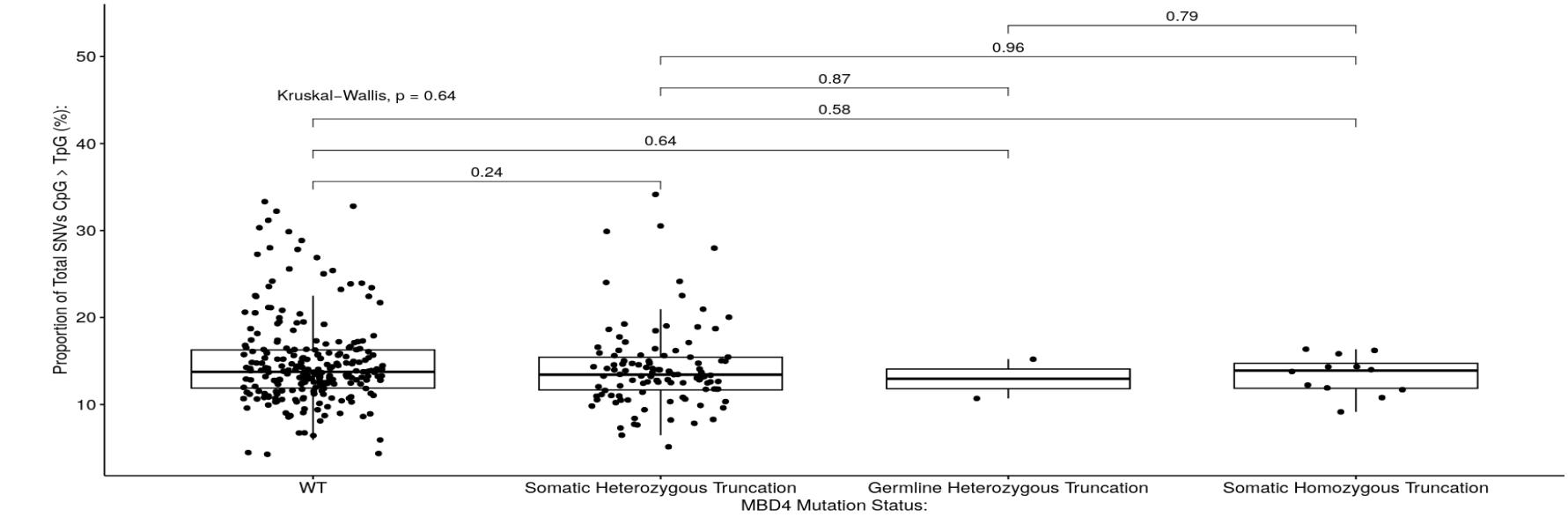
b)


Kruskal–Wallis, p = 0.69
0.41
0.46
0.27
0.57
0.33
0.88
Age at Sampling (Years):
100
50
WT
Somatic Heterozygous Truncation
Germline Heterozygous Truncation
Somatic Homozygous Truncation
MBD4 Mutation Status:

c)


Kruskal–Wallis, p = 0.24
0.2
0.23
0.27
0.12
0.31
0.35
125000
100000
75000
50000
25000
0
Total CpG > TpG Mutation Count:
WT
Somatic Heterozygous Truncation
Germline Heterozygous Truncation
Somatic Homozygous Truncation
MBD4 Mutation Status:

d)


Kruskal–Wallis, p = 0.64
0.79
0.96
0.87
0.58
0.64
0.24
50
40
30
20
10
Proportion of Total SNVs CpG > TpG (%):
WT
Somatic Heterozygous Truncation
Germline Heterozygous Truncation
Somatic Homozygous Truncation
MBD4 Mutation Status:

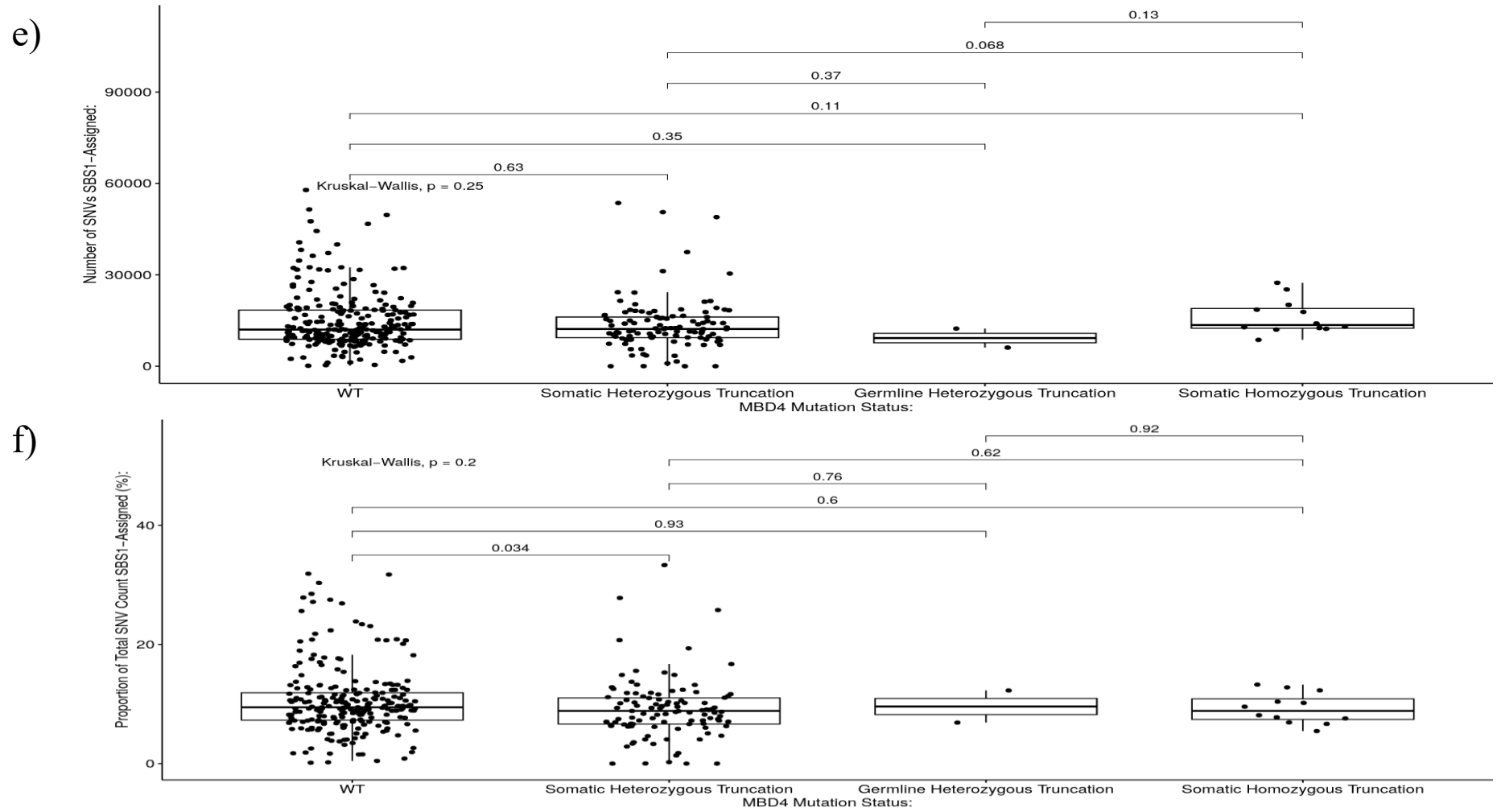


Figure 4.12 – The Mutation Profile of MSI$^+$ Colorectal Cancers with *MBD4* Mutations: Boxplots comparing the total single-nucleotide variation (SNV) burden (a), sampling age (b), number of C → T mutations at CpG sites (c), the proportion of the total SNV burden that are C → T mutations at CpG sites (d), the number of SNVs attributed to the COSMIC signature SBS1 (e) and the proportion of SNVs attributed to SBS1 (f) of microsatellite unstable colorectal cancers that are either *MBD4* wild-type (WT) or harbour somatic heterozygous *MBD4* truncations, germline heterozygous *MBD4* truncations or likely biallelic somatic *MBD4* truncations.

Cancer #:	*MBD4* Mutation:	Tumour Purity (%):	Alternate Reads:	Expected Reads if Heterozygous:	$P_{(\chi^2)}$:	Expected Reads if Homozygous:	$P_{(\chi^2)}$:
1	$MBD4^{E314fs*}$	39	35	19	0.000051	38	0.5047
2	$MBD4^{E314fs*}$	58	50	25	< 0.00001	49	0.8777
3	$MBD4^{E314fs*}$	78	52	28	< 0.00001	57	0.1628
4	$MBD4^{E314fs*}$	53	59	31	< 0.00001	61	0.2128
5	$MBD4^{E314fs*}$	61	69	34	< 0.00001	68	0.8951
6	$MBD4^{E314fs*}$	67	55	32	< 0.00001	64	0.0591
7	$MBD4^{E314fs*}$	53	69	32	< 0.00001	64	0.375
8	$MBD4^{E314fs*}$	44	61	25	< 0.00001	50	NA
9	$MBD4^{E314fs*}$	12	20	7	< 0.00001	15	0.146
10	$MBD4^{E314fs*}$	34	54	22	< 0.00001	45	0.081
11	$MBD4^{E314fs*}$	37	53	26	< 0.00001	51	0.7323
12	$MBD4^{E314fs*}$	48	51	29	< 0.00001	58	0.2279

Table 4.3 – Somatic Homozygous *MBD4* Truncations in 100,000 Genomes Project Samples: Details of the samples with homozygous truncations in the *MBD4* gene in the colorectal cancer domain (V14) of the 100,000 Genomes Project. Included are the specific *MBD4* mutation present within the sample, the estimated tumour purity of the sample and the number of reads that aligned to the mutant allele (Alternate Reads). Also included are the expected number of reads if the mutation was either heterozygous or homozygous and the associated chi-squared (χ^2) p-values ($p_{(\chi^2)}$).

and harboured the *MBD4*^{E314fs*} mutation, a common mutation within the $A_{(10)}$ tract of the *MBD4* protein-coding sequence.

A total of three MSS CRCs harboured somatic heterozygous truncations in *MBD4* (see Figure 4.11), but these cancers presented with no differences in SNV burden (p = 0.98, Figure 4.11a) or age (p = 0.88, Figure 4.11b) compared to *MBD4*-WT cancers. These cancers also had no differences in the number (p = 0.69, Figure 4.11c) or proportion (p = 0.64, Figure 4.11d) of C → T mutations at CpG sites compared to their *MBD4*-WT counterparts. The number (p = 0.82, Figure 4.11e) and proportion (p = 0.82, Figure 4.11f) of SNVs attributed to SBS1 was also not significantly different to *MBD4*-WT cancers, overall indicating that MSS cancers with somatic heterozygous truncations in *MBD4* did not present with increased C → T mutagenesis at CpG sites compared to their *MBD4*-WT counterparts.

In MSI^+ CRCs, a total of 107 cancers had somatic heterozygous *MBD4* truncations, with an additional twelve harbouring likely homozygous truncations (see Table 4.3). Interestingly, both of these groups of cancer presented with an increased SNV burden compared to their *MBD4*-WT counterparts (p = 0.025 and p = 0.027 respectively, Figure 4.12a). However, there were no differences in the ages of the MSI^+ cancers with somatic heterozygous (p = 0.88, Figure 4.12b) or homozygous (p = 0.57, Figure 4.12b) *MBD4* truncations compared to their *MBD4*-WT counterparts. Similarly, these cancers had no differences in the number (p = 0.35 and p = 0.12 respectively, Figure 4.12c) or proportion (p = 0.24 and p = 0.58 respectively, Figure 4.12d) of C → T mutations at CpG sites compared to *MBD4*-WT CRCs. Following mutation signature extraction, MSI^+ CRCs with somatic heterozygous or homozygous *MBD4* truncations presented with no differences in the number of SNVs assigned to SBS1 compared to *MBD4*-WT cancers (p = 0.63 and p = 0.11 respectively, Figure 4.12e). Surprisingly, despite not having a significant difference in the number of SNVs assigned to SBS1 compared to *MBD4*-WT cancers, CRCs with somatic heterozygous *MBD4* truncations presented with a significantly lower proportion of SNVs attributed to SBS1 than *MBD4*-WT cancers (p = 0.034, Figure 4.12f). However, there was no significant difference in the CRCs with somatic homozygous *MBD4* truncations (p = 0.6, Figure 4.12f). Therefore, this data suggests that somatic truncations in *MBD4,* including likely biallelic truncations, have no effect on C → T mutagenesis at CpG sites in MSI^+ CRC or the prevalence of SBS1 within these cancers.

Overall, this data suggests that germline mutations in *MBD4* and the subsequent loss of the WT allele drives a significant increase in C → T mutagenesis at CpG sites in a number of cancer types, including breast cancer, sarcoma and uveal melanoma, which could not be explained by patient age. In addition to this, colorectal polyps with a germline biallelic mutation in *MBD4* also presented with this increase in C → T mutagenesis at CpG sites compared to *MBD4*-WT CRCs. However, this same phenomenon could not be replicated in CRCs with somatic mutations in *MBD4,* as even cancers with likely biallelic truncations in *MBD4* showed no significant difference in the number of C → T mutations at CpG sites compared to *MBD4*-WT cancers. Therefore, it may be the case that germline mutations in *MBD4*, either biallelic or heterozygous with subsequent loss of the WT allele, may represent a novel cancer predisposition syndrome, acting via an increase in C → T mutagenesis at CpG sites. It would therefore be prudent to characterise the driver gene profile of these *MBD4*-mutant cancers in order to investigate the mechanism by which this novel predisposition syndrome may operate.

4.3.2 – Germline *MBD4* Mutations May Drive Pathogenic Mutations in Cancer Driver Genes

In section 4.3.1, it was suggested that germline mutations in *MBD4,* either biallelic or monoallelic with loss of the WT allele, drives an increase in the number of C → T mutations at CpG sites – potentially via an accumulation of unrepaired spontaneous deaminations as a consequence of *MBD4* loss. Interestingly, this increase in C → T mutagenesis at CpG sites could not be explained by age, as there was some evidence to suggest that *MBD4* loss may be correlated with cancer development at a younger age. Therefore, it is possible that germline mutations in *MBD4* represent a novel cancer predisposition syndrome as suggested by Palles *et al.*, driving tumorigenesis via the accumulation of pathogenic C → T mutations at CpG sites within the protein-coding sequence of key cancer-specific driver genes (221).

In order to test this theory, the *MBD4*-mutant cancers identified by Degasperi *et al.* were searched for pathogenic C → T mutations at CpG sites in the protein-coding sequence of cancer-specific driver genes. The results of this analysis are presented in Table 4.4. Interestingly, several pathogenic C → T mutations at CpG sites were detected across a range of cancer-specific driver genes. In ductal breast cancer, point mutations were identified in *TP53* and *MRE11A* – genes where pathogenic mutations have been linked to disease pathogenesis (467,468). Following variant effect prediction, both $TP53^{R248W}$ and $MRE11A^{R351C}$ were categorised as "deleterious" and "probably damaging" by the Sorting Intolerant from Tolerant (SIFT) and Polymorphism Phenotyping (PolyPhen) algorithms respectively. In addition to this, a second *MBD4*-mutant ductal breast cancer presented with a truncation in Ring Finger Protein 31 (*RNF31*), a gene which has been previously suggested to alter *TP53* signalling in breast cancer (469). In *MBD4*-mutant lobular breast cancer, a pathogenic mutation in *PTEN* was identified according to both the SIFT and Polyphen algorithms (Table 4.4). The role of *PTEN* as a breast cancer driver gene has been well documented, while the $PTEN^{R130Q}$ mutation present in this *MBD4*-mutant cancer has been previously reported in invasive breast cancer by Yang *et al.* (470).

Furthermore, *MBD4*-mutant myxofibrosarcoma and sarcoma of unspecified sub-type presented with truncations in the *TP53* gene (Table 4.4). Mutations in *TP53* have previously been suggested to be present in 44% of myxofibrosarcomas while sarcomas in general are estimated to make up 25% of cancers with inherited *TP53* deficiency – further indicating the role of *TP53* as a sarcoma driver gene (471,472). In uveal melanoma, a cancer where *MBD4* mutations are thought to be drivers, the *MBD4*-mutant cancer presented with a truncation in *BRCA1*-Associated Protein 1 (*BAP1*) and a pathogenic missense variant in *GNA11* (Table 4.4). Both of these genes have been reported as drivers in uveal melanoma, with germline *BAP1* mutations associated with a genetic predisposition to uveal melanoma and pathogenic mutations in *GNA11* identified in 32% of primary uveal melanomas and 57% of uveal melanoma metastases (473,474).

When the same driver gene analysis was performed on the *MBD4*-mutant colorectal polyps (Table 4.5), each of the polyps was found to harbour pathogenic mutations in the protein-coding sequence of *APC*. Given the previously reported role of *APC* as a CRC driver gene (see Chapter I of this thesis for a full description), this is perhaps not unexpected. As seen in

Cancer #:	Cancer Type:	Gene Mutation:	Position (hg38):	Reference Allele:	Alternate Allele:	SIFT:	PolyPhen:
1	Ductal Breast Cancer	*MRE11A*^{R351C}	Chr11:94,467,860	G	A	Deleterious	Probably Damaging
1	Ductal Breast Cancer	*TP53*^{R248W}	Chr17:7,674,221	G	A	Deleterious	Probably Damaging
2	Ductal Breast Cancer	*RNF31*^{R176*}	Chr14:24,150,230	C	T	NA	NA
3	Lobular Breast Cancer	*PTEN*^{R130Q}	Chr10:87,933,148	G	A	Deleterious	Probably Damaging
4	Myxofibrosarcoma	*TP53*^{R174*}	Chr17:7,673,704	G	A	NA	NA
5	Sarcoma of Unspecified Sub-Type	*TP53*^{R81*}	Chr17:7,674,894	G	A	NA	NA
6	Uveal Melanoma	*BAP1*^{R60*}	Chr3:52,408,551	G	A	NA	NA
6	Uveal Melanoma	*GNA11*^{R183C}	Chr19:3,115,014	C	T	Deleterious	Probably Damaging

Table 4.4 – C → T Mutations at CpG Sites in *MBD4*-Mutant Cancer Driver Genes: Details of pathogenic C → T mutations in the coding regions of known cancer driver genes. Included are the type of *MBD4*-mutant cancer, the specific driver gene mutation, the genomic co-ordinates of the driver gene mutation (hg38), the reference and alternate alleles of the mutation and the consequences of the mutation according to the Sorting Intolerant from Tolerant (SIFT) or Polymorphism Phenotyping (Polyphen) variant effect predictors. NA = Not Applicable.

Table 4.5, each polyp harboured an *APC*^{R1450*} truncation, as well as some polyps harbouring unique *APC* truncations. For example, one of the polyps also presented with *APC*^{R234*} and *APC*^{R387*} in addition to the common *APC*^{R1450*} mutation. Furthermore, two of the polyps also presented with pathogenic mutations in the suspected CRC driver genes Epidermal Growth Factor Receptor (*EGFR*), *APC* Membrane Recruitment Protein 1 (*AMER1*) and F-Box & WD Repeat Domain-Containing 7 (*FBXW7*). A list of the CRC-specific driver gene mutations present in these polyps is presented in Table 4.5.

The presence of pathogenic C → T mutations at CpG sites in the protein-coding sequence of key cancer driver genes in these *MBD4*-mutant cancers and colorectal polyps perhaps suggests a mechanism by which germline loss of *MBD4* may act to drive tumorigenesis in a number of different cancer types. If this is truly the case, germline *MBD4* mutations may therefore represent a novel cancer predisposition syndrome associated with the development of several tumours, including breast cancer, sarcoma, uveal melanoma and possibly CRC – given that the *MBD4*-mutant polyps presented with C → T mutations at CpG sites within the *APC* coding sequence. It is possible, given enough time for unrepaired spontaneous deaminations of 5-mC to accumulate, that these polyps may develop into CRC as a result of C → T mutations at CpG sites in the protein-coding regions of other CRC driver genes. While this data potentially provide a mechanism by which cancer-specific driver genes may mutate in individuals with germline mutations in *MBD4,* further work may be required to determine if there are any factors that influence the likelihood of a CpG site undergoing spontaneous deamination, factors which can be used to identify CpG sites – and potentially genes – that are at high risk of mutation via this mechanism in individuals with germline *MBD4* mutations.

4.3.3 – C → T Mutations are Enriched at Highly-Methylated CpG Sites

Following the analyses presented above, it could be concluded that germline mutations in *MBD4*, both biallelic and monoallelic with loss of the WT allele, drive an increase in C → T mutagenesis at CpG sites, as is the case in *MBD4*-mutant breast cancer, sarcoma, uveal melanoma and colorectal polyps (see section 4.3.1). However, the same is not seen in cancers with somatic mutations in *MBD4,* including cancers with likely biallelic somatic truncations. These C → T mutations at CpG sites seen in cancers with germline mutations in *MBD4* and subsequent LoH included pathogenic mutations in cancer-specific driver genes, suggesting that the loss of *MBD4* may drive tumorigenesis in these cancers via unrepaired spontaneous deamination of 5-mC within the protein-coding sequences of cancer driver genes. However, as discussed in section 4.1.4, there are a number of factors that have been suggested to influence the likelihood of an unrepaired spontaneous deamination being propagated into a C → T mutation at a CpG site. These include DNA methylation, as it is thought that more highly-methylated CpG sites are at higher risk of deamination (452,453). Other factors include replication timing – where CpGs in later-replicating regions of the genome are thought to be at higher risk due to less efficient repair mechanisms and the transcription strand a CpG site lies on – where CpGs on the coding strand are thought to be more

Polyp:	Gene Mutation:	Position (hg38):	Reference Allele:	Alternate Allele:	SIFT:	PolyPhen:
Descending Colon 1	*APC*^{R232*}	Chr5:112,792,494	C	T	NA	NA
Descending Colon 1	*APC*^{R405*}	Chr5:112,819,245	C	T	NA	NA
Descending Colon 1	*APC*^{R1450*}	Chr5:112,839,942	C	T	NA	NA
Descending Colon 2	*APC*^{R564*}	Chr5:112.828,919	C	T	NA	NA
Descending Colon 2	*APC*^{R876*}	Chr5:112,838,220	C	T	NA	NA
Descending Colon 2	*APC*^{R1450*}	Chr5:112,839,942	C	T	NA	NA
Descending Colon 2	*EGFR*^{R962C}	Chr7:55,200,351	C	T	Deleterious	Probably Damaging
Descending Colon 3	*FBXW7*^{R222*}	Chr4:152,346,992	G	A	NA	NA
Descending Colon 3	*APC*^{R1450*}	Chr5:112,839,942	C	T	NA	NA
Descending Colon 3	*AMER1*^{R626*}	ChrX:64,191,411	G	A	NA	NA
Sigmoid Colon	*APC*^{R283*}	Chr5:112,815,507	C	T	NA	NA
Sigmoid Colon	*APC*^{R405*}	Chr5:112,819,245	C	T	NA	NA
Sigmoid Colon	*APC*^{R1450*}	Chr5:112,839,942	C	T	NA	NA

Table 4.5 – Pathogenic C → T Mutations at CpG Sites in *MBD4*-Mutant Colorectal Polyps: Details of pathogenic C → T mutations in the coding regions of known colorectal cancer driver genes in *MBD4*-mutant colorectal polyps. Included are the polyp number, the specific driver gene mutation, the genomic co-ordinates of the driver gene mutation (hg38), the reference and alternate alleles of the mutation and the consequences of the mutation according to the Sorting Intolerant from Tolerant (SIFT) or Polymorphism Phenotyping (Polyphen) variant effect predictors. NA = Not Applicable.

vulnerable to deamination as they are exposed as single-stranded DNA compared to the template strand, which has been suggested to be shielded from deamination by transcription-associated proteins (435,453,455,456).

To further investigate this, fractional methylation data from whole-genome bisulphite sequencing of the normal sigmoid colon was combined with replication timing data from the CRC cell line HCT116 or transcription strand data obtained from Gencode (465,475). Tissue-specific fractional methylation and replication timing data could not be obtained for breast, soft tissue or melanocytes – therefore methylation and replication timing analysis was restricted to CRC and the *MBD4*-mutant colorectal polyps. Using this data, each CpG site in the genome could be assigned to one of twelve DNA methylation bins, ranging from 0 (non-methylated) to 1 (highly-methylated), and C → T mutations in cancers or polyps could subsequently be assigned to one of these bins. As seen in Figure 4.13, there was a significant positive correlation between DNA methylation and the rate of C → T mutagenesis at CpG sites in MSS CRCs and the *MBD4*-mutant colorectal polyps. These significant correlations were observed in MSS *MBD4*-WT CRCs (r^2 =0.7275, p = 0.000421), CRCs with somatic heterozygous *MBD4* truncations (r^2 = 0.6546, p = 0.001435) and MSS CRCs with germline heterozygous *MBD4* truncations (r^2 = 0.7508, p = 0.000266). Interestingly, the most significant association between DNA methylation and the rate of C → T mutagenesis at CpG sites was seen in the *MBD4*-mutant colorectal polyps (r^2 = 0.9319, p < 0.00001 – Figure 4.13). Furthermore, the regression slope (α) for the association between DNA methylation and the C → T mutation rate at CpG sites was significantly higher in these *MBD4*-mutant polyps (α = 1130.6) compared to *MBD4*-WT CRCs (α = 78.6, p < 0.0001). The slopes of the MSS CRCs with somatic (α = 84.1) or germline (α = 116.1) were not significantly greater than their *MBD4*-WT counterparts (p = 0.827 and p = 0.166 respectively). Despite the difference in regression slope, the regression constant, which represents the C → T mutation rate when methylation is zero, was not significantly different in the *MBD4*-WT cancers and the *MBD4*-mutant colorectal polyps (p = 0.516), suggesting that it is more highly-methylated CpG sites that are at greatest risk of C → T mutagenesis in these polyps. The fact that both the r^2 correlation and the regression slope between DNA methylation and the rate of C → T mutagenesis at CpG sites was greater in the *MBD4*-mutant polyps than in *MBD4*-WT CRCs suggests that highly-methylated CpG sites may be more at risk of C → T mutagenesis via unrepaired spontaneous deaminations of 5-mC propagating into C → T mutations as a consequence of *MBD4* deficiency.

When the same analysis was performed in MSI^+ CRCs (see Figure 4.14), significant positive correlations between DNA methylation and the rate of C → T mutagenesis at CpG sites were identified in *MBD4*-WT CRCs (r^2 = 0.755, p = 0.000244), CRCs with somatic (r^2 = 0.7284, p = 0.000413) or germline (r^2 = 0.826, p = 0.000043) heterozygous *MBD4* truncations and cancers with likely biallelic somatic *MBD4* truncations (r^2 = 0.8002, p = 0.000086). However, the regression slope of MSI^+ *MBD4*-WT CRCs (α = 390) was not significantly different compared to CRCs with somatic heterozygous truncations (α = 396.1, p = 0.953), germline heterozygous truncations (α = 342.4, p = 0.587) or somatic likely biallelic truncations (α = 516.3, p = 0.254) in *MBD4.* This indicates that while there may be a significant correlation between DNA methylation and the rate of C → T mutagenesis at CpG sites in MSI^+ CRCs, truncations in *MBD4* do not increase the strength of this association,

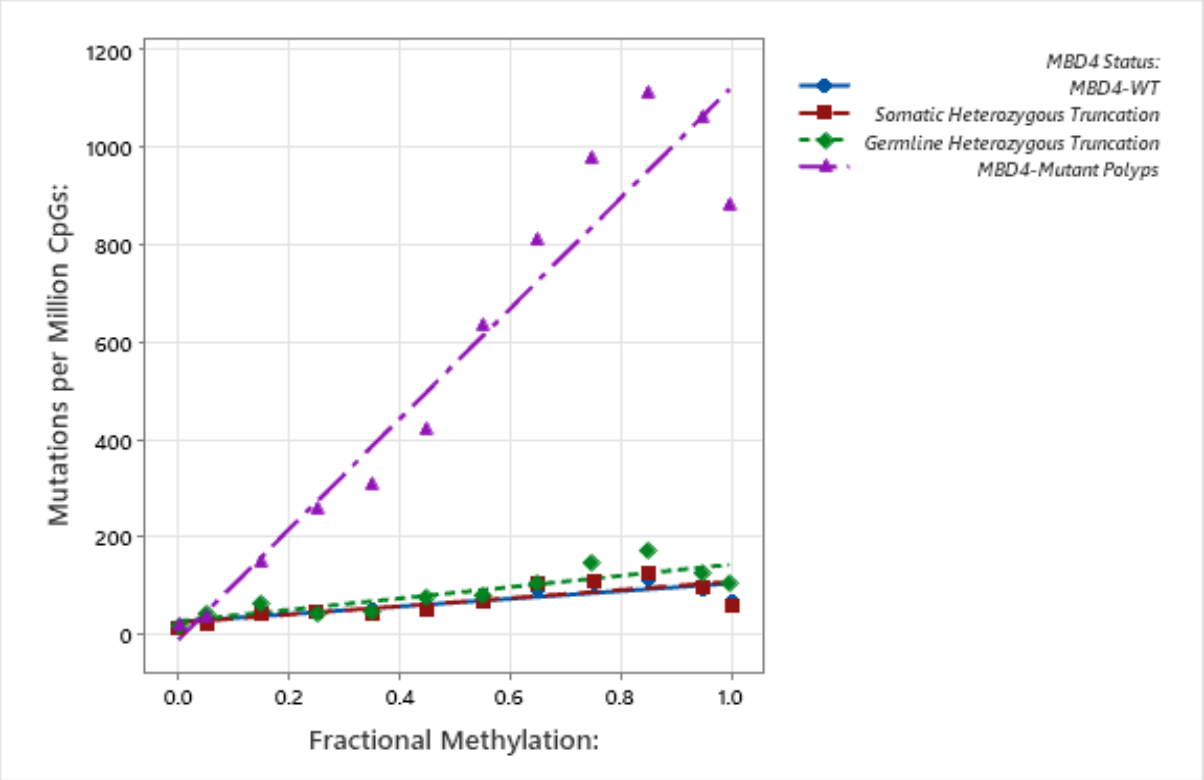


Figure 4.13 – The Relationship Between DNA Methylation and C → T Mutation Rates in MSS Colorectal Cancer & *MBD4*-Mutant Colorectal Polyps: The association between fractional DNA methylation and the rate of C → T mutagenesis at CpG sites in MSS *MBD4* wild-type colorectal cancer, colorectal cancers with somatic heterozygous *MBD4* truncations, colorectal cancers with germline heterozygous *MBD4* truncations and *MBD4*-mutant colorectal polyps.

further indicating that only biallelic germline loss of *MBD4* drives an increase of C → T mutagenesis of highly-methylated CpG sites.

4.3.4 – C → T Mutations at CpG Sites are Enriched in Late-Replicating DNA

In addition to the DNA methylation status of a CpG site, recent evidence has suggested that DNA in later-replicating regions of the genome is at greater risk of mutation than DNA in early-replicating regions (454). This is thought to be due to the differential activity of the DNA MMR pathway throughout the cell cycle, with early-replicating DNA being thought to benefit from the highest degree of MMR activity (454). To investigate if DNA replication timing was also a factor that could influence the likelihood of a C → T mutation at a CpG site, Repli-Seq replication timing data was combined with the fractional DNA methylation described above to divide the genome into four replication timing bins – Earliest, Early Late and Latest.

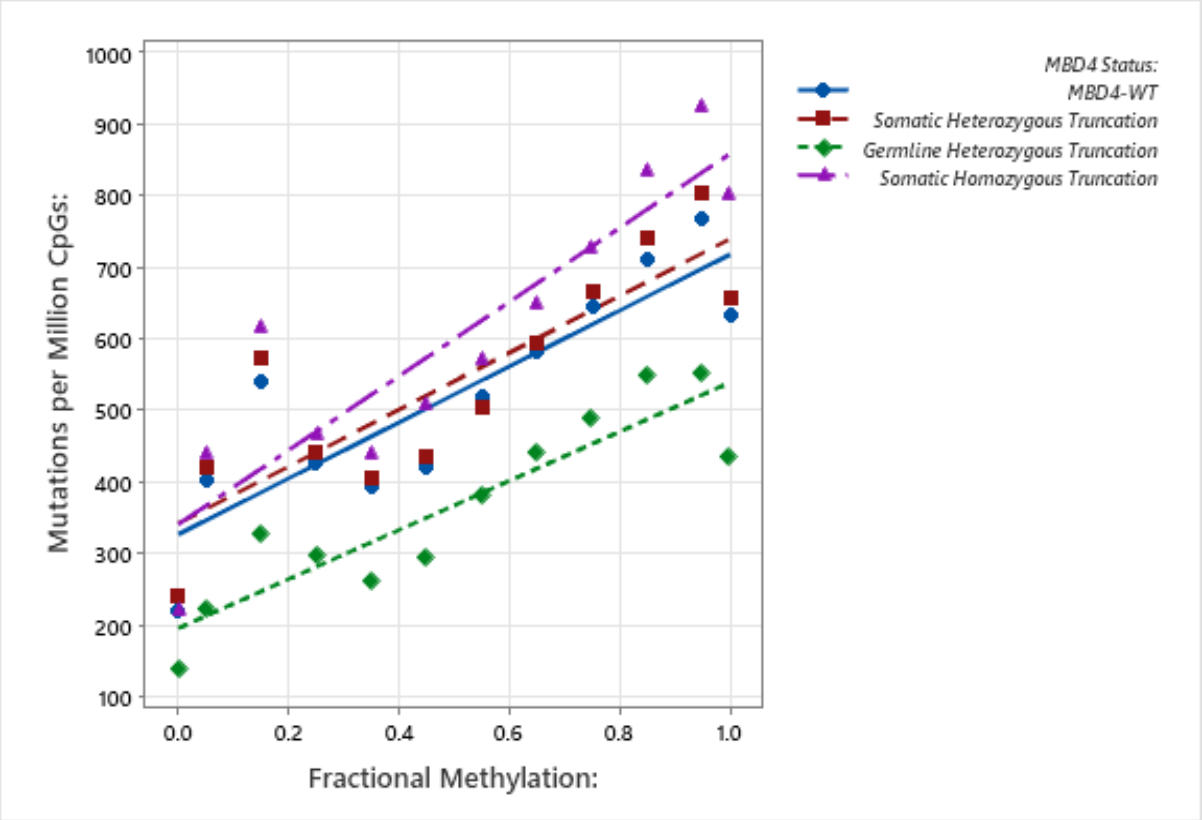


Figure 4.14 – The Relationship Between DNA Methylation and C → T Mutation Rates in MSI$^+$ Colorectal Cancer: The association between fractional DNA methylation and the rate of C → T mutagenesis at CpG sites in MSI$^+$ *MBD4* wild-type colorectal cancer, colorectal cancers with somatic heterozygous *MBD4* truncations, colorectal cancers with germline heterozygous *MBD4* truncations and colorectal cancers with likely somatic biallelic *MBD4* truncations.

In MSS CRCs and the *MBD4*-mutant colorectal polyps, a significant enrichment of mutations in later-replicating DNA was identified across all groups. In *MBD4*-WT MSS CRCs, there were significant correlations between fractional DNA methylation and C → T mutation rates in all four replication timing bins (Figure 4.15a, Table 4.6). Further analysis of these correlations revealed a significant cumulative increase in the methylation-mutation rate regression slope in each of the three other replication timing bins compared to the earliest replicating bin (Methylation x Replication Timing = 28.95, $p < 0.0001$, Table 4.7). The same correlations between methylation and the rate of C → T mutagenesis at CpG sites were identified in all four replication timing bins of MSS CRCs with somatic heterozygous *MBD4* truncations (Figure 4.15b, Table 4.6). Like *MBD4*-WT cancers, there was also a significant cumulative increase in regression slope in the other replication timing bins compared to earliest-replicating bin (Methylation x Replication Timing = 33.88, $p = 0.001$, Table 4.7). In CRCs with germline heterozygous *MBD4* truncations, significant correlations were identified between methylation and C → T mutagenesis at CpG sites in all four replication timing bins (Figure 4.15c, Table 4.6). There was also a significant cumulative increase in the regression slope of this association in later-replicating regions of the genome (Methylation x Replication Timing = 35, $p = 0.002$, Table 4.7).

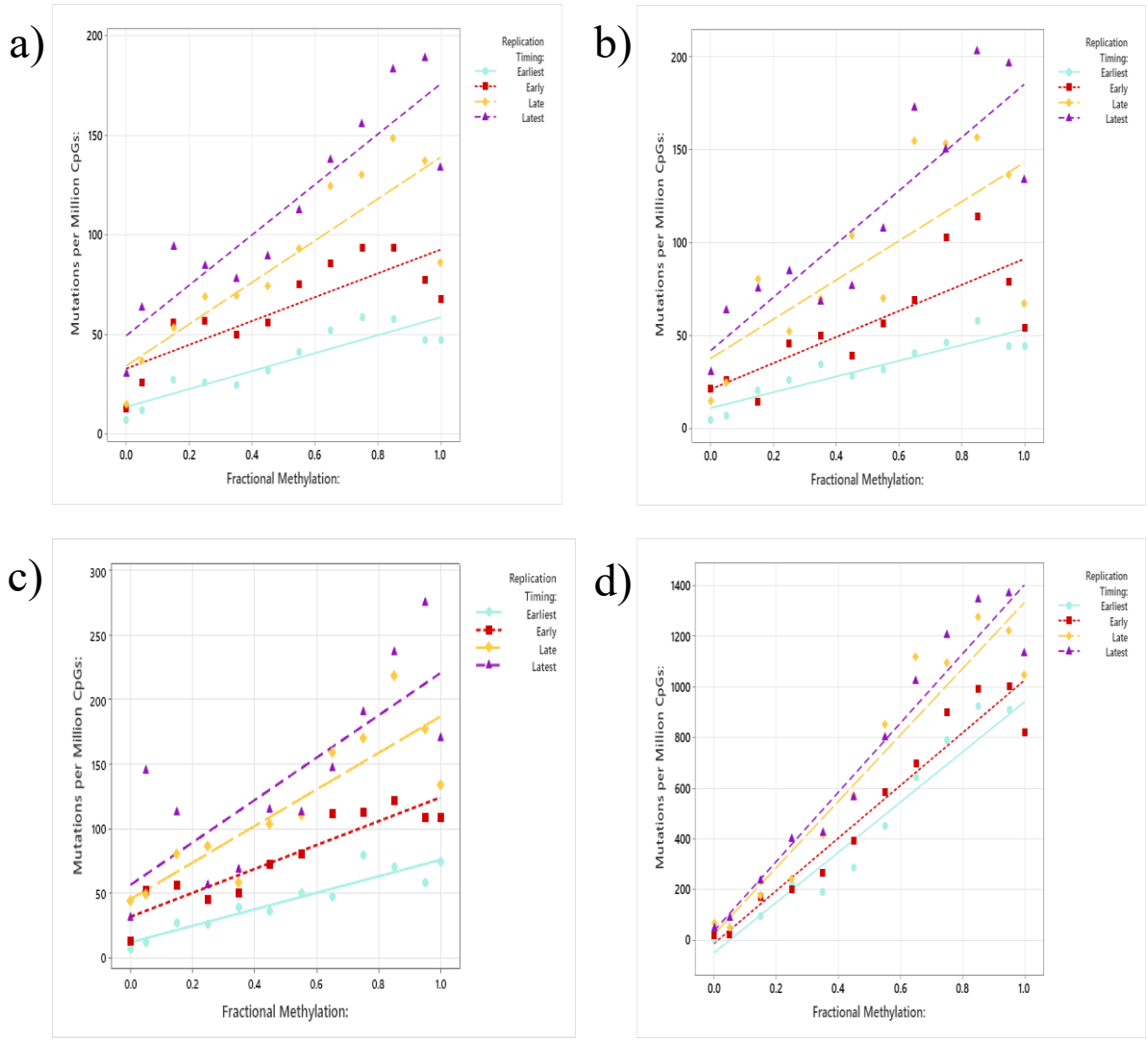


Figure 4.15 – The Effect of DNA Replication Timing on C → T Mutagenesis at CpG Sites of MSS Colorectal Cancers and *MBD4*-Mutant Colorectal Polyps: Graphs comparing the relationship between fractional DNA methylation and C → T mutation rates in the earliest (blue), early (red), late (yellow) and latest (purple) replicating regions of the genome. Included are analyses of MSS *MBD4* wild-type colorectal cancer (a), colorectal cancers with somatic heterozygous *MBD4* truncations (b), colorectal cancers with germline heterozygous *MBD4* truncations (c) and *MBD4*-mutant colorectal polyps (d).

In the *MBD4*-mutant colorectal polyps, there were significant correlations between DNA methylation and C → T mutation rates in all four replication timing bins (Figure 4.15d, Table 4.6), with replication timing significantly increasing the slope of this regression equation (Methylation x Replication Timing = 140.2, p = 0.003 – Table 4.7). The detailed regression analysis of each group of cancers are presented in Tables 4.6 and 4.7.

As discussed above, there is a suggestion that the MMR pathway is not as active in late-replicating regions of the genome as it is in early-replicating regions, which is thought to

MSS *MBD4*-WT (n = 1,524):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	***r*^{2}:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	45	13.51	0.8097	0.8998	0.000067
Early	59.8	32.58	0.676	0.8222	0.00103
Late	104.8	33.96	0.7829	0.8848	0.000131
Latest	126.4	49.2	0.807	0.8983	0.000072
MSS *MBD4* Somatic Heterozygous Truncation (n = 3):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	***r*^{2}:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	42.4	10.8	0.836	0.9143	0.000031
Early	70.3	20.7	0.5968	0.7725	0.003228
Late	105.6	37.3	0.509	0.7134	0.009186
Latest	143.3	41.6	0.7541	0.8684	0.000248
MSS *MBD4* Germline Heterozygous Truncation (n = 5):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	***r*^{2}:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	63.9	11.8	0.8738	0.9348	< 0.00001
Early	92.5	31.3	0.8479	0.9208	0.000021
Late	141.6	44.9	0.763	0.8735	0.000206
Latest	164.2	56	0.6217	0.7885	0.002308
***MBD4*-Mutant Colorectal Polyps (n = 4):**					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	***r*^{2}:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	990.8	-50.2	0.9316	0.9652	< 0.00001
Early	1040.1	-15.1	0.9211	0.9597	< 0.00001
Late	1311.2	21.5	0.8983	0.9478	< 0.00001
Latest	1367.8	34.9	0.9289	0.9638	< 0.00001

Table 4.6 – Regression Analysis of MSS Colorectal Cancers & *MBD4*-Mutant Colorectal Polyps: The regression equations of the relationship between DNA methylation and C → T mutation rates at CpG sites in MSS colorectal cancers and *MBD4*-mutant colorectal polyps. Included are the replication timing bin, the regression slope, the regression constant, the methylation-mutation rate r^2 correlation, the Pearson's R correlation measure of this relationship and the p-value associated with the Pearson's R statistic ($p_{(Pearson's\ R)}$).

MSS *MBD4*-WT (n = 1,524):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	40.6	0.002
Replication Timing	10.84	0.008
Methylation x Replication Timing Interaction Term	28.95	< 0.0001
Constant	16.05	0.034
MSS *MBD4* Somatic Heterozygous Truncation (n = 3):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	39.6	0.03
Replication Timing	10.89	0.061
Methylation x Replication Timing Interaction Term	33.88	0.001
Constant	11.3	0.294
MSS *MBD4* Germline Heterozygous Truncation (n = 5):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	63.1	0.003
Replication Timing	14.62	0.0028
Methylation x Replication Timing Interaction Term	35	0.002
Constant	14.1	0.249
***MBD4*-Mutant Colorectal Polyps (n = 4):**		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	967.1	< 0.0001
Replication Timing	29.2	0.276
Methylation x Replication Timing Interaction Term	140.2	0.003
Constant	-46	0.358

Table 4.7 – The Effect of Replication Timing on Methylation-Mutation Rate Associations in MSS Colorectal Cancer & *MBD4*-Mutant Colorectal Polyps: Regression equations for the effect of fractional DNA methylation and replication timing on C → T mutation rates at CpG sites. Included are the variables of the equation – DNA methylation, replication timing, the interaction term between the two, the regression constant and the p-value of each ($p_{(Coefficient)}$).

drive the increase in mutation rates in later-replicating regions. Therefore in MSI^+ CRCs, which possess defective DNA MMR, this replication timing dependent increase in methylation-mutation rate regression slopes was not observed – due to the absence of functional MMR in all replication timing bins in these cancers. In MSI^+ *MBD4*-WT CRCs, while there were significant correlations between DNA methylation and C → T mutation rates in three replication timing bins (Figure 4.16a, Table 4.8), there was not a significant correlation in the latest replication timing bin (r^2 = 0.1728, p = 0.178955). In addition to this, unlike in MSS CRCs, replication timing had no effect on the slope of the methylation-mutation rate association (Methylation x Replication Timing = -48.1, p = 0.279 – Table 4.9).

The same trend was identified in MSI^+ CRCs with somatic heterozygous *MBD4* truncations (Figure 4.16b), where all replication timing bins except the latest (r^2 = 0.1333, p = 0.243228 – Table 4.8) presented with significant methylation-mutation rate correlations. There was also no significant effect of replication timing on this association (Methylation x Replication Timing = -56.6, p = 0.24 – Table 4.9). In MSI^+ CRCs with germline heterozygous *MBD4* truncations, significant correlations between DNA methylation and C → T mutation rates were observed in all four replication timing bins (Figure 4.16c, Table 4.8), but replication timing had no effect on the slope of these correlations (Methylation x Replication Timing = 22.1, p = 0.46 – Table 4.9). Finally, the MSI^+ CRCs with likely somatic biallelic *MBD4* truncations (Figure 4.16d) presented with significant associations in three of the four

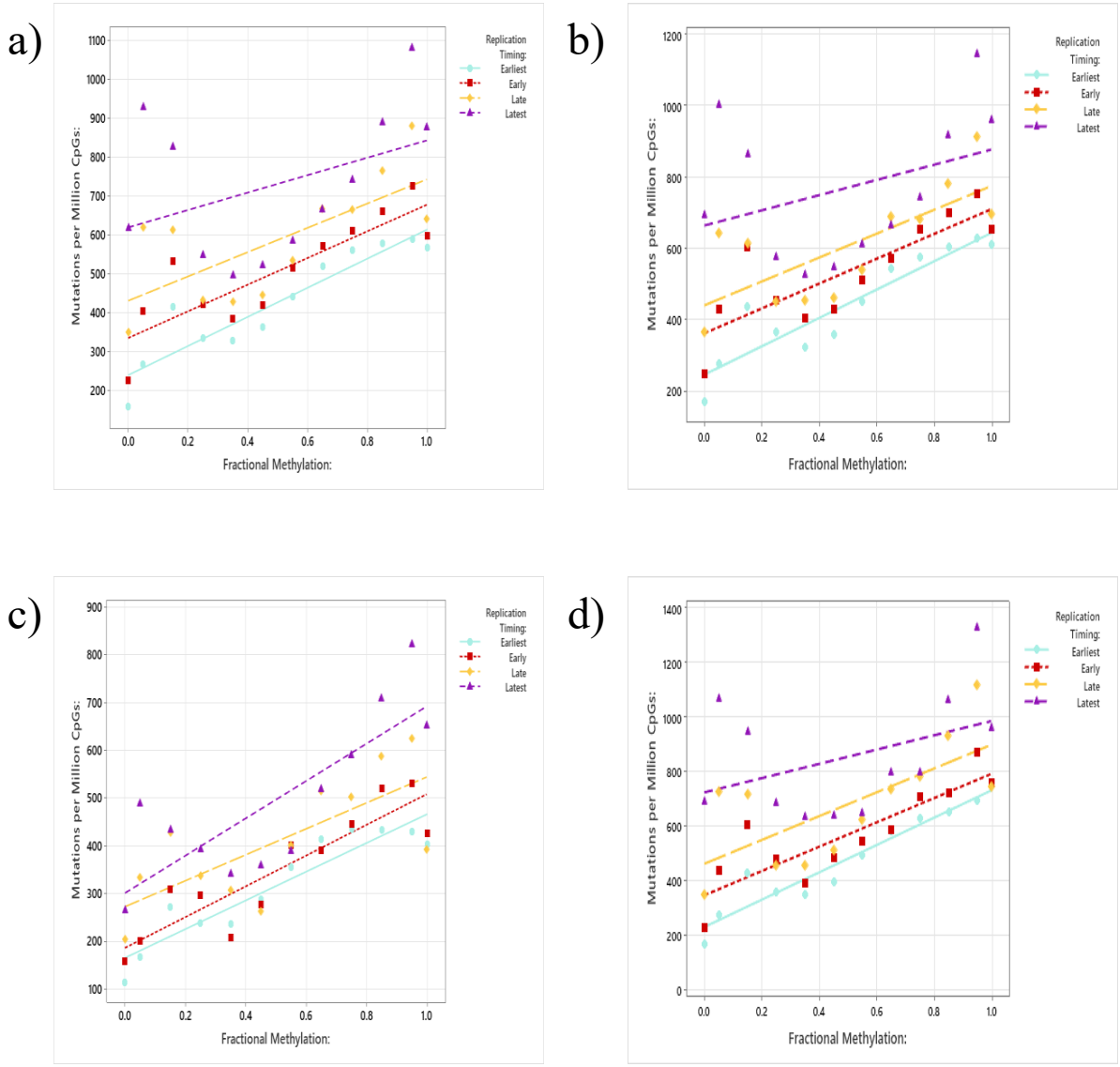


Figure 4.16 – The Effect of DNA Replication Timing on C → T Mutagenesis at CpG Sites of MSI$^+$ Colorectal Cancer: Graphs comparing the relationship between fractional DNA methylation and C → T mutation rates in the earliest (blue), early (red), late (yellow) and latest (purple) replicating regions of the genome. Included are analyses of MSI *MBD4* wild-type colorectal cancer (a), colorectal cancers with somatic heterozygous *MBD4* truncations (b), colorectal cancers with germline heterozygous *MBD4* truncations (c) and colorectal cancers with likely biallelic somatic *MBD4* truncations (d).

replication timing bins, with a non-significant association in the latest replicating DNA (r^2 = 0.169, p = 0.184304 – Table 4.8). Like all the MSI$^+$ CRCs before, there was no significant effect of replication timing on the methylation-mutation rate regression equation (Methylation x Replication Timing = -72.6, p = 0.189 – Table 4.9). The complete regression analysis of these MSI$^+$ CRCs are presented in Tables 4.8 and 4.9.

Overall, this data suggests that C → T mutations are not only more likely at highly-methylated CpG sites but also CpG sites in late-replicating regions of the genome. In MSS CRC this was apparent by the significant increase in the methylation-mutation rate slope in later-replicating DNA. However, in MSI$^+$ CRC, the effect of replication timing appears to be

MSI$^+$ *MBD4*-WT (n = 236):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	374	239.3	0.8633	0.9291	0.000013
Early	343	334.1	0.7318	0.8555	0.000387
Late	313	430	0.4882	0.6987	0.011475
Latest	224	618.9	0.1728	0.4157	0.178955
MSI$^+$ *MBD4* Somatic Heterozygous Truncation (n = 107):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	397	245.8	0.8551	0.9247	0.000017
Early	348	360.7	0.6775	0.8231	0.001005
Late	335	439.1	0.5317	0.7292	0.007124
Latest	213	663.4	0.1333	0.3651	0.243228
MSI$^+$ *MBD4* Germline Heterozygous Truncation (n = 2):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	301	165.1	0.8672	0.9312	0.000011
Early	322	185.3	0.8001	0.8945	0.000086
Late	272	271.6	0.5213	0.722	0.008015
Latest	391	300.3	0.6566	0.8103	0.001393
MSI$^+$ *MBD4* Somatic Homozygous Truncation (n = 12):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	501	232.3	0.9201	0.9592	< 0.00001
Early	445	347.2	0.7462	0.8638	0.000292
Late	436	463	0.4862	0.6973	0.011714
Latest	262	725	0.169	0.4111	0.184304

Table 4.8 – Regression Analysis of MSI$^+$ Colorectal Cancer: The regression equations of the relationship between DNA methylation and C → T mutation rates at CpG sites in MSI$^+$ colorectal cancers. Included are the replication timing bin, the regression slope, the regression constant, the methylation-mutation rate r^2 correlation, the Pearson's R correlation measure of this relationship and the p-value associated with the Pearson's R statistic ($p_{(Pearson's\ R)}$).

MSI$^+$ *MBD4*-WT (n = 236):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	385.8	< 0.0001
Replication Timing	123.5	< 0.0001
Methylation x Replication Timing Interaction Term	-48.1	0.279
Constant	220.4	< 0.0001
MSI$^+$ *MBD4* Somatic Heterozygous Truncation (n = 107):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	408.3	< 0.0001
Replication Timing	133.1	< 0.0001
Methylation x Replication Timing Interaction Term	-56.6	0.24
Constant	227.6	< 0.0001
MSI$^+$ *MBD4* Germline Heterozygous Truncation (n = 2):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	288.4	< 0.0001
Replication Timing	49.2	0.008
Methylation x Replication Timing Interaction Term	22.1	0.46
Constant	156.8	< 0.0001
MSI$^+$ *MBD4* Somatic Homozygous Truncation (n = 12):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	520	< 0.0001
Replication Timing	159.4	< 0.0001
Methylation x Replication Timing Interaction Term	-72.6	0.189
Constant	202.8	0.002

Table 4.9 – The Effect of Replication Timing on Methylation-Mutation Rate Associations in MSI$^+$ Colorectal Cancer: Regression equations for the effect of fractional DNA methylation and replication timing on C → T mutation rates at CpG sites. Included are the variables of the equation – DNA methylation, replication timing, the interaction term between the two, the regression constant and the p-value of each ($p_{(Coefficient)}$).

of less importance. While replication timing may not have a significant effect on the slope of the association between DNA methylation and mutation rate in MSI$^+$ CRCs, there were still more mutations in these late-replicating regions of the genome compared to the early-replicating regions. Therefore, it could be concluded increasing DNA methylation and later replication timing, particularly in MSS CRCs, are two factors that may increase the rate of C → T mutagenesis at CpG sites.

4.3.5 – Spontaneous Deamination of 5-mC is not Influenced by Transcription Strand

As well as the DNA methylation and replication timing bin of a CpG site, previous data has suggested that the transcription strand of a CpG site may also influence the likelihood of C → T mutagenesis via unrepaired spontaneous deamination (437,456). As discussed previously, single-stranded DNA is at higher risk of spontaneous deamination than double-stranded DNA. It has previously been suggested that during transcription, the single-stranded DNA of the template strand is shielded by transcription-associated proteins, reducing the likelihood of deamination (437,455,456). Therefore, it has been hypothesised that single-stranded DNA of

the transcriptional coding strand, which is not protected by these proteins, remains vulnerable to deamination.

To investigate if this was the case in CRC and the *MBD4*-mutant colorectal polyps, transcription strand data obtained from Gencode was combined with the fractional DNA methylation data used previously (463,465,475). This provided an estimate of transcription direction for 15,316,904 CpG sites – allowing C → T mutations to be assigned to either the coding or template transcription strand. The total number of C → T mutations at CpG sites assigned to either the coding or template transcription strand for each cancer or polyp could then be used to calculate the $\log_2$(Coding/Template) ratio of mutations on each transcription strand. Surprisingly, in MSS CRCs and the *MBD4*-mutant colorectal polyps, this ratio of transcription strand mutations was not significantly different in any group (see Figure 4.17). This included the *MBD4*-mutant colorectal polyps, which presented with a $\log_2$(Coding/Template) ratio that was not significantly different to the MSS *MBD4*-WT CRCs ($p = 0.7$, Figure 4.17). When this analysis was extended to MSI^+ cancers, the same result was observed. As seen in Figure 4.18, the $\log_2$(Coding/Template) ratio was not significantly different in any of the *MBD4*-mutant MSI^+ CRCs compared to their *MBD4*-WT counterparts (Figure 4.18). This data may suggest that spontaneous deaminations are not more common on the transcriptional coding strand as first hypothesised, as there appears to be no difference in the transcription strand bias between *MBD4*-mutant and *MBD4*-WT samples.

In order to more closely investigate this phenomenon, the transcription strands of C → T mutations at the CpG sites of the *MBD4*-mutant breast cancers, sarcomas and uveal melanoma described in sections 4.3.1 and 4.3.2 were calculated and compared to the *MBD4*-mutant colorectal polyps. Surprisingly, there was no difference in the $\log_2$(Coding/Template) ratios of the *MBD4*-mutant colorectal polyps and any of the other *MBD4*-mutant cancer types ($p_{(Kruskal\text{-}Wallis)} = 0.8$, Figure 4.19). However, both the myxofibrosarcoma and sarcoma of unspecified sub-type did appear to show more of a bias towards the coding strand than any of the other *MBD4*-mutant cancers – although these differences were not significant.

In order to conclude if the transcription strand distribution of C → T mutations was affected by the DNA methylation status of a CpG site, C → T mutations on each transcription strand were binned into one of the twelve DNA methylation bins described above. Interestingly, no significant differences were identified between regression equations of the coding and template transcription strands in MSS CRCs or the *MBD4*-mutant polyps (Figure 4.20). In *MBD4*-WT cancers (Figure 4.20a), while there were still significant associations between C → T mutations and DNA methylation for both the coding ($r^2 = 0.6925$, $p = 0.000785$) and template ($r^2 = 0.6448$, $p = 0.001661$) strands, there was no difference in the slope of the regressions for the coding ($\alpha = 32.39$) and template ($\alpha = 33.44$) strands ($p = 0.918$). The same observations were made in CRCs with somatic heterozygous *MBD4* truncations (Figure 4.20b). Significant correlations between DNA methylation and C → T mutation rates were observed for the coding ($r^2 = 0.609$, $p = 0.002745$) and template ($r^2 = 0.5763$, $p = 0.004194$) transcription strands, however there was no significant difference in the regression slopes of the coding ($\alpha = 34.5$) or template ($\alpha = 34.83$) strands ($p = 0.979$). As seen in Figure 4.19b, there may be more mutations on the coding strand in these cancers due to an apparent difference in regression constant between the transcription strands. However, this difference was not significant ($p = 0.638$). In MSS CRCs with germline heterozygous *MBD4* truncations (Figure 4.20c), there were also significant correlations between DNA methylation and C → T

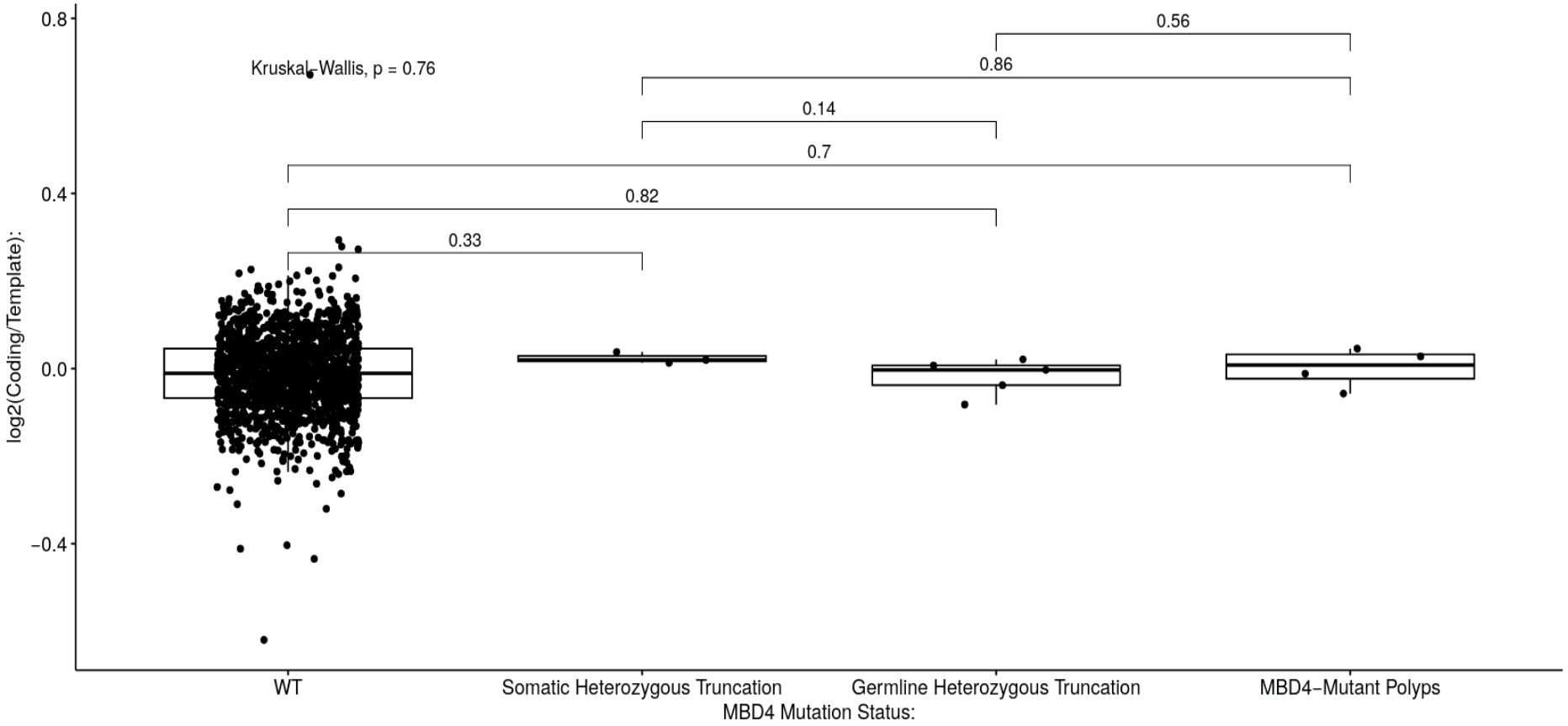


Figure 4.17 – Transcription Strand Bias of MSS Colorectal Cancer & *MBD4*-Mutant Colorectal Polyps: Boxplots showing the $\log_2$(Coding/Template) ratios of the number of C → T mutations at CpG sites on each transcription strand. Included are MSS *MBD4* wild-type colorectal cancer (WT), colorectal cancers with somatic heterozygous *MBD4* truncations, colorectal cancers with germline heterozygous *MBD4* truncations and *MBD4*-mutant colorectal polyps.

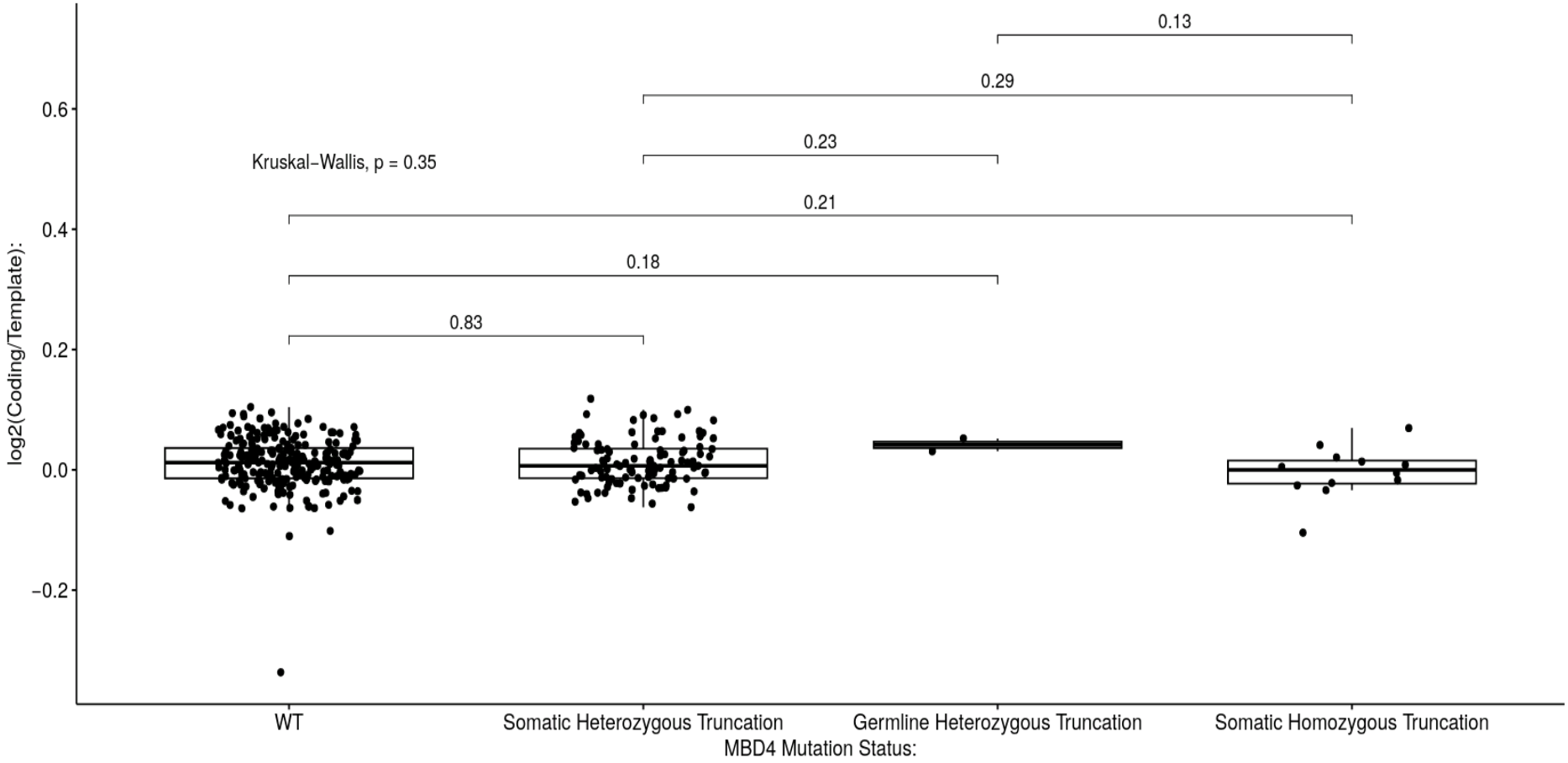


Figure 4.18 – Transcription Strand Bias of MSI$^+$ Colorectal Cancer: Boxplots showing the $\log_2$(Coding/Template) ratios of the number of C → T mutations at CpG sites on each transcription strand. Included are MSI$^+$ *MBD4* wild-type colorectal cancer (WT), colorectal cancers with somatic heterozygous *MBD4* truncations, colorectal cancers with germline heterozygous *MBD4* truncations and colorectal cancers with likely somatic biallelic *MBD4* truncations.

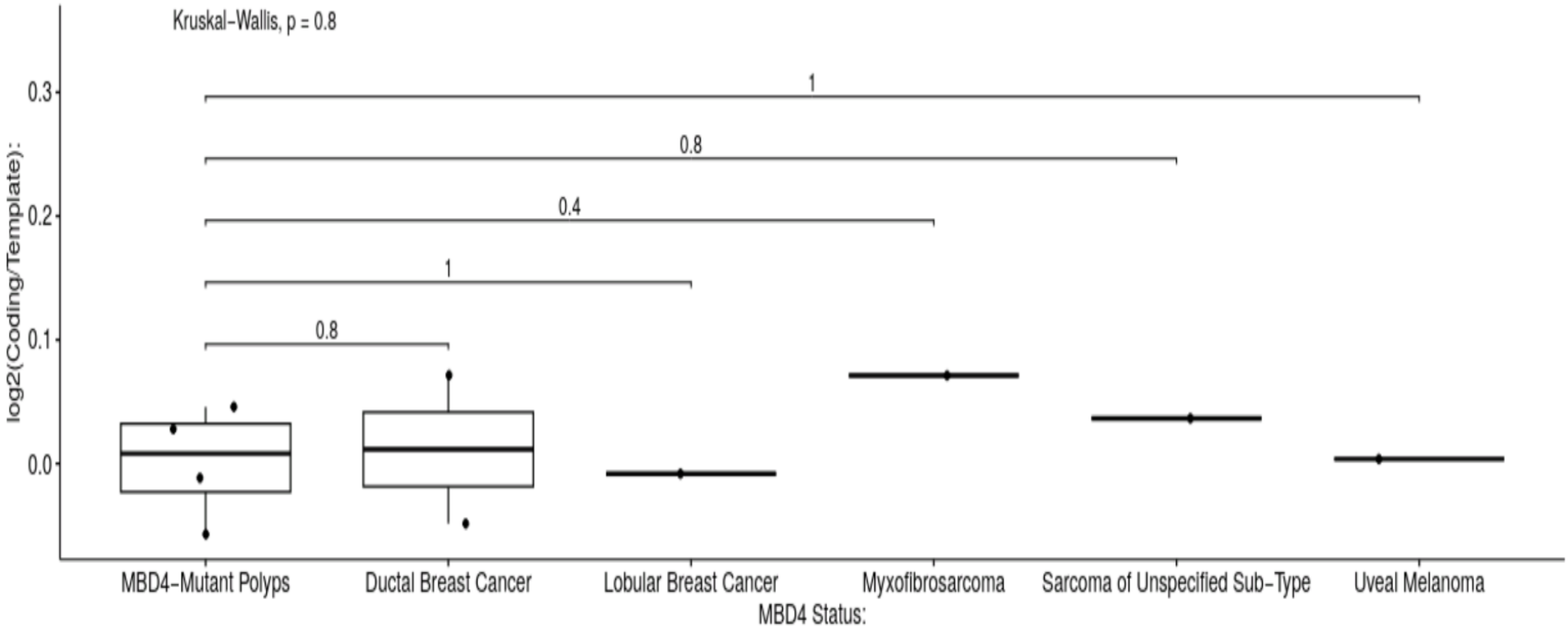


Figure 4.19 – Transcription Strand Biases of *MBD4*-Mutant Cancers & Colorectal Polyps: The $\log_2$(Coding/Template) ratio of C → T mutations at CpG sites of differing transcription strands. Shown are the transcription strand mutation ratios of *MBD4*-mutant colorectal polyps, ductal breast cancer, lobular breast cancer, myxofibrosarcoma, sarcoma of unspecified sub-type and uveal melanoma.

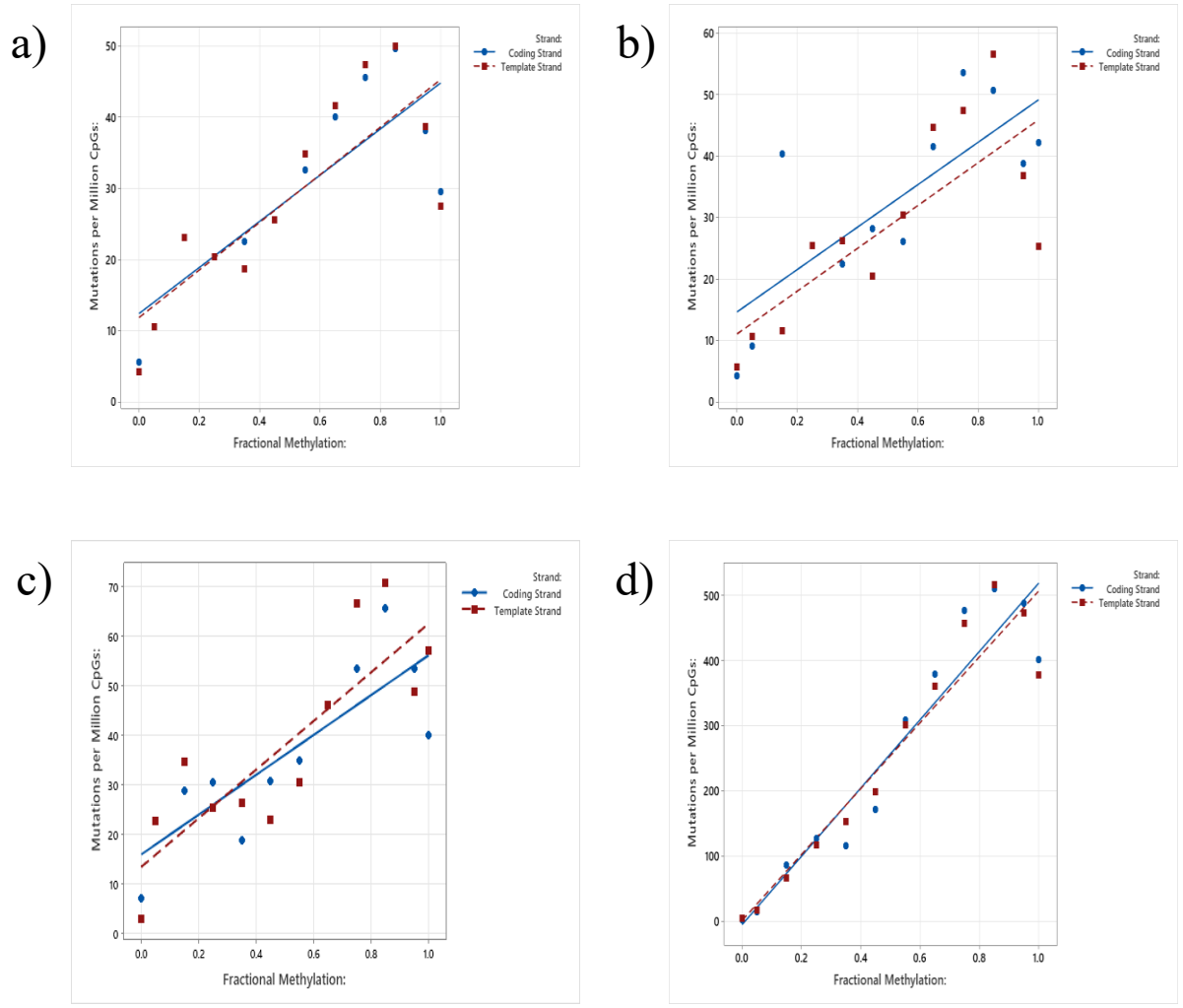


Figure 4.20 – Transcription Strand Bias Association with DNA Methylation in MSS Colorectal Cancer & *MBD4*-Mutant Colorectal Polyps: Charts showing the relationship between fractional DNA methylation and C → T mutation rates on the coding (blue) and template (red) transcription strands. Included are charts for MSS *MBD4* wild-type colorectal cancer (a), colorectal cancers with somatic heterozygous *MBD4* truncations (b), colorectal cancers with germline heterozygous *MBD4* truncations (c) and *MBD4*-mutant colorectal polyps (d).

mutation rates for both the coding (r^2 = 0.7066, p = 0.000616) and template (r^2 = 0.7119, p = 0.000562) strands. Similarly to the cancers described above, there was no significant difference in the regression slope of the coding (α = 40.23) and template (α = 49.14) transcription strands (p = 0.495). When the same analysis was performed on the *MBD4*-mutant colorectal polyps (Figure 4.20d), significant correlations were observed between DNA methylation and the rate of C → T mutagenesis at CpG sites on both the coding (r^2 = 0.8902, p < 0.00001) and template (r^2 = 0.8916, p < 0.00001) transcription strands. However, similarly to the MSS CRCs, there was no significant difference between the regression slopes of the coding (α = 523.5) and template (α = 505.3) strands (p = 0.804).

When this analysis was also performed in the context of MSI^+ CRCs (Figure 4.21), there were significant associations between C → T mutations and DNA methylation on both transcription strands but no significant differences in the regression slopes of each strand. In MSI^+ *MBD4*-WT CRCs (Figure 4.21a), there were significant correlations between DNA

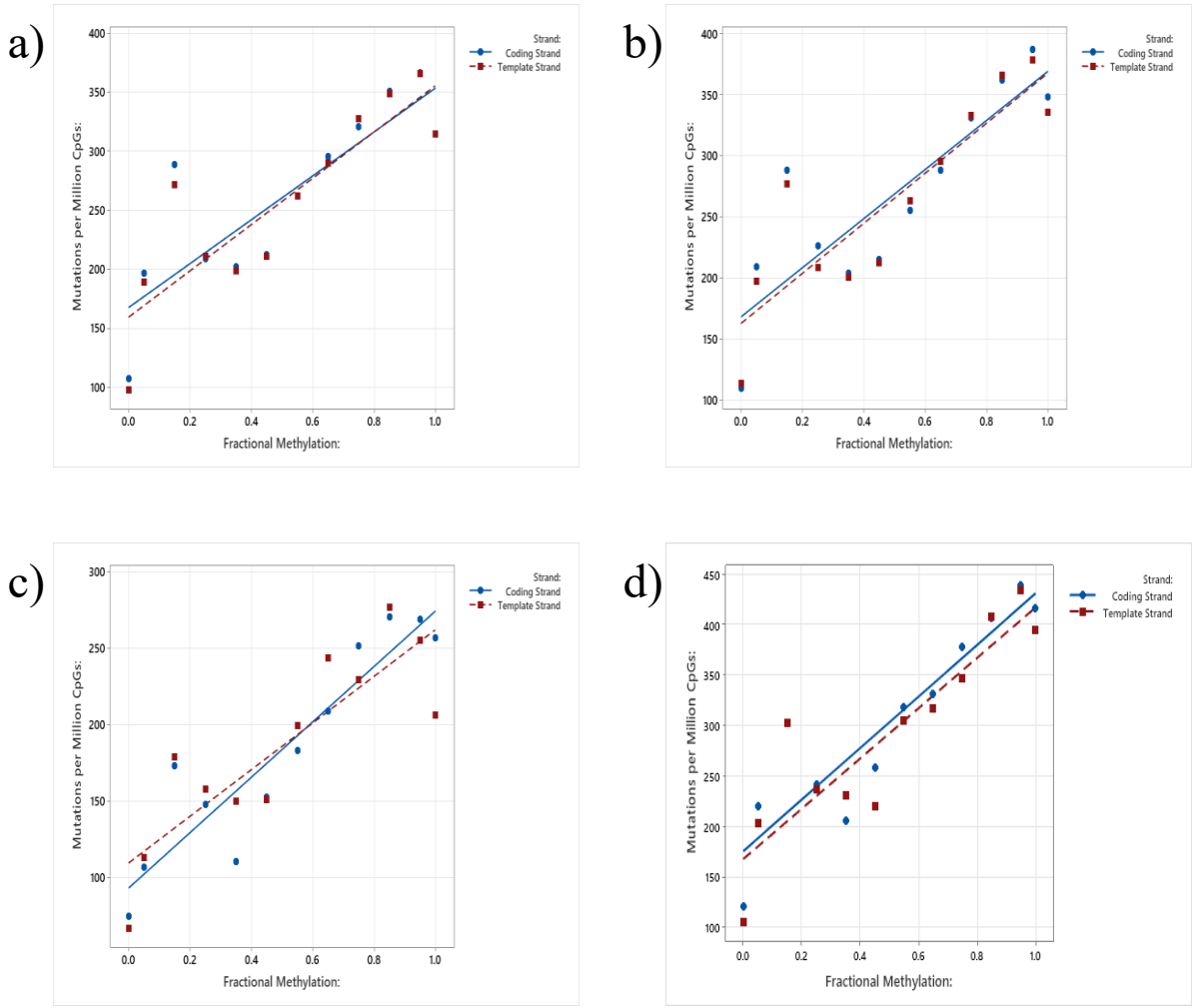


Figure 4.21 – Transcription Strand Bias Association with DNA Methylation in MSI$^+$ Colorectal Cancer: Charts showing the relationship between fractional DNA methylation and C → T mutation rates on the coding (blue) and template (red) transcription strands. Included are charts for MSI$^+$ *MBD4* wild-type colorectal cancer (a), colorectal cancers with somatic heterozygous *MBD4* truncations (b), colorectal cancers with germline heterozygous *MBD4* truncations (c) and colorectal cancers with likely biallelic somatic *MBD4* truncations (d).

methylation and C → T mutation rates on both the coding (r^2 = 0.7208, p = 0.000477) and template (r^2 = 0.7627, p = 0.000207) transcription strands. However the regression slopes of the coding (α = 185.8) and template (α = 196.1) strands were not significantly different from each other (p = 0.842). In MSI$^+$ CRCs with somatic heterozygous *MBD4* truncations (Figure 4.21b), there were also significant correlations between DNA methylation and C → T mutation rates on both the coding (r^2 = 0.7486, p = 0.000278) and template (r^2 = 0.7775, p = 0.000149) strands. Like *MBD4*-WT cancers, there was no significant difference in the regression slope of the coding (α =197.8) and template (α =201.1) strands (p = 0.948). The same trend was apparent in MSI$^+$ CRCs with germline heterozygous *MBD4* truncations (Figure 4.21c), where there were significant correlations between DNA methylation and the rate of C → T mutagenesis at CpG sites on both the coding (r^2 = 0.8542, p = 0.000017) and template (r^2 = 0.7348, p = 0.000366) transcription strands. While in these cancers the regression slope of the coding strand (α = 181.1) was greater than that of the template strand

(α = 152.6), this difference was not significant (p = 0.45). The cancers with likely somatic biallelic truncations in *MBD4* (Figure 4.21d) also presented with significant correlations between DNA methylation and C → T mutation rates on both the coding (r^2 = 0.8307, p = 0.000037) and template (r^2 = 0.8001, p = 0.000086) transcription strands. Similarly to Figure 4.21c, the regression slope of the coding strand (α =255.2) was greater than the regression slope of the template strand (α =249.1), but this difference was not significant (p = 0.91).

In summary, it appears that the likelihood of spontaneous deamination of 5-mC is not influenced by transcription strand. If CpG sites on the coding strand were indeed more vulnerable to spontaneous deamination than those on the template strand as previously suggested, a large coding strand bias would have been expected in the *MBD4*-mutant colorectal polyps compared to the *MBD4*-WT cancers. However this bias was not observed. Therefore, from all the factors studied, it appears that only the DNA methylation and replication timing associated with a CpG site influence the likelihood of C → T mutagenesis at CpG sites via unrepaired spontaneous deaminations of 5-mC.

4.4 – Discussion

The spontaneous deamination of 5-mC to thymine represents a key mechanism of C → T mutagenesis at CpG sites in the human genome (295). As discussed in section 4.1.2, C → T mutations at CpG sites are associated with the mutation signature SBS1, a clock-like signature thought to be a consequence of the age-dependent accumulation of unrepaired spontaneous deaminations of 5-mC (425,476). This mutation signature is so prevalent in human cancer that, in a recent landscape paper of 2,023 whole-genome sequenced CRCs, 99% of cancers presented with SBS1 to some degree, with this signature on average accounting for 15.5% of all SNVs (295). In most instances, these spontaneous deaminations associated with SBS1 are repaired by a combination of *MBD4, TDG* and the BER pathway (see Figure 4.3). However, patients identified by Sanders *et al.* with germline biallelic truncations in *MBD4* presented with a greatly elevated mutation burden, predominantly consisting of C → T mutations at CpG sites (446). Interestingly, these AMLs presented with C → T mutations at CpG sites in the protein-coding sequences of cancer driver genes, indicating that *MBD4* mutations may drive tumorigenesis via the propagation of unrepaired spontaneous deaminations into pathogenic driver mutations.

In order to investigate this, a series of *MBD4*-mutant cancers and colorectal polyps were compared to their *MBD4*-WT counterparts. Recently, Degasperi *et al.* identified a novel mutation signature associated with *MBD4* deficiency, which they termed SBS96 (451). This mutation signature was remarkably similar to SBS1 but presented with a greater proportion of C → T mutations in the context C**C**G than in the context G**C**G. When the mutation signatures of the *MBD4*-mutant breast cancers, sarcomas and uveal melanoma identified by Degasperi *et al.* from 100KGP data and the *MBD4*-mutant colorectal polyps were extracted, each presented with a signature that resembled SBS96. In addition to this, these cancers and polyps presented with a greatly increased number of C → T mutations at CpG sites compared to their *MBD4*-WT counterparts – indicating that unrepaired spontaneous deaminations within these cancers and polyps were propagating into C → T mutations at CpG sites. Furthermore,

these *MBD4*-mutant cancers either showed no age difference or were younger than their *MBD4*-WT counterparts, meaning that this increase in C → T mutagenesis at CpG sites could not be explained by patient age.

However, CRCs presenting with somatic *MBD4* truncations, including twelve with likely homozygous truncations, did not present with SBS96 or an increased number of C → T mutations at CpG sites compared to *MBD4*-WT cancers. This data is consistent with what has previously been reported by Poulos *et al.,* where MSI^+ CRCs with somatic *MBD4* truncations had no difference in the number of C → T mutations at CpG sites compared to *MBD4*-WT cancers (453). This data indicates that only cancers with germline *MBD4* mutations, either biallelic or monoallelic with subsequent loss of the WT allele, present with the increased number of C → T mutations at CpG sites. This may be a consequence of the age-dependent mechanism by which these mutations are thought to accumulate. Patients with germline mutations in *MBD4* can be assumed to have a longer period of *MBD4* deficiency than cancers with somatic mutations in *MBD4*, representing a longer period of time over which unrepaired spontaneous deaminations of 5-mC can accumulate. However, the use of DNA sequencing read counts to determine if a somatic *MBD4* truncation was heterozygous or homozygous represents one of the limitations of this analysis. While other sources of DNA-sequencing data (e.g. TCGA) have accompanying RNA-sequencing data, this was not available for cancers from 100KGP data. Should suitable RNA-sequencing data have been available for these cancers, a more reliable assessment of these somatic *MBD4* mutations could have been performed.

Further analysis of these *MBD4*-mutant cancers and colorectal polyps identified pathogenic C → T mutations at CpG sites within the protein-coding sequences of cancer-specific driver genes. These genes included *PTEN* and *TP53,* which have been identified as drivers in several types of cancer (61,212,472,477,478). It is therefore plausible that the mechanism by which germline deficiencies in *MBD4* may drive tumorigenesis is via spontaneous deaminations of 5-mC, which may otherwise be repaired in *MBD4*-proficient individuals, propagating into pathogenic C → T mutations in cancer driver genes. As discussed in section 4.1.3, previous work by Sanders *et al.* characterised this scenario in *MBD4*-mutant AML – where pathogenic C → T mutations were identified at CpG sites of the coding sequence of *DNMT3A, IDH1* and *IDH2* (446). Similarly, the *MBD4*-mutant uveal melanoma investigated in this chapter presented with pathogenic C → T mutations at CpG sites within the coding sequences of the driver genes *GNA11* and *BAP1*. This confirms previous work by Derrien *et al.,* who identified six uveal melanoma patients with germline *MBD4* mutations (479). These *MBD4*-mutant cancers presented with a mutation signature similar to SBS96 identified by Degasperi *et al.,* while two-thirds of these cancers presented with pathogenic *GNA11* and *BAP1* mutations (479). In addition to this, the *MBD4*-mutant colorectal polyps investigated in this chapter presented with pathogenic C → T mutations at CpG sites of the *APC* protein-coding sequence. Tanakaya *et al.* also reported colorectal polyposis in a patient with germline *MBD4* deficiency, indicating a potential role for *MBD4* in driving this phenotype (447).

The study by Palles *et al.* suggested that patients with germline *MBD4* mutations were predisposed to the development of colorectal polyps, AML and other haematological malignancies (221). However, the data presented in this chapter suggest that these patients may also be predisposed to the development of breast cancer and sarcoma. Although an *in situ* breast ductal carcinoma was reported in an *MBD4*-mutant patient in the study by Palles *et*

al., the data presented in this chapter is, to the best of my knowledge, the first time *MBD4* deficiency has been implicated in disease predisposition for either of these cancer types. Palles *et al.* recommended the use of *MBD4* in diagnostic testing panels for colorectal polyposis, AML and uveal melanoma, in order to improve patient prognosis via earlier detection and screening of at risk individuals (221). The data presented in this chapter suggest that *MBD4* may also have the same clinical utility in breast cancer and sarcoma, potentially improving patient outcomes in the contexts of these diseases too. In addition to this, previous studies have suggested that *MBD4*-mutant cancers have improved responses to certain anti-cancer therapies (480). The study by Saint-Ghislain *et al.* suggests that *MBD4*-mutant metastatic uveal melanomas show better response to immune-checkpoint blockade therapy than *MBD4*-WT cancers (60% vs 4% respectively) (480). Furthermore, *MBD4*-mutant cancers showed improved overall survival and progression-free survival compared to their *MBD4*-WT counterparts (480). This suggests that patients with cancer predispositions driven by germline mutations in *MBD4* may benefit from a different course of anti-cancer therapies than *MBD4*-WT cancers, potentially providing an opportunity for the use of personalised medicine in these patients.

In the context of CRC, the data presented here and in previous studies suggest that germline mutations in *MBD4* predispose individuals to colorectal polyposis, however there is little data suggesting that it also predisposes individuals to the subsequent development of CRC. The study of *Mbd4*$^{-/-}$ mice by Wong *et al.* found no evidence of enhanced colorectal tumorigenesis compared to *Mbd4*$^{+/+}$ controls, despite a marked increase in the number of C → T mutations at CpG sites of the *Mbd4*$^{-/-}$ animals (450). From this data, it could be concluded that *MBD4* is not involved in driving CRC pathogenesis and that the colorectal polyps observed in *MBD4*-mutant patients may never develop into CRC. However, given the proposed indirect mechanism described above by which *MBD4* deficiency may predispose individuals to cancer, it is possible that the twenty-four month timescale used in the study by Wong *et al.* was not long enough for unrepaired spontaneous deaminations to accumulate in these *Mbd4*$^{-/-}$ animals. While twenty-four months is considered an extended lifespan for mice, in humans the timescale for unrepaired spontaneous deaminations to accumulate in *MBD4*-mutant patients is far longer – with the youngest *MBD4*-mutant patient in the study by Palles *et al.* presenting with colorectal polyps at eighteen years of age (221). Furthermore, when the *Mbd4*$^{-/-}$ mice in the study by Wong *et al.* were crossed onto an *Apc*-deficient background, intestinal tumorigenesis was enhanced in the *Mbd4*$^{-/-}$ animals compared to controls – with most animals presenting with pathogenic C → T mutations at CpG sites of the remaining *Apc* allele (450). This suggests that, on the appropriate background, *Mbd4* deficiency may drive CRC tumorigenesis in mice. As discussed in Chapter I of this thesis, the classical pathogenesis of CRC is thought to be driven by the sequential acquisition of pathogenic mutations in key CRC driver genes, including *APC, BRAF* and *TP53*. The *MBD4*-mutant colorectal polyps described in this chapter presented with a number of C → T mutations at CpG sites that resulted in *APC* truncations, which is not unexpected given the well-documented role of *APC* deficiency in the colorectal polyposis syndrome FAP (118). While the most common CRC-associated mutation in *BRAF* (*BRAF*^{V600E}) is driven by an A → T mutation in the context C**A**C, approximately half of pathogenic *TP53* truncations are C → T mutations at CpG sites, suggesting *MBD4*-mutant cancers may also develop these mutations (447). Overall, this it is possible that the *MBD4* deficiency of the affected individuals with colorectal polyps may subsequently drive the progression of these polyps into CRC.

Alternatively to mutations in the protein-coding sequence of cancer driver genes, an additional mechanism to modify the expression of these genes may also exist in patients with germline *MBD4* deficiency. As discussed in Chapter II of this thesis, aside from pathogenic mutations in the coding sequence of genes, there are other mechanisms by which an SNV may alter the expression of genes. SNVs outside the coding region of genes may act as eQTLs, which may either up-regulate or down-regulate the expression of a gene. While this chapter has focussed on C → T mutations at CpG sites within the coding sequence of driver genes, the majority of the C → T mutations at CpG sites in *MBD4*-mutant cancers lie outside the protein-coding regions of these genes. Therefore, it is plausible that these mutations may act to indirectly alter the expression of cancer-specific driver genes, either by altering transcription factor binding or the local chromatin conformation. It would be interesting to map these C → T mutations at CpG sites outside of coding regions to genomic features, similar to the *in silico* analysis performed in Chapter II. In the future this could be combined with RNA-sequencing data to potentially explain the altered gene expression of any driver genes without direct changes to the protein-coding sequence of that gene.

In addition to describing the role of germline *MBD4* mutations in cancer predisposition, this chapter also assessed a number of factors previously suggested to influence the likelihood of spontaneous deamination of an individual CpG site. These factors included the DNA methylation status, replication timing and transcription strand of the CpG site, with more highly-methylated CpG sites, CpG sites in late-replicating regions of the genome and CpG sites on the coding transcription strand thought to be the most at risk. When this was investigated using normal sigmoid colon fractional methylation data, a positive correlation was identified between DNA methylation and the C → T mutation rate at CpG sites. This correlation was identified in *MBD4*-mutant colorectal polyps and also in *MBD4*-WT CRCs and CRCs with germline or somatic *MBD4* truncations. While MSI^+ CRCs with somatic *MBD4* truncations showed no difference in the relationship between DNA methylation and the C → T mutation rate at CpG sites compared to *MBD4*-WT cancers, both the regression slope and the r^2 measure of this correlation was greater in the *MBD4*-mutant colorectal polyps than in MSS *MBD4*-WT CRCs, indicating that the correlation between DNA methylation and C → T mutagenesis at CpG sites was stronger in these polyps. From this data, it could be concluded that highly-methylated CpG sites are more likely to undergo spontaneous deamination, as the *MBD4*-mutant polyps are deficient in a key component of the repair pathway for these deaminations and present with a higher C → T mutation rate at highly-methylated CpG sites than *MBD4*-proficient cancers. The positive correlation between DNA methylation and C → T mutation rates at CpG sites has previously been reported in the studies by Poulos *et al.* and Fang *et al.,* who performed similar analyses using whole-genome sequencing data available from other sources (452,453). However, a limitation of the data presented in this chapter and the previous studies described above is the use of reference fractional methylation data from the normal sigmoid colon. As described in Chapter I of this thesis, alterations in DNA methylation patterns represents one of the key alterations that underpin the adenoma-carcinoma sequence of CRC pathogenesis (195). These changes can include hyper-methylation of regions of DNA that were previously hypo-methylated, especially at the promoter regions of key tumour suppressor genes (195). Therefore, the binning of C → T mutations based on the DNA methylation status of a CpG site in the normal sigmoid colon may not provide an accurate representation of the true DNA methylation status of that CpG site. Ideally, DNA methylation analysis would be performed

on each individual cancer or colorectal polyp, however this is often unpractical – while the use of methylation arrays would not provide the depth of information that the whole-genome bisulphite sequencing data used in this chapter was able to provide.

It was also identified in these colorectal polyps and cancers that the C → T mutation rate at CpG sites was highest in late-replicating genomic regions. Previous studies have suggested that the DNA MMR pathway is less active in late-replicating DNA, consequently increasing the mutation rate in these regions (454). In MSS CRCs and *MBD4*-mutant colorectal polyps, the association between DNA methylation and the C → T mutation rate at CpG sites was significantly stronger in late-replicating DNA compared to early-replicating DNA, whereas this difference was non-significant in MSI^+ CRCs. This indicates that the suggestions made in this previous study may be accurate, with the weaker association with replication timing in MSI^+ CRCs a consequence of the MMR deficiencies present in these cancers. This resembles data from the study by Poulos *et al.,* who also identified significant differences in the association between DNA methylation and mutation rates in late-replicating DNA in MSS CRCs, while not replicating this in MSI^+ cancers (453). However, the data presented in this chapter extends this analysis by utilising a significantly expanded sample size in addition to tissue-appropriate replication timing data with a finer-scale of resolution compared to the 1Mb windows from lymphoblastoid cell lines used in this previous study (453). Furthermore, the study of *MBD4*-mutant AML by Sanders *et al.* also identified an enrichment of C → T mutations at CpG sites in late-replicating DNA (446). There could be a number of mechanisms to explain this enrichment, the first being a simplistic model where spontaneous deaminations in late-replicating genomic regions have less time to be repaired before cell division, making them more likely to propagate into mutations. Alternatively, in line with previous data, the enrichment in late-replicating DNA may be linked to the MMR pathway. The study by Bellacosa *et al.* has previously suggested that *MBD4* is able to interact with the MMR protein *MLH1,* which has in turn been associated with altered expression of other MMR genes (481,482). Given this interaction between *MBD4* and the MMR pathway, it is possible that these proteins may act as a vehicle for the transport of *MBD4* to the sites of DNA mismatches produced by the spontaneous deamination of 5-mC. If so, it may be that in late-replicating DNA *MBD4* is less likely to be transported to the site of spontaneous deaminations as a consequence of reduced MMR activity. This could be tested in the future via chromatin immunoprecipitation and testing for co-localisation of *MBD4* and MMR proteins. If this was to be the case, it might be expected that the replication timing profile of the *MBD4*-mutant colorectal polyps may more closely resemble MSI^+ CRCs. However, it is possible that to some degree the loss of *MBD4* is compensated for by *TDG,* which performs a similar role in the BER pathway, in early-replicating regions of the genome. Currently, there is very little data investigating the potential association between *TDG* and MMR proteins, meaning this hypothesis would require further investigation.

Contrary to what has been previously reported regarding transcription strand bias of spontaneous deaminations, neither the CRCs nor *MBD4*-mutant colorectal polyps studied in this chapter presented with transcription strand bias. The same trend was also reported in *MBD4*-mutant AML by Sanders *et al.,* which implies that spontaneous deamination may not be affected by transcription strand as previously suggested (446). While transcription may not influence the likelihood of spontaneous deamination, there are other circumstances where single-stranded DNA may be exposed within cells, thus making the DNA more vulnerable to

deamination. One such example of this is during DNA replication, where the double-stranded DNA helix is unwound by DNA helicase to reveal single-stranded DNA to act as a template for the replicative DNA polymerases (483). If the process of DNA replication is indeed associated with enhanced risk of deamination, tissues with the highest rate of cellular turnover – and therefore the highest rate of DNA replication – may present with more deaminations than tissues with lower cellular turnover. As discussed in section 4.1.2, intestinal tissues have the highest rate of cellular turnover of any tissue, suggesting this tissue may be at the greatest risk of spontaneous deamination (439,484).

Overall, the data presented in this chapter suggests that DNA methylation and replication timing influence the likelihood of a CpG site undergoing spontaneous deamination. Identifying these factors and potentially others (e.g. those hypothesised above) may be of clinical relevance in patients with germline *MBD4* mutations. For example, it has been shown that highly-methylated CpG sites in late-replicating regions of the genome are at greater risk of deamination, therefore CpG sites within the protein-coding sequence of cancer-specific driver genes that fit these criteria may consequently be at greater risk of deamination. Therefore, it may be possible to predict from this which driver genes are likely to mutate in these patients, providing opportunities for surveillance and potentially clinical intervention in some circumstances.

In summary, the data presented in this chapter suggest that germline, but not somatic, truncations in *MBD4* predispose affected individuals to breast cancer, sarcoma, uveal melanoma and colorectal polyposis, with potential subsequent progression to CRC. The data in this chapter has identified a potential mechanism that underpins this predisposition and also suggest a number of factors that may affect the rate of spontaneous deamination of 5-mC – as well as suggesting additional factors that merit further investigation. Therefore, this chapter has expanded on previous studies to provide a comprehensive analysis of the consequences of germline *MBD4* truncations in a number of cancer types and how this information could be translated to clinically benefit affected individuals.

Chapter V – DNA Replication Errors as an Alternative Mechanism of C → T Mutagenesis at CpG Sites

5.1 – Background

5.1.1 – Other Mutation Signatures Characterised by C → T Mutations at CpG Sites

Chapter IV of this thesis discussed the mutation signature SBS1, a signature almost exclusively characterised by C → T mutations at CpG sites (425). SBS1 has previously been attributed to the failure to repair the spontaneous deamination of 5-mC to thymine, a repair process that is dependent on *MBD4* and/or *TDG* (425,441,450). As seen in Chapter IV of this thesis, a variety of cancers with germline mutations in *MBD4,* as well as colorectal polyps from a patient with a germline biallelic truncation in *MBD4,* present with increased C → T mutagenesis at CpG sites, thereby supporting the hypothesis that SBS1 is characterised by the accumulation of C → T mutations arising from unrepaired deaminations of 5-mC.

However, in addition to SBS1 (see Figure 5.1a), there are a number of other mutation signatures described in the COSMIC database that are at least partially characterised by C → T mutations at CpG sites. For example, the mutation signature SBS10b, which is associated with defective DNA "proofreading" – caused by mutations in the exonuclease domain of DNA *POL-ε* (485). As presented in Figure 5.1b, SBS10b is primarily characterised by C → T mutations in the context T**C**G, with nearly half of all SNVs attributed to this trinucleotide context. Mutation signatures SBS6 (Figure 5.1c) and SBS15 (Figure 5.1d) also present with a high number of C → T mutations at CpG sites. Both of these mutation signatures are associated with defective DNA MMR with particular enrichment of C → T mutations in the context G**C**G. The presence of other mutation signatures characterised by C → T mutations at CpG sites implies there may also be an alternative mechanism to unrepaired spontaneous deamination of 5-mC by which C → T mutagenesis occurs at CpG sites.

5.1.2 – DNA Polymerase Mutations in Cancer

5.1.2.1 – DNA Polymerases δ & ε

The accurate replication of the three-billion base-pairs of the human genome is essential for the maintenance of genomic stability and the prevention of potentially tumour-initiating mutations. This is, in part, achieved by the high-fidelity of cellular replicative polymerases working in concert with active DNA "proofreading" and other DNA repair mechanisms (486,487). In addition to the MMR pathway, the role of which in maintaining genomic integrity will be described in section 5.1.3, DNA polymerase (POL) proteins themselves also play a key role in preventing DNA replication errors propagating into potentially pathogenic mutations (452,486,488). In eukaryotic cells, there are fifteen known POL enzymes – however, most DNA replication requires only three (489–491). These are the DNA primase *POL-α* and the bulk DNA polymerases *POL-ε* and *POL-δ* – which are responsible for synthesising the leading and lagging strand respectively during DNA replication (489).

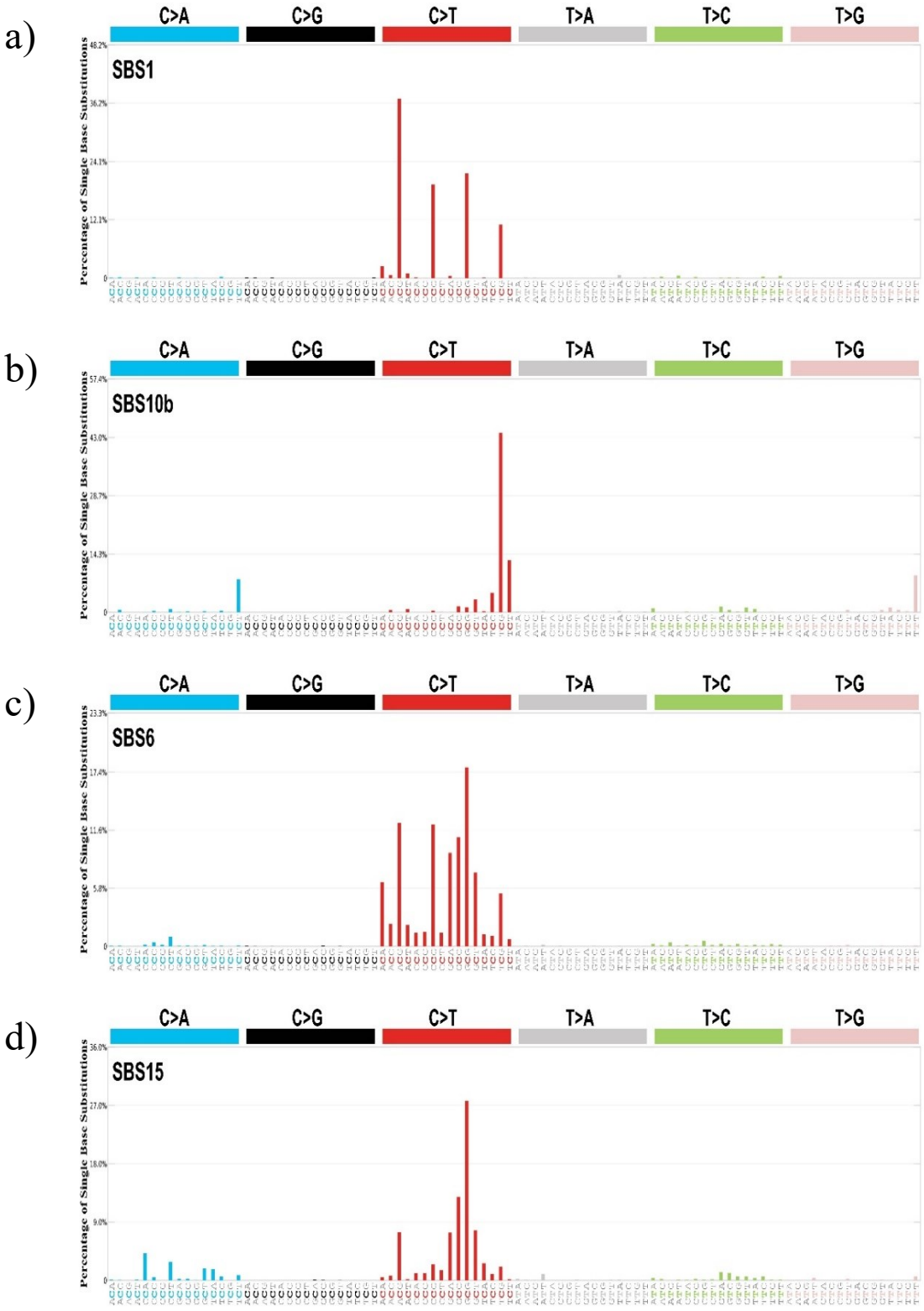


Figure 5.1 – Other Mutation Signatures Characterised by C → T Mutations at CpG Sites: Other mutation signatures from the COSMIC database that present with an enrichment of C → T mutations at CpG sites. Presented in each plot are the enrichments of each of the ninety-six potential trinucleotide context of a single-base substitution mutation. Signatures presented include SBS1 (a), SBS10b (b), SBS6(c) and SBS15 (d). All signature plots obtained from the COSMIC database (https://cancer.sanger.ac.uk/signatures/sbs/).

In mammals, the *POL-δ* enzyme is comprised of four sub-units – each encoded by a different gene (489). The catalytic *p125* sub-unit is encoded by the *POLD1* gene and the *p50, p68* and *p12* sub-units are encoded by the *POLD2, POLD3* and *POLD4* genes respectively (489). Similarly, the *POL-ε* protein is also made up of four sub-units, with the 225kDa catalytic sub-unit *p261* encoded by the *POLE* gene, whereas the additional sub-units *p59, p12* and *p17* are the products of the *POLE2, POLE3* and *POLE4* genes respectively (489).

5.1.2.2 – DNA Polymerase Exonuclease Activity

In mammals, *POL-δ, POL-ε* and *POL-γ* possess intrinsic 3' → 5' exonuclease activity, due to their critical roles in the synthesis of new DNA strands during replication (492). This exonuclease activity allows these polymerases to "proofread" the newly-synthesised DNA strand and correct any errors that may arise during the process of DNA replication (492). The activity of this exonuclease domain is critical to maintain the low intrinsic replication error rates of *POL-δ* and *POL-ε*, which have error rates of 1 in 10^5 – 10^6 base-pairs and 1 in 10^6 – 10^7 base-pairs respectively (493). The importance of this DNA "proofreading" is highlighted by the sixty-fold increase in mutation rates in cells lacking a functional *POL-δ* exonuclease domain (494).

The exonuclease domains of these DNA polymerases are located at the N-terminus of the protein, where conserved amino acid residues play a critical role in both DNA binding and catalysing the removal of nucleotides from the newly-synthesised DNA strand at the site of a base-pair mismatch arising from a replication error – allowing the error to be corrected (495). The process of "proofreading" the newly-synthesised strand during DNA replication is summarised in Figure 5.2.

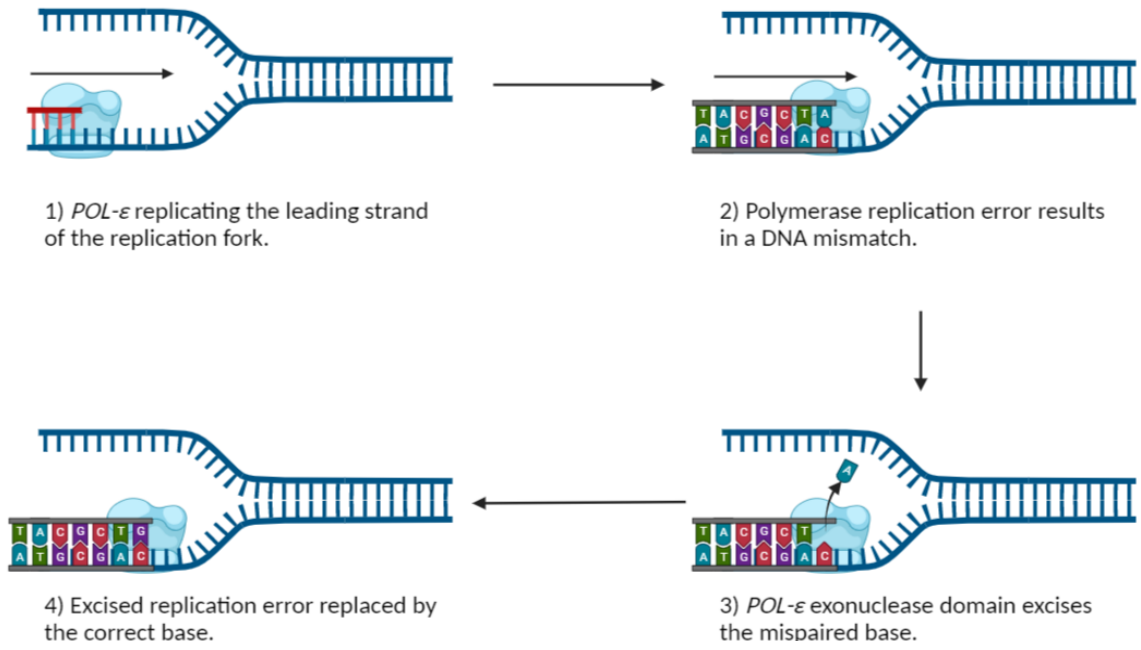


Figure 5.2 – The Mechanism of Polymerase-Mediated "Proofreading" During DNA Replication: A diagrammatic illustration by which DNA polymerases repair replication errors via intrinsic exonuclease domain "proofreading" activity. Firstly, DNA polymerase epsilon (*POL-ε*), synthesising the leading strand during DNA replication (1) may make an error during replication – resulting in a DNA mismatch (2). The mispaired base is then excised by the exonuclease domain of *POL-ε* (3) and is replaced by the correct base (4). Created with BioRender.com (https://app.biorender.com/).

5.1.2.3 – DNA Polymerase Mutations in Colorectal Cancer

Mutations in the exonuclease domain of *POL-δ* and *POL-ε* have been associated with the development of several types of cancer, including endometrial cancer and CRC (156,496). Whole-exome sequencing of CRCs from TCGA revealed for the first time hyper-mutated cancers with somatic *POL-ε* EDMs (69). Subsequently, Palles *et al.* identified germline mutations in the exonuclease domain of *POL-δ* and *POL-ε* that predisposed affected individuals to the development of colorectal adenomas (156). The adenomas presented by these carriers of $POL\text{-}\delta^{S478N}$ or $POL\text{-}\varepsilon^{L424V}$ showed chromosomal instability as well as pathogenic mutations in a number of known CRC driver genes, including *APC, BRAF* and *KRAS* (156). Interestingly, these adenomas showed no signs of MSI, indicating that MMR deficiencies were not the cause of these driver gene mutations (156). When these equivalent mutations were studied in yeast ($POL\text{-}\delta^{L479S}$ and $POL\text{-}\varepsilon^{C462S}$), the exonuclease activity of these polymerases was abolished – resulting in a greatly increased mutation rate compared to POL-WT yeast (156).

Further studies have associated cancers with pathogenic *POL-ε* exonuclease domain mutations (EDMs) with an "ultra-mutagenic" phenotype on an otherwise MSS background (452,497,498). These cancers present with an exceptionally high mutation burden of more than 100 mutations per Mb, primarily attributed to unrepaired DNA replication errors as a consequence of abolished DNA "proofreading" mechanisms (452,497,498). Following the previous study by Palles *et al.,* subsequent studies have identified a number of additional pathogenic *POL-ε* EDMs associated with this ultra-mutagenic phenotype, including $POL\text{-}\varepsilon^{P286R}$, $POL\text{-}\varepsilon^{V411L}$ and $POL\text{-}\varepsilon^{S459F}$ – which have been identified in endometrial cancer and CRC (452,499,500). In total, *POL-ε* EDMs are present in approximately 7% of all endometrial cancers and approximately 3% of CRCs (501,502). Despite the associated hyper-mutated phenotype described above, mutations in the *POL-ε* exonuclease domain may be associated with more favourable patient outcomes in endometrial cancer (501). In the study by Church *et al.,* patients harbouring *POL-ε* EDMs were associated with fewer instances of disease recurrence (6.2% vs 14.1%) and cancer-related death (2.3% vs 9.7%) compared to their *POL-ε* WT counterparts (501). The study by Li *et al.* identified a significant correlation between tumour mutation burden and survival in endometrial cancer, as well as a significant correlation between *POL-ε* EDMs and tumour mutation burden, suggesting that cancers with *POL-ε* EDMs may be associated with improved survival (503).

In the context of CRC, the study by Domingo *et al.* identified sixty-six cancers with pathogenic *POL-ε* EDMs and found that they were significantly younger (median age of 54.5 years vs 67.2 years) and more common in males (75.8% vs 55.5%) than MSS *POL-ε* WT CRCs (504). In addition to this, CRCs with *POL-ε* EDMs presented with increased tumour infiltration by $CD8^+$ lymphocytes and increased cytotoxic T-cell markers than MSS *POL-ε*

WT CRCs (504). This, coupled with the significantly reduced risk of disease recurrence of *POL-ε* mutant CRCs compared to MSS *POL-ε* WT cancers, suggests that *POL-ε* EDMs may be associated with a more favourable prognosis in CRC (504). In the recent CRC landscape study, mutation signatures associated with defective *POL-ε* "proofreading" were present in 1.84% (SBS10a) and 0.84% (SBS10b) of CRCs respectively. When combined with the previous work by Palles *et al.,* it could be concluded that *POL-ε* EDMs represent a driver mutation in a subset of CRCs.

Overall, pathogenic mutations in the exonuclease domain of *POL-ε* are associated with a hyper-mutated phenotype in CRC and endometrial cancer. Interestingly, these EDMs are also associated with the mutation signature SBS10b, which is predominantly characterised by C → T mutations in the context T**C**G, suggesting that unrepaired DNA replication errors associated with defective "proofreading" may represent an alternative mechanism to unrepaired spontaneous deaminations of 5-mC for C → T mutagenesis at CpG sites.

5.1.3 – Defective DNA Mismatch Repair in Colorectal Cancer

5.1.3.1 – DNA Mismatch Repair

In addition to polymerase-mediated "proofreading" during DNA replication, the DNA MMR pathway represents another mechanism by which cells are able to maintain genomic stability. Like DNA "proofreading", the MMR pathway is able to correct DNA mismatches which are erroneously produced during DNA replication, preventing their propagation into potentially tumour-initiating mutations (505). Overall, the removal of DNA mismatches associated with replication errors by either polymerase-mediated "proofreading" or the MMR pathway is estimated to improve the fidelity of DNA replication 100-fold (494). The MMR pathway involves a number of proteins, including heterodimers between the eight eukaryotic MutS homologue (MSH) proteins, referring to MutS-directed MMR that has been previously identified in bacteria (506). Heterodimers between *MSH2* and *MSH6* form the MutS-α complex, which is involved in repair of single-base DNA mismatches (507). In addition to this, heterodimers between *MSH2* and *MSH3* form the MutS-β complex, which work in tandem with MutS-α to repair smaller insertion-deletion loops (IDLs) (507). The MutS-β complex is also involved in the repair of larger IDLs between two and eight base-pairs in size (507). *MSH1* is thought to be involved in the MMR pathway for mitochondrial DNA, while the *MSH4* and *MSH5* proteins are related to meiosis and not the repair of DNA mismatches produced by DNA replication errors (507).

In addition to these MutS heterodimers, another important component of the eukaryotic MMR pathway are MutL homologue (MLH) proteins, named after similar proteins first identified in *E. coli* (508). The *MLH1* protein forms heterodimers with *PMS2, PMS1* and *MLH3* to form the MutL-α, MutL-β and MutL-γ complexes respectively, with the MutL-α complex being the most prominent of these three (508). These MutL complexes have critical endonuclease activity that form part of the DNA MMR pathway (127,509,510). Upon the identification of a DNA mismatch by the MutS complex, MutL complexes are recruited to the site of the mismatch and generate a nick in the DNA either 5' or 3' of the mismatch

(509,510). If this nick is 5' of the mismatch, MutS-α, replication protein A and the DNA exonuclease 1 (*EXO1*) are required for MMR, whereas if the nick is 3' of the mismatch, proliferating cell nuclear antigen and replication factor C are also required for repair (511–513). Following the generation of this DNA nick, DNA is excised by *EXO1* past the point of the DNA mismatch and re-synthesised by *POL-δ* (514). A summary of the DNA MMR pathway is provided in Figure 5.3.

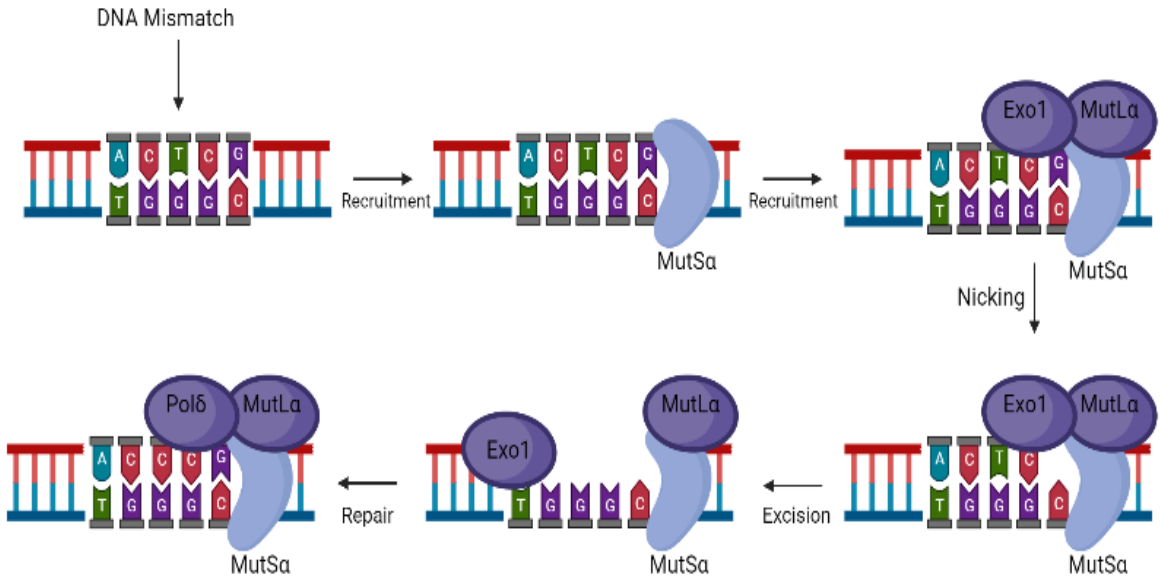


Figure 5.3 – An Overview of the Mismatch Repair Pathway: A diagrammatic illustration of the human mismatch repair pathway. Starting with mispaired bases (T:G mismatch), mismatch repair complexes MutS-α and MutL-α are recruited alongside DNA exonuclease 1 (*EXO1*). MutL-α oversees the nicking of a single base near the mismatch allowing *EXO1* to excise the DNA past the point of the mismatch before repair is completed by DNA polymerase δ. Created with BioRender.com (https://app.biorender.com/).

One of the major challenges facing the MMR pathway is ensuring the correct DNA strand is repaired. Given that DNA mismatches are often the result of DNA replication errors, repair proteins should be directed to the newly-synthesised DNA strand and not the parental strand (507). During DNA replication, this newly-synthesised daughter strand is unmethylated, allowing the MutH complex in prokaryotes to discriminate this strand from the parental strand by nicking at hemi-methylated sites around the DNA mismatch (515). However, in eukaryotes this DNA nicking is performed by MutL (see Figure 5.3). In addition to this, there are no MutH proteins in eukaryotes, meaning our understanding of the mechanisms underlying strand discrimination remains incomplete (515).

5.1.3.2 – DNA MMR Pathway Deficiency in Colorectal Cancer

As described in Chapter I of this thesis, pathogenic mutations in genes involved in the MMR pathway are a driver of MSI in several cancer types, including CRC, gastric cancer and

endometrial cancer (516). Approximately 15% of all CRCs are classified as MSI^+ and present with a higher somatic mutation burden than MSS cancers – further exemplifying the importance of the MMR pathway in ensuring the fidelity of DNA replication (486,507). When studied in yeast, mutations in components of the MMR pathway increased the rate of frameshift mutations at small microsatellite repeat regions of the genome, which are typically between one and six base-pairs in length and account for approximately 3% of the genome (517). The clinical utility of the MSI status of a cancer is described in Chapter I of this thesis, as is the CRC predisposition syndrome Lynch Syndrome (LS) – which is a consequence of germline pathogenic mutations in components of the MMR pathway.

In a recent CRC landscape paper, whole-genome sequencing analysis revealed a number of mutation signatures associated with MMR deficiency were present in CRC, including SBS15, SBS26 and SBS44 (295). As seen in Figure 5.1d, the mutation signature SBS15 is characterised by a large number of C → T mutations at CpG sites and was identified in 6.77% of all CRCs by the recent study (295). This suggests that, in addition to unrepaired spontaneous deaminations of 5-mC (SBS1) and unrepaired replication errors caused by defective DNA "proofreading" (SBS10b), defective DNA MMR also has a role in C → T mutagenesis at CpG sites. Interestingly, previous studies have suggested that both unrepaired spontaneous deaminations of 5-mC and unrepaired *POL-ε* replication errors contribute to C → T mutagenesis at CpG sites in MSI^+ cancers (518,519). The study by Fang *et al.* suggested that deficiencies in the MutS complex are associated with an increase in unrepaired spontaneous deaminations, whereas mutations in components of the MutL complex are associated with the propagation of unrepaired DNA replication errors (518). In addition to this, the study suggests that cancers with MutS deficiencies present with a greater number of C → T mutations at CpG sites and a greater representation of SBS1 than cancers with mutations in MutL proteins – thereby suggesting that MutS proteins, like *MBD4,* are involved in the repair of spontaneous deaminations (518). This increase was also seen in the study by Sanders *et al.,* who studied patients with constitutive deficiencies in MMR proteins (519).

5.1.4 – Chapter Aims

The identification of mutation signatures in CRC with an enrichment of C → T mutations at CpG sites other than SBS1 further emphasises the importance of these mutations in disease pathogenesis. The specific aetiologies of these additional signatures, including SBS10b and SBS15, also suggest that there may be an alternative mechanism driving C → T mutagenesis at CpG sites in addition to unrepaired spontaneous deaminations of 5-mC. These mutation signatures are attributed to defective *POL-ε* "proofreading" and MMR respectively, indicating that unrepaired replication errors may represent the underlying mechanism behind the enhanced C → T mutagenesis at CpG sites in these cancers. However, previous studies have suggested that MSI^+ cancers may accumulate these mutations via a combination of both unrepaired spontaneous deaminations and replication errors (see section 5.1.3.2). Therefore, this chapter will explore the extent to which DNA replication errors drive C → T mutagenesis at CpG sites in CRC, with particular emphasis placed on MSI^+ cancers. Similarly to work performed in Chapter IV of this thesis, this chapter will also assess if there are any other factors that influence the likelihood of a DNA replication error occurring at a CpG site,

including DNA methylation and replication timing. In order to achieve this, this chapter aims to:

- Assess the number of C → T mutations at CpG sites and mutation signatures of CRCs with *POL-ε* EDMs and/or MSI in comparison to MSS, POL-WT cancers.
- Compare the number of C → T mutations at CpG sites and mutation signatures of MSI^+ CRCs with mutations in MutS or MutL complex proteins.
- Assess the replication strand bias of these mutations in these cancers using previously published replication strand data (520) and use this to determine the likely mechanisms behind replication error-induced C → T mutagenesis at CpG sites.
- Assess the relationship between DNA methylation / replication timing and the C → T mutation rate at CpG sites in these cancers.

The above aims are accompanied by the following hypotheses:

- MSI^+ CRCs and CRCs with pathogenic *POL-ε* EDMs will present with a higher C → T mutation burden at CpG sites than MSS POL-WT CRCs.
- CRCs with *POL-ε* EDMs will show a strong bias of C → T mutations at CpG sites on the leading strand template, indicating that these mutations arise from unrepaired DNA replication errors.
- MSI^+ CRCs with MutS deficiencies will present with a higher number of C → T mutations at CpG sites than MSI^+ cancers with MutL deficiencies.
- MSI^+ CRCs with MutL deficiencies will show a bias of C → T mutations at CpG sites of the leading strand template, whereas MSI^+ CRCs with MutS deficiencies will show no strand bias, suggesting that these mutations are a consequence of unrepaired spontaneous deaminations of 5-mC.

5.2 – Materials & Methods

5.2.1 – 100,000 Genomes Project Data

Similarly to the analysis in Chapter IV of this thesis, primary, treatment-naïve CRCs with a PCR-free library preparation were identified from the CRC domain of the 100KGP (see section 4.2.1 for a detailed description of these cancers). Of these 1,907 cancers, 359 were categorised as MSI^+ via mSINGS analysis and eighteen presented with somatic mutations in the exonuclease domain of either *POL-δ* (n = 1) or *POL-ε* (n = 17). A complete list of these POL-mutant cancers is presented in Table 5.1. There were no cancers with germline mutations in the exonuclease domain of either *POL-δ* or *POL-ε.* SNVs were extracted from the somatic VCFs associated with each individual cancer and included in downstream analysis only if they presented with the "PASS" flag in the filter field.

Cancer #:	Chromosome #:	Position (hg38):	MSI Status:	Reference Allele:	Alternate Allele:	Mutation:	Previously Reported?
1	12	132,673,261	MSS	G	A	*POL-ε*^{S459F}	Yes (521)
2	12	132,673,261	MSS	G	A	*POL-ε*^{S459F}	Yes (521)
3	12	132,673,271	MSS	C	G	*POL-ε*^{A456P}	Yes (521)
4	12	132,673,703	MSS	C	G	*POL-ε*^{V411L}	Yes (521)
5	12	132,673,703	MSS	C	A	*POL-ε*^{V411L}	Yes (521)
6	12	132,673,703	MSS	C	A	*POL-ε*^{V411L}	Yes (521)
7	12	132,673,703	MSS	C	G	*POL-ε*^{V411L}	Yes (521)
8	12	132,673,703	MSS	C	A	*POL-ε*^{V411L}	Yes (521)
9	12	132,676,598	MSI^{+}	G	A	*POL-ε*^{P286L}	Yes (522)
10	12	132,676,598	MSS	G	C	*POL-ε*^{P286R}	Yes (499)
11	12	132,676,598	MSS	G	C	*POL-ε*^{P286R}	Yes (499)
12	12	132,676,598	MSS	G	C	*POL-ε*^{P286R}	Yes (499)
13	12	132,676,598	MSS	G	C	*POL-ε*^{P286R}	Yes (499)
14	12	132,676,598	MSS	G	C	*POL-ε*^{P286R}	Yes (499)
15	12	132,676,598	MSS	G	C	*POL-ε*^{P286R}	Yes (499)
16	12	132,676,598	MSS	G	C	*POL-ε*^{P286R}	Yes (499)
17	12	132,676,598	MSS	G	C	*POL-ε*^{P286R}	Yes (499)
18	19	50,402,733	MSI^{+}	G	A	*POL-δ*^{G321D}	No

Table 5.1 – Pathogenic DNA Polymerase Exonuclease Domain Mutations in Colorectal Cancer: A list of primary, treatment-naïve colorectal cancers with pathogenic mutations in the exonuclease domain of DNA polymerase-δ (*POL-δ*) or DNA polymerase-ε (*POL-ε*) identified from the 100,000 Genomes Project. Presented are the chromosome number and co-ordinates (hg38) of the exonuclease domain mutation, the MSI status of the cancer, the reference and alternate alleles associated with the mutation, the associated amino acid change and whether this mutation has been previously reported in human cancer (with reference number in brackets)..

5.2.2 – MutS & MutL Classification

Following the identification of the 359 MSI^+ CRCs described above, the 357 MSI^+ POL-WT CRCs were then classified as either MutS-deficient or MutL-deficient. A list of CRCs with truncations or pathogenic missense variants in the MMR proteins *MLH1, PMS2, MSH2* and *MSH6* was provided by Güler Gül (University of Edinburgh). MSI^+ cancers with germline pathogenic mutations in *MSH2* or *MSH6* could be classified as LS patients with MutS deficiency and individuals with germline *MLH1* mutations were categorised as LS patients with MutL deficiency. Furthermore, cancers with likely somatic biallelic inactivation of *MSH2* or *MSH6,* as determined by χ^2 analysis, were categorised as cancers with somatic MutS deficiency. Similarly, cancers with likely somatic biallelic inactivation of *MLH1* were categorised as cancers with somatic MutL deficiency. In the instances where an otherwise MSI^+ cancer presented with no mutations in any of the above MMR genes, it was assumed that the MSI^+ phenotype of the cancer was a consequence of *MLH1* promoter hyper-methylation, a common mechanism of *MLH1* inactivation (see Chapter I of this thesis).

5.2.3 – Mutation Spectrum & Signature Extraction

The trinucleotide context of each SNV and the mutation signatures of the 1,907 CRCs described in section 5.2.1 were extracted in the same way as described in Chapter IV of this thesis. Briefly, only somatic SNVs with the "PASS" flag in the filter field of the VCF were included in mutation signature extraction. SigProfilerExtractor was used to extract the trinucleotide context of each input SNV and also to extract the pre-defined COSMIC mutation signature(s) (COSMIC V3.2) present in these cancers (383).

5.2.4 – DNA Methylation, Replication Timing, Bivalent Promoter & Replication Strand Mutation Mapping

Similarly to Chapter IV of this thesis, fractional DNA methylation data from the normal sigmoid colon and replication timing data from the CRC cell line HCT116 were used to bin C → T mutations at CpG sites according to the DNA methylation and replication timing bins described previously (see section 4.2.5 and Table 4.1). Methylation and replication timing binning was performed as described in section 4.2.5, with the same twelve DNA methylation bins and four replication timing bins being used for the 27,099,859 CpG sites. In addition to this, a total of 855,851 CpG sites mapped to the bivalent promoter regions described in Chapter III of this thesis (371). The number of CpG sites in each DNA methylation bin is presented in Table 5.2. C → T mutations at CpG sites could then be binned to these bivalent promoter regions using BEDTools (v2.30).

Previously published replication strand data from the study by Haradhvala *et al.* was provided by Dr Marketa Tomkova (University of Oxford) (520). This data uses the replication timing profiles of six lymphoblastoid cell lines described by Koren *et al.* – comprising of mother-father-offspring trios of either European or West-African heritage (523). From this, 20,000 base-pair regions of the genome could be characterised as either left-replicating or right-

replicating, indicating the direction of travel of the replication fork from the origin of replication (520). Replication strand co-ordinates were lifted to genome build hg38 using the UCSC LiftOver tool (372). This replication direction data could then be combined with the fractional DNA methylation data to give a replication direction estimate for 9,638,264 CpG sites. The number of CpG sites in each methylation bin categorised as left-replicating or right-replicating is presented in Table 5.3. Following this, C → T mutations at CpG sites could be assigned to either left-replicating or right-replicating DNA using BEDTools (v2.30).

From this data, the ratio of C → T to G → A mutations could be calculated in both left-replicating and right-replicating DNA in order to identify the mechanism by which C → T mutations at CpG sites might be produced. This data could then be combined with fractional methylation data to produce an association between DNA methylation and replication strand, with the number of C → T mutations at CpG sites normalised against the number of CpG sites in each bin (see Table 5.3) in a similar method to what is described in section 4.2.5. In addition to this, the number of C → T mutations at CpG sites on each replication strand was used to calculate a $\log_2$(Leading/Lagging) ratio of mutations for each cancer.

5.2.5 – Statistical Analysis

Similarly to Chapter IV of this thesis, comparisons between the total SNV burdens, the number of C → T mutations at CpG sites, ages and mutation signature compositions of each group of cancers was performed using a Wilcoxon test via the R package *ggpubr* (365). In the instances where there were multiple groups being compared, a Kruskal-Wallis test was used in addition to pairwise Wilcoxon tests between groups, again via the R package *ggpubr.* In addition to this, comparisons between the regression equations for the correlation of DNA methylation and C → T mutation rate at CpG sites were performed using interaction terms between regression constants. To compare the number of mutations on the template leading and lagging strands in each cancer, a non-parametric paired Wilcoxon rank test was performed by the *PairedData* package (370).

5.3 – Results

5.3.1 – Characteristics of MSI$^+$ CRCs and CRCs with *POL-ε* EDMs

CRCs with pathogenic EDMs in DNA *POL-δ* or *POL-ε* are known to present with a hyper-mutated phenotype, characterised by potentially millions of SNVs within the genomes of tumour cells (156,159,499). Similarly, MSI$^+$ CRCs also present with an elevated number of SNVs compared to their MSS counterparts (295). In the recent CRC landscape study, CRCs with *POL-ε* EDMs presented with more than 100 mutations per Mb and MSI$^+$ CRCs on average presented with more than 50 – while MSS POL-WT CRCs presented on average with

# CpG Sites:	0	0.01 – 0.1	0.11 – 0.2	0.21 – 0.3	0.31 – 0.4	0.41 – 0.5	0.51 – 0.6	0.61 – 0.7	0.71 – 0.8	0.81 – 0.9	0.91 – 0.99	1	Total:
Bivalent Promoters	242,408	366,518	73,792	37,783	27,998	23,027	18,614	15,809	15,151	17,310	16,925	516	855,851

Table 5.2 – CpG Site DNA Methylation & Bivalent Promoter Distribution: The number of CpG sites at bivalent promoter regions found within each DNA methylation bin, according to fractional methylation data of the normal sigmoid colon and bivalent promoter data obtained from Court & Arnaud. Each column represents one of the twelve possible DNA methylation bins of the genome (ranging from 0 – 1).

# CpG Sites:	0	0.01 – 0.1	0.11 – 0.2	0.21 – 0.3	0.31 – 0.4	0.41 – 0.5	0.51 – 0.6	0.61 – 0.7	0.71 – 0.8	0.81 – 0.9	0.91 – 0.99	1	Total:
Left-Replicating	154,118	172,984	55,190	68,191	94,876	140,584	170,437	247,611	433,266	1,004,062	2,219,461	114,466	4,875,156
Right-Replicating	142,573	158,942	52,851	65,573	92,009	137,319	167,269	243,752	426,276	983,212	2,180,055	113,277	4,763,108
Total:	296,691	331,926	108,041	133,764	186,795	277,903	337,706	491,363	859,542	1,987,274	4,399,516	227,743	9,638,264

Table 5.3 – CpG Site DNA Methylation & Replication Strand Distribution: The number of CpG sites found within each DNA methylation bin and replication strand, according to fractional methylation data of the normal sigmoid colon and replication strand data obtained from Haradhvala *et al*. Each column represents one of the twelve possible DNA methylation bins of the genome (ranging from 0 – 1) and each row represents whether a CpG site is located within a Left-Replicating or Right-Replicating region of the genome.

less than 10 mutations per Mb (295). Therefore, it could be inferred that these cancers may also present with a greater number of C → T mutations at CpG sites than their MSS POL-WT counterparts. In order to investigate this, the mutation spectrum of 1,907 primary, treatment-naïve CRCs was extracted from the 100KGP dataset. Of these cancers, 1,548 were MSS and 359 were MSI^+, while a total of 18 presented with pathogenic mutations in the exonuclease domain of either *POL-δ* or *POL-ε*. A detailed breakdown of these cancers is presented in Figure 5.4.

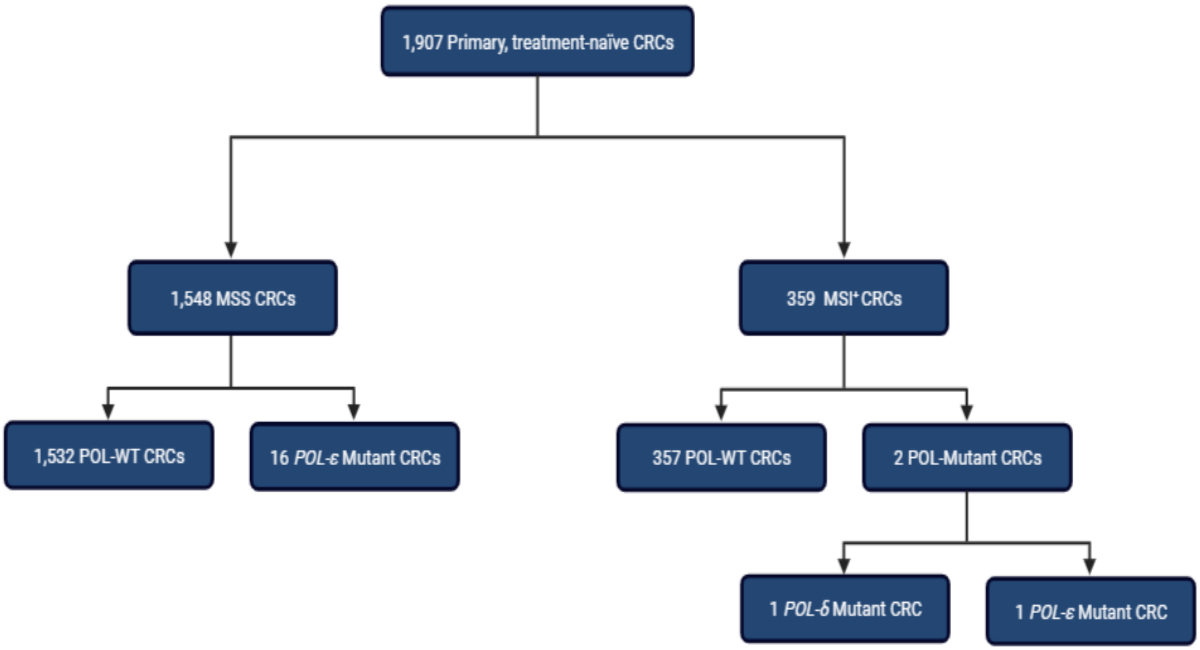
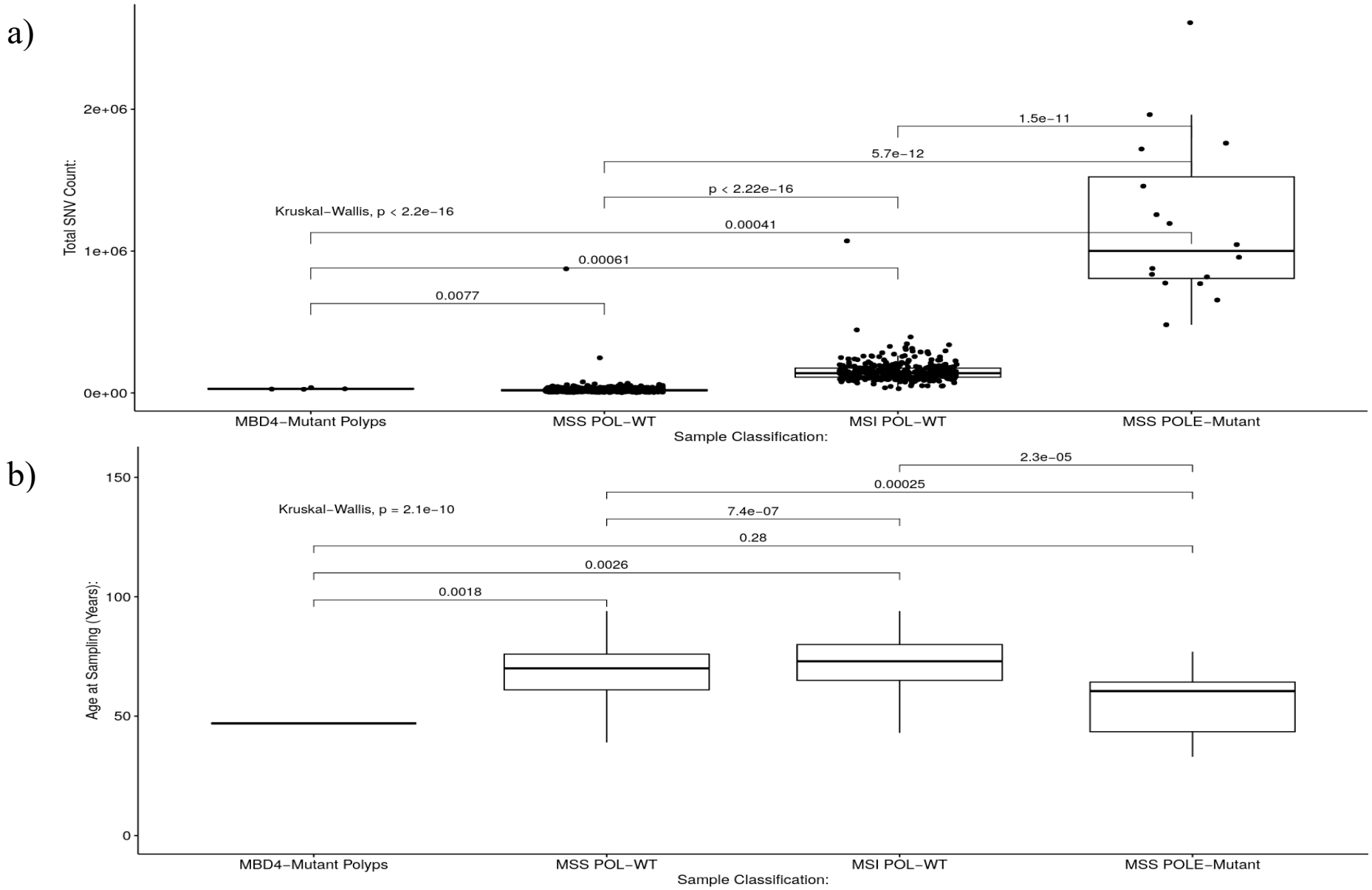


Figure 5.4 – Colorectal Cancers with Microsatellite Instability & DNA Polymerase Exonuclease Domain Mutations: A flowchart summarising the microsatellite instability (MSI) and DNA polymerase (POL) mutation statuses of the 1,907 primary, treatment-naïve colorectal cancers (CRCs) identified in V14 of the 100,000 Genomes Project Data. Created with BioRender.com (https://app.biorender.com/).

As seen in Figure 5.4, two MSI^+ CRCs presented with POL EDMs, one cancer with a *POL-*δ^{G321D} mutation and one cancer with a *POL-*ε^{P286L} mutation (see Table 5.1 for details). This *POL-δ* mutant MSI^+ cancer did not present with any of the mutation signatures indicative of *POL-δ* "proofreading" deficiency (data not shown), so was excluded from further analyses. In addition to this, while the *POL-ε* mutant MSI^+ cancer presented with the expected mutation signatures (data not shown), the *POL-*ε^{P286L} mutation was not observed in any MSS cancer (see Table 5.1), making comparisons between MSS and MSI^+ cancers with *POL-ε* EDMs difficult. Therefore, this cancer was also excluded from downstream analyses.

As seen in Figure 5.5a, the total SNV burdens of MSI^+ POL-WT CRCs and MSS CRCs with *POL-ε* EDMs was significantly greater than that of MSS POL-WT cancers ($p < 2.2 \times 10^{-16}$ and $p = 5.7 \times 10^{-12}$ respectively) and the *MBD4*-mutant colorectal polyps described in Chapter IV of this thesis ($p = 0.00061$ and $p = 0.00041$ respectively). Furthermore, the SNV burden of the MSS *POL-ε* mutant CRCs was significantly greater than the SNV burden of


a)
Total SNV Count:
2e+06
1e+06
0e+00
Kruskal–Wallis, p < 2.2e–16
1.5e–11
5.7e–12
p < 2.22e–16
0.00041
0.00061
0.0077
MBD4–Mutant Polyps
MSS POL–WT
MSI POL–WT
MSS POLE–Mutant
Sample Classification:
b)
Age at Sampling (Years):
150
100
50
0
Kruskal–Wallis, p = 2.1e–10
2.3e–05
0.00025
7.4e–07
0.28
0.0026
0.0018
MBD4–Mutant Polyps
MSS POL–WT
MSI POL–WT
MSS POLE–Mutant
Sample Classification:

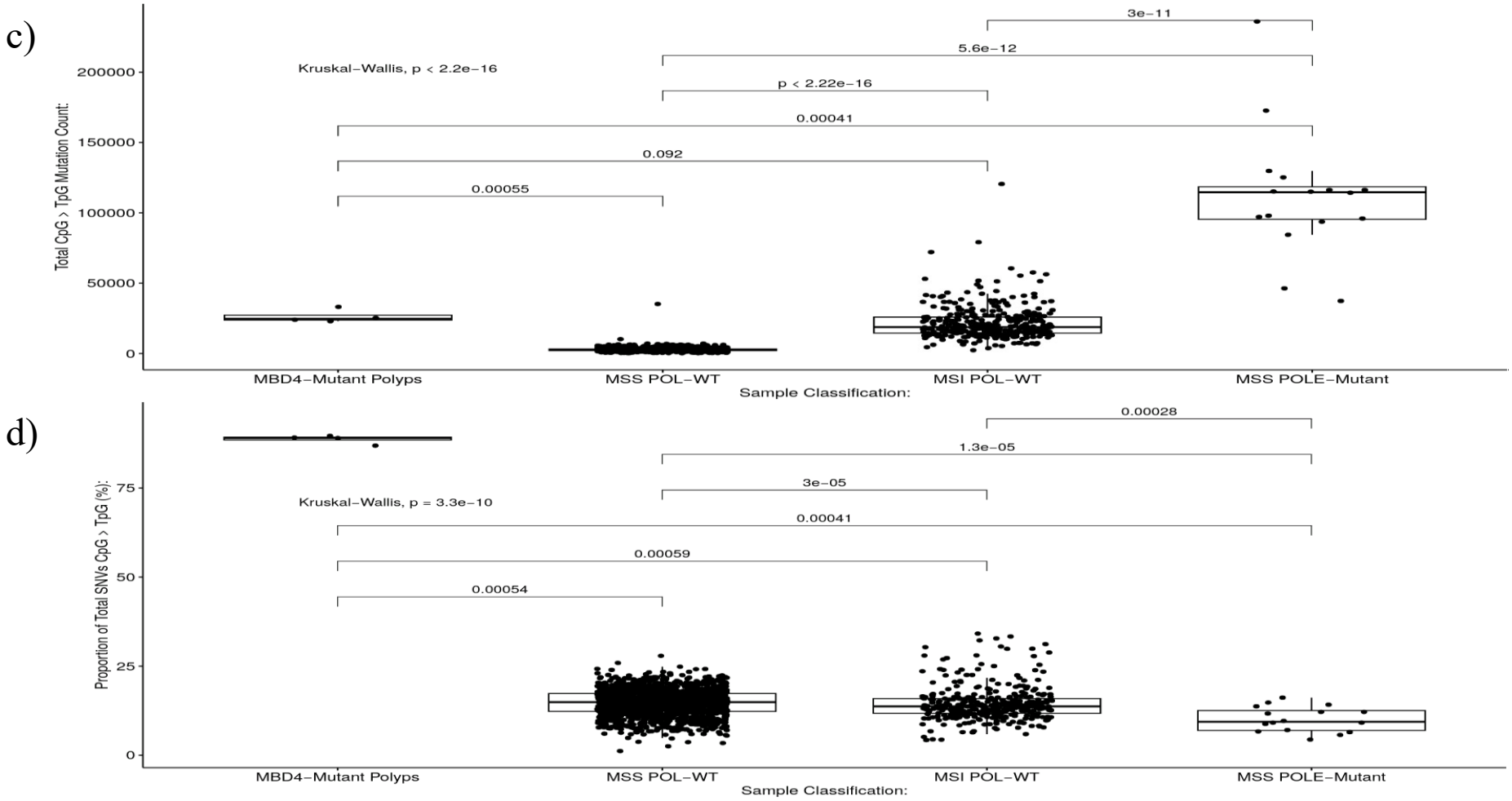


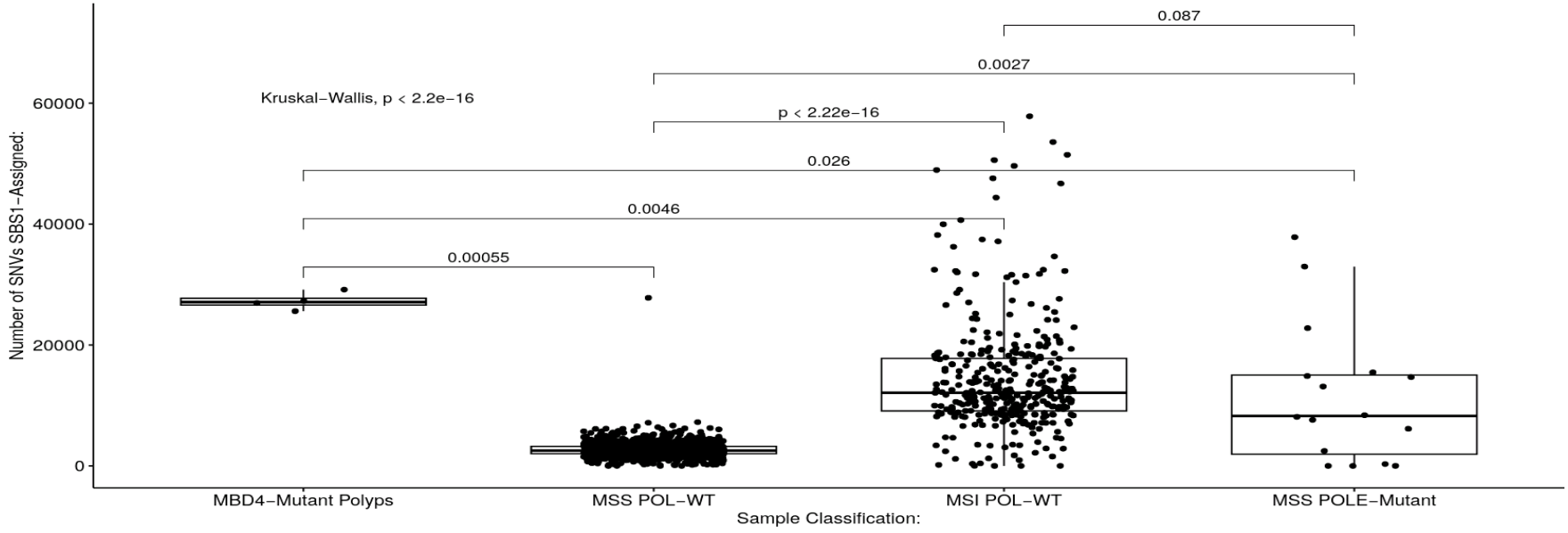
Figure 5.5 – The Mutation Profile of Colorectal Cancers with Microsatellite Instability & DNA Polymerase Exonuclease Domain Mutations: Characteristics of colorectal polyps extracted from a patient with a germline biallelic *MBD4* truncation and colorectal cancers from the 100,000 Genomes Project. Colorectal cancers were classified as microsatellite stable (MSS) or unstable (MSI^+) without mutations in the exonuclease domain of DNA polymerase-ε (POL-WT) and MSS cancers with *POL-ε* exonuclease domain mutations (*POLE*-Mutant). Included are the total single-nucleotide variation (SNV) counts from these individuals (a), the ages of the participants (b), the number of C → T mutations at CpG sites (c) and the proportion of the total SNV burden comprised by these mutations (d).

MSI^+ POL-WT cancers ($p = 1.5 \times 10^{-11}$). It was also observed that MSI^+ POL-WT CRCs were significantly older than their MSS POL-WT counterparts ($p = 7.4 \times 10^{-7}$, Figure 5.5b) and the *MBD4*-deficient individual from whom the polyps were extracted ($p = 0.0026$, Figure 5.5b). MSS CRCs with *POL-ε* EDMs were significantly younger than their MSS ($p = 0.00025$, Figure 5.5b) and MSI^+ ($p = 2.3 \times 10^{-5}$, Figure 5.5b) POL-WT counterparts. Interestingly, MSI^+ POL-WT and MSS *POL-ε* mutant cancers had a significantly greater number of C → T mutations at CpG sites than MSS POL-WT CRCs ($p < 2.2 \times 10^{-16}$ and $p = 5.6 \times 10^{-12}$ respectively – Figure 5.5c). Similarly to the total SNV burden, the number of C → T mutations at CpG sites was significantly greater in MSS *POL-ε* mutant CRCs than MSI^+ POL-WT cancers ($p = 3 \times 10^{-11}$, Figure 5.5c). However, the proportion of the total SNV burden made up of C → T mutations at CpG sites was significantly smaller in MSI^+ POL-WT and MSS *POL-ε* mutant CRCs compared to MSS POL-WT cancers ($p = 3 \times 10^{-5}$ and $p = 1.3 \times 10^{-5}$ respectively – Figure 5.5d) and the *MBD4*-mutant colorectal polyps ($p = 0.00059$ and $p = 0.00041$ respectively – Figure 5.5d). This was possibly a consequence of the increased number of SNVs in other contexts to C → T mutations at CpG sites in these cancers, including the C → A mutations common in cancers with *POL-ε* EDMs (SBS10a) and T → C mutations common in MSI^+ cancers (SBS26).

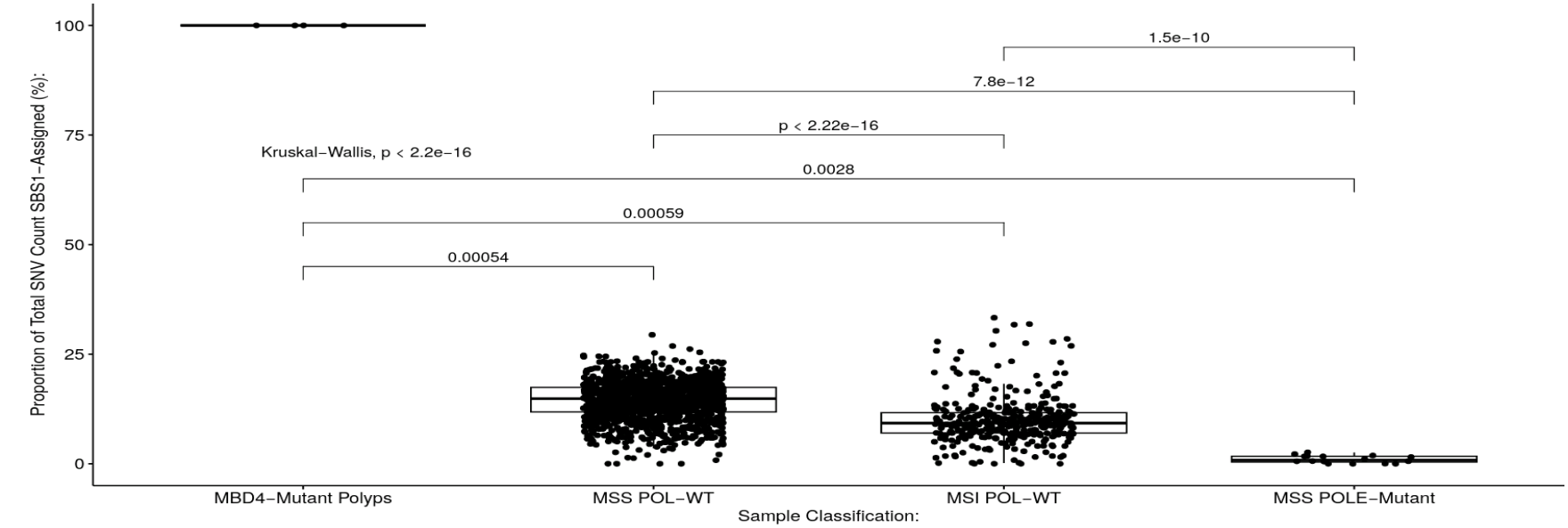
Following mutation signature extraction from the *MBD4*-mutant colorectal polyps and CRCs described above, the number of SNVs attributed to SBS1 was significantly greater in the *MBD4*-mutant colorectal polyps ($p = 0.00055$, Figure 5.6a), MSI^+ POL-WT CRCs ($p < 2.2 \times 10^{-16}$, Figure 5.6a) and MSS *POL-ε* mutant CRCs ($p = 0.0027$, Figure 5.6a) compared to the MSS POL-WT cancers. However, only the *MBD4*-mutant polyps ($p = 0.00054$, Figure 5.6b) presented with a greater proportion of SNVs attributed to SBS1 than the MSS POL-WT cancers. The MSI^+ POL-WT ($p < 2.2 \times 10^{-16}$, Figure 5.6b) and MSS *POL-ε* mutant ($p = 7.8 \times 10^{-12}$, Figure 5.6b) CRCs presented with a significantly smaller proportion of SNVs attributed to SBS1 than MSS POL-WT cancers. Unsurprisingly, only the MSS *POL-ε* mutant CRCs presented with SBS10b (Figure 5.6c and Figure 5.6d), whereas only the MSI^+ POL-WT cancers presented with SBS15 (Figure 5.6e and Figure 5.6f), indicating that there were differences in the mutation signatures of these groups of CRCs.

In order to characterise the association between DNA methylation and the rate of C → T mutagenesis at CpG sites in these groups of cancers, C → T mutation at CpG sites were binned into the same twelve DNA methylation bins described in Chapter IV of this thesis. As seen in Figure 5.7, there were significant correlations between DNA methylation and the rate of C → T mutagenesis at CpG sites in the *MBD4*-mutant colorectal polyps ($r^2 = 0.9319$, $p < 0.00001$), MSS POL-WT CRCs ($r^2 = 0.7269$, $p = 0.000426$), MSI^+ POL-WT CRCs ($r^2 = 0.7549$, $p = 0.000245$) and MSS CRCs with *POL-ε* EDMs ($r^2 = 0.6902$, $p = 0.000816$). As previously discussed in Chapter IV of this thesis, the regression slope for the association between DNA methylation and the rate of C → T mutagenesis at CpG sites was significantly steeper in the *MBD4*-mutant colorectal polyps ($\alpha = 1130.6$) than in the MSS POL-WT cancers ($\alpha = 78.7$, $p < 0.0001$). Interestingly, the regression slopes of the MSI^+ POL-WT ($\alpha = 396.8$) and MSS *POL-ε* mutant ($\alpha = 4{,}070$) CRCs were also significantly steeper than that of the MSS POL-WT cancers ($p < 0.0001$). This may suggest that as well as in the *MBD4*-mutant polyps, highly-methylated CpG sites in MSI^+ POL-WT and MSS *POL-ε* mutant CRCs may also be at greater risk of C → T mutagenesis than more lowly-methylated CpG sites. The regression constant of the MSS POL-WT cancers was not significantly different to that

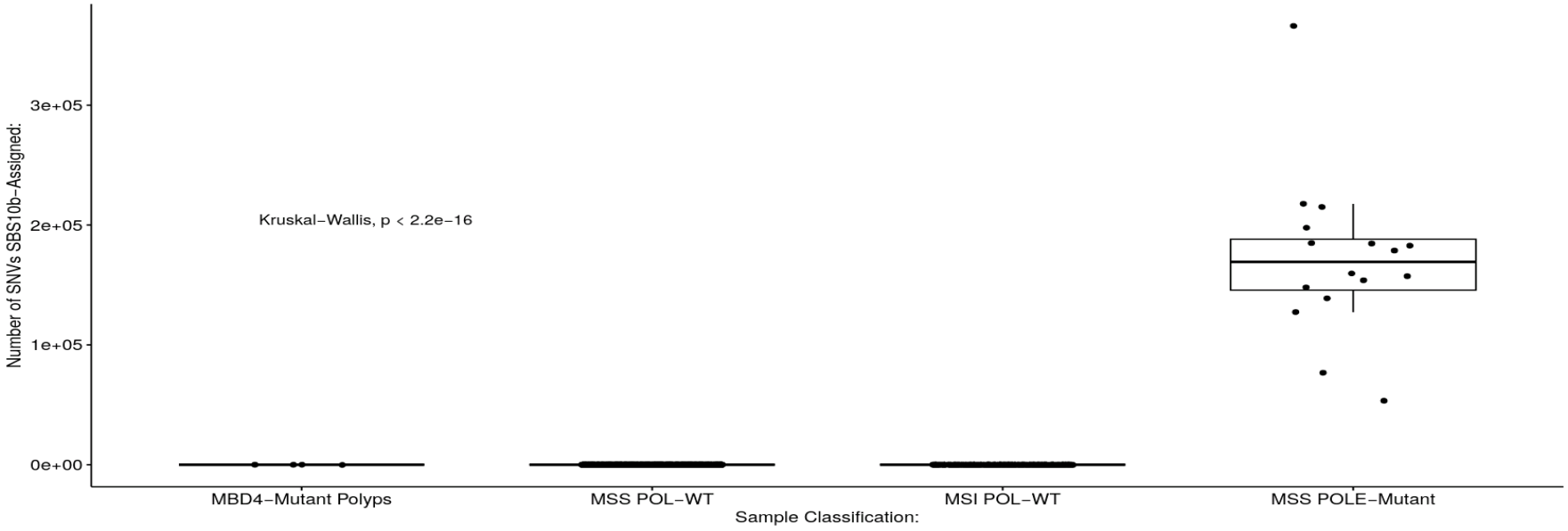
a)


0.087
0.0027
Kruskal–Wallis, p < 2.2e–16
p < 2.22e–16
0.026
0.0046
0.00055
60000
40000
20000
0
Number of SNVs SBS1–Assigned:
MBD4–Mutant Polyps
MSS POL–WT
MSI POL–WT
MSS POLE–Mutant
Sample Classification:

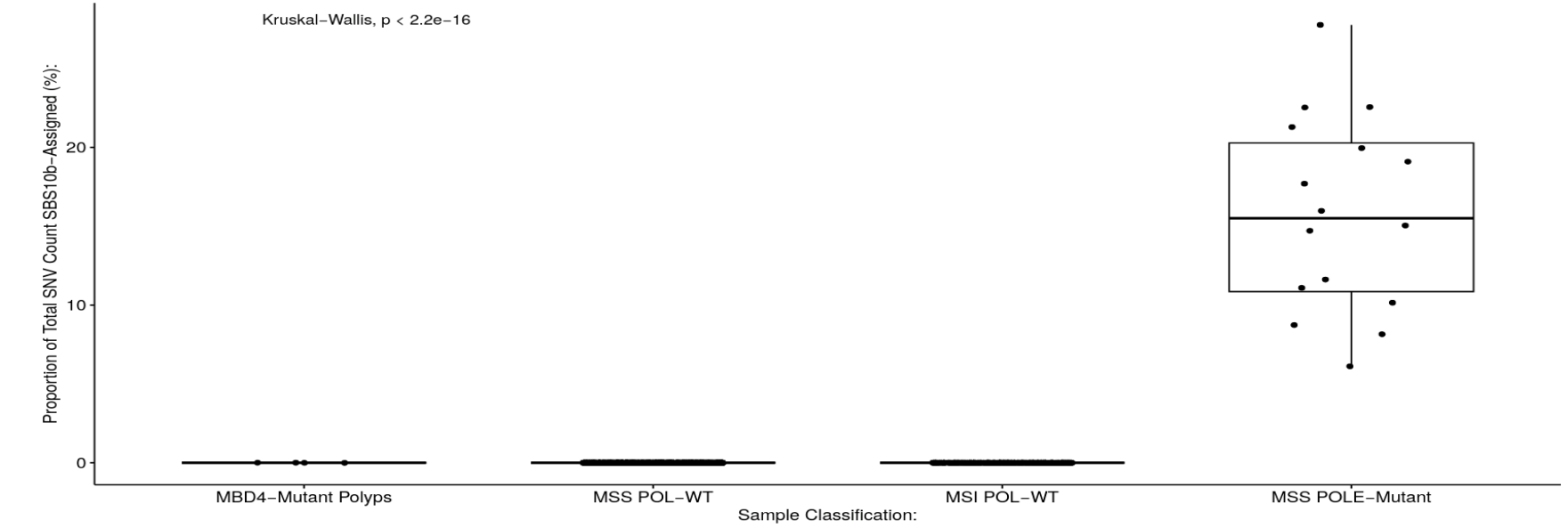
b)


1.5e–10
7.8e–12
p < 2.22e–16
Kruskal–Wallis, p < 2.2e–16
0.0028
0.00059
0.00054
100
75
50
25
0
Proportion of Total SNV Count SBS1–Assigned (%):
MBD4–Mutant Polyps
MSS POL–WT
MSI POL–WT
MSS POLE–Mutant
Sample Classification:

c)


Number of SNVs SBS10b-Assigned:
3e+05
2e+05
1e+05
0e+00
Kruskal–Wallis, p < 2.2e–16
MBD4–Mutant Polyps
MSS POL–WT
MSI POL–WT
MSS POLE–Mutant
Sample Classification:

d)


Kruskal–Wallis, p < 2.2e–16
Proportion of Total SNV Count SBS10b-Assigned (%):
20
10
0
MBD4–Mutant Polyps
MSS POL–WT
MSI POL–WT
MSS POLE–Mutant
Sample Classification:

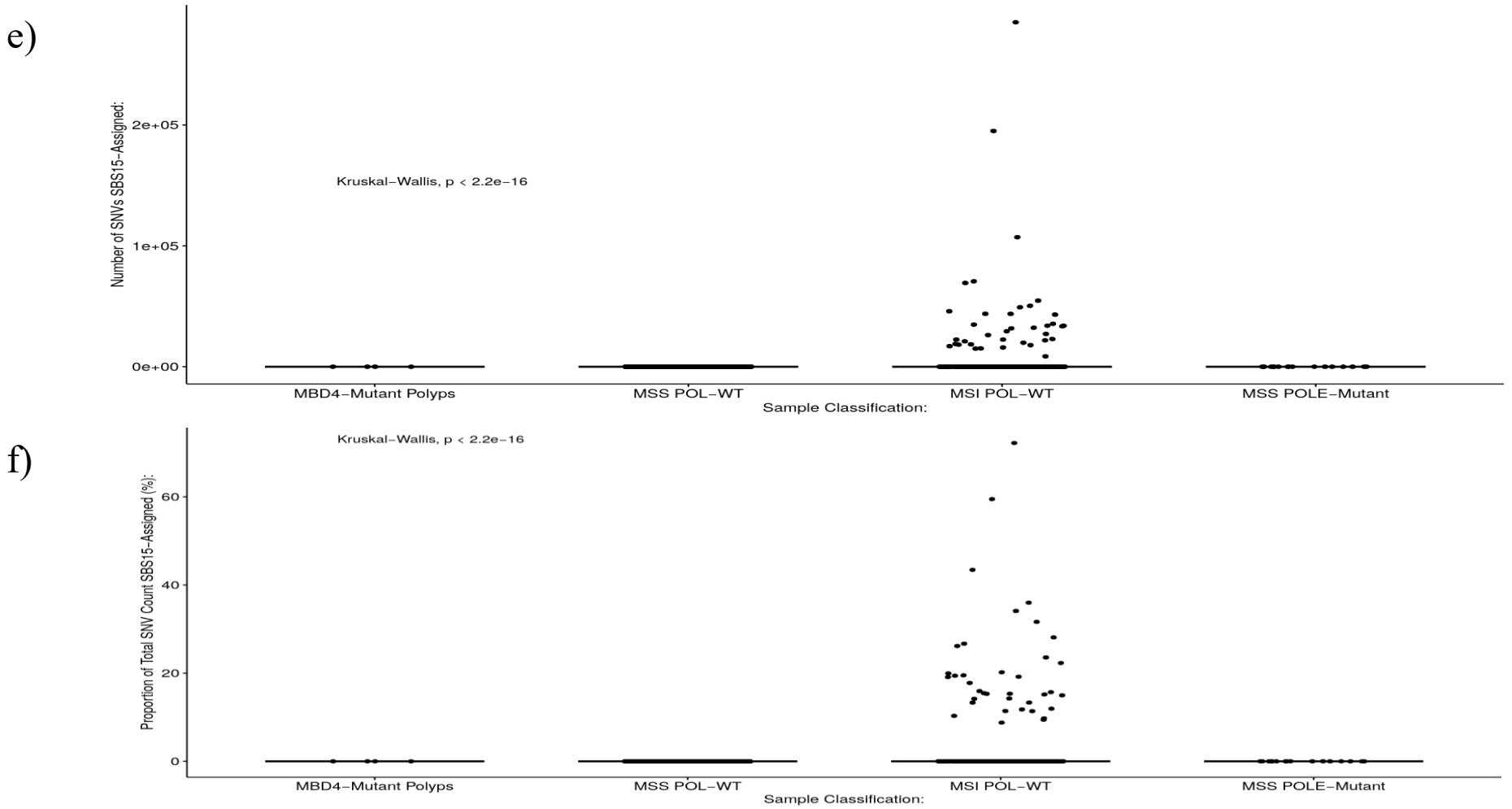


Figure 5.6 – Mutation Signature Distribution of Colorectal Cancers with Microsatellite Instability & DNA Polymerase Exonuclease Domain Mutations: The number and proportion of single-nucleotide variations (SNVs) attributed to specific mutation signatures in colorectal polyps extracted from an individual with a germline biallelic *MBD4* truncation, microsatellite stable (MSS) or unstable (MSI^{+}) colorectal cancers without DNA polymerase exonuclease domain mutations (POL-WT) and MSS colorectal cancers with *POL-ε* exonuclease domain mutations. Included are the number (a) and proportion (b) of SNVs attributed to SBS1, the number (c) and proportion (d) of SNVs attributed to SBS10b and the number (e) and proportion (f) of SNVs attributed to SBS15.

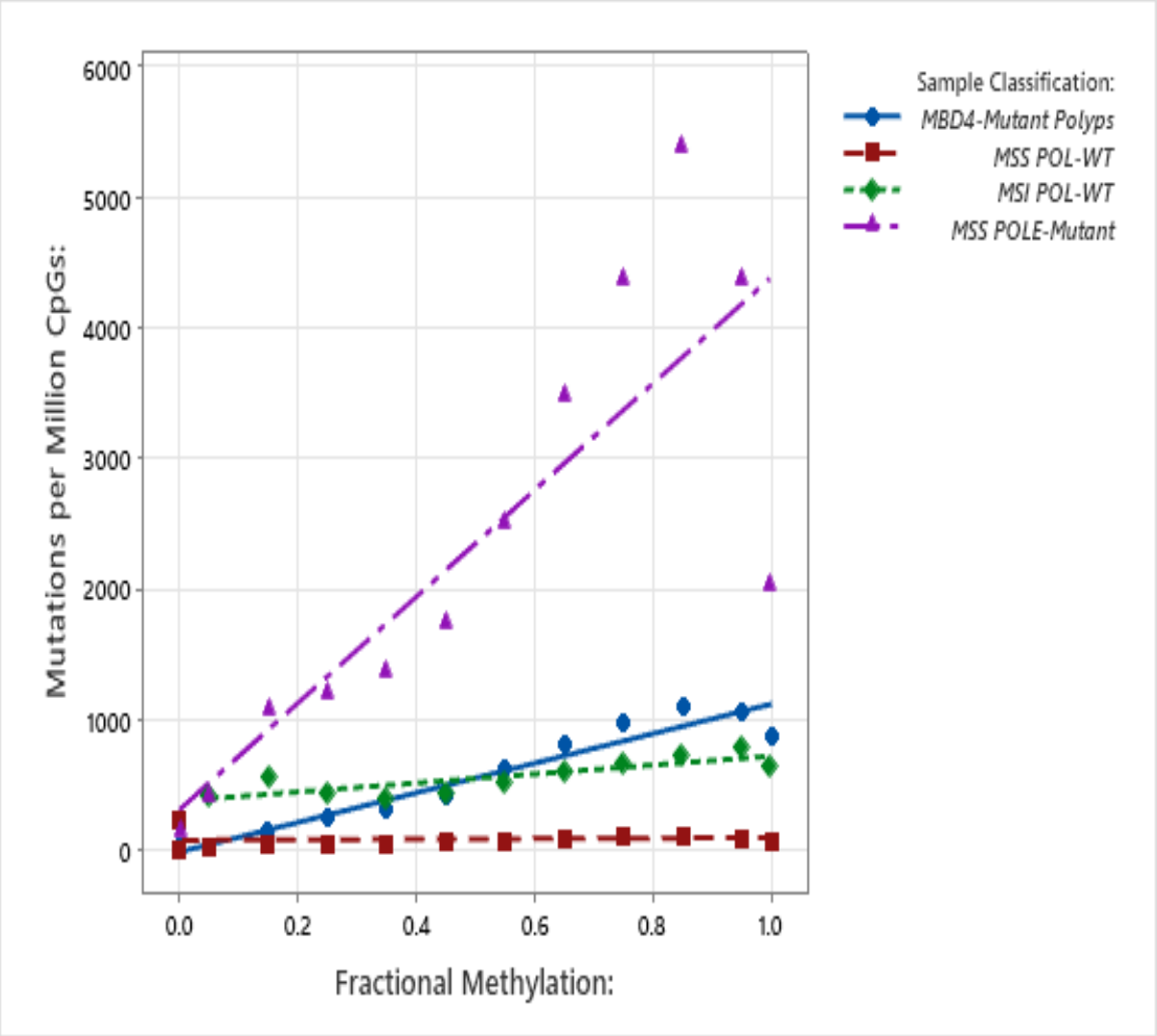


Figure 5.7 – The Effect of DNA Methylation on C → T Mutagenesis at CpG Sites in *MBD4*-Mutant Colorectal Polyps & Colorectal Cancers: The association between the C → T mutation rate at CpG sites and DNA methylation in colorectal cancer (CRC) whole-genome sequencing data obtained from the 100,000 Genomes Project and data from colorectal polyps. Included are colorectal polyps from a patient with a germline biallelic *MBD4* truncation, microsatellite stable (MSS) DNA polymerase (POL) wild-type (WT) CRCs, microsatellite unstable (MSI^+) POL-WT CRCs and MSS cancers with pathogenic mutations in the exonuclease domain of *POL-ε* (*POLE*-Mutant).

of the *MBD4*-mutant colorectal polyps ($p = 0.516$) or the MSS *POL-ε* mutant CRCs ($p = 0.602$) – indicating that, at the point where DNA methylation is zero, there was no difference in the rate of C → T mutagenesis at CpG sites between these groups. The regression constant of the MSI^+ POL-WT cancers was significantly greater than in the MSS POL-WT cancers ($p < 0.0001$), suggesting that the rate of C → T mutagenesis at unmethylated CpG sites was higher in these cancers.

The association between DNA replication timing and the rate of C → T mutagenesis at CpG sites was then investigated in these polyps and CRCs (see Chapter IV of this thesis for a detailed description of the method). As seen in Figure 5.8 and Table 5.4, there were

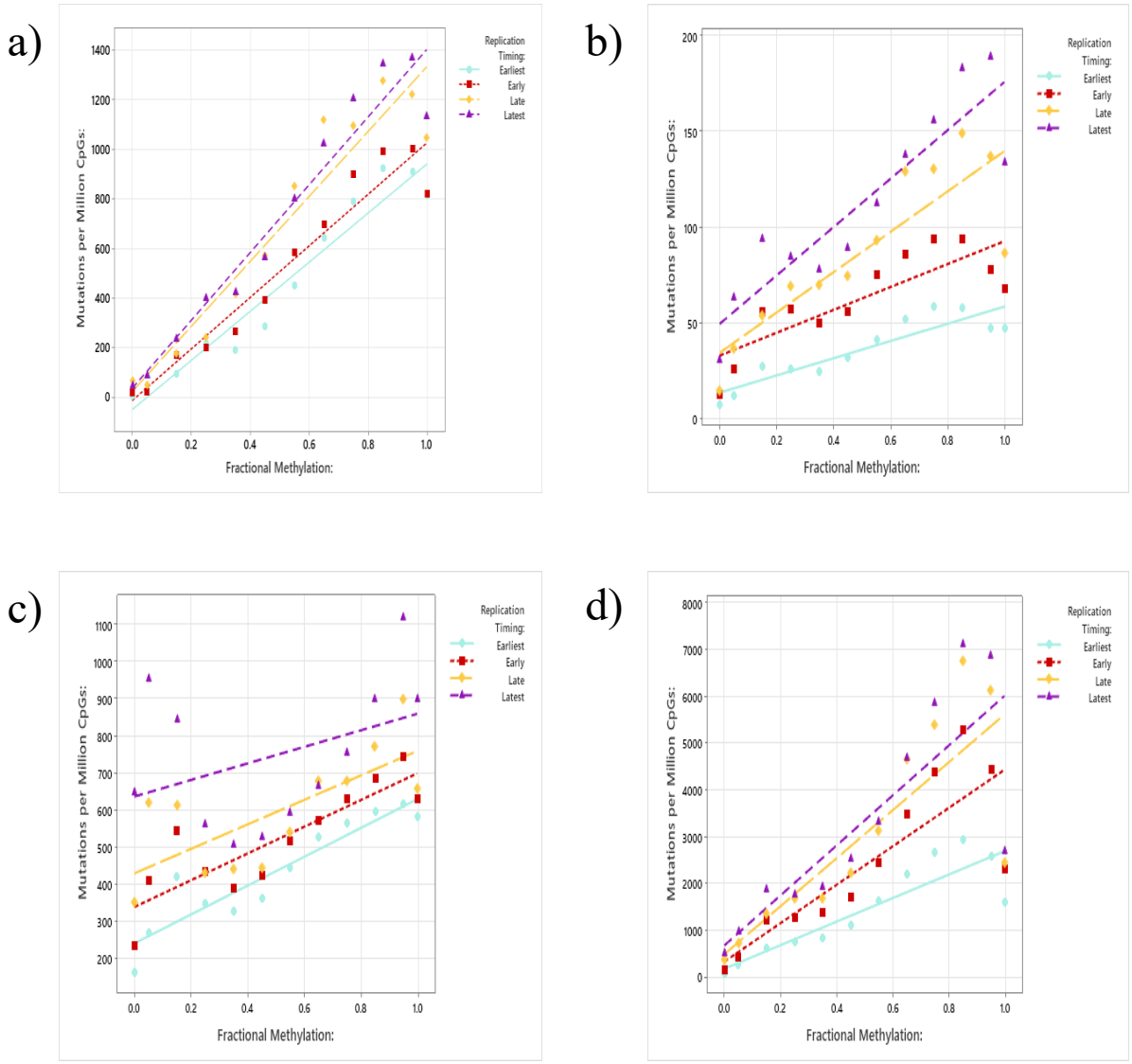


Figure 5.8 – The Effect of Replication Timing on C → T Mutagenesis at CpG Sites in *MBD4*-Mutant Colorectal Polyps & Colorectal Cancers: The association between DNA methylation and the rate of C → T mutagenesis at CpG sites in the earliest (blue), early (red), late (yellow) and latest (purple) replicating regions of the genome in *MBD4*-mutant colorectal polyps (a), microsatellite stable (MSS) DNA polymerase wild-type (POL-WT) colorectal cancers (b), microsatellite unstable POL-WT colorectal cancers (c) and MSS *POL-ε* mutant colorectal cancers (d).

significant positive correlations between DNA methylation and the rate of C → T mutagenesis at CpG sites in all four replication timing bins in the *MBD4*-mutant colorectal polyps (Figure 5.8a), MSS POL-WT CRCs (Figure 5.8c) and MSS *POL-ε* mutant cancers (Figure 5.8d). However, this significant correlation was only seen in three replication timing bins in MSI^+ POL-WT CRCs (Figure 5.8c). In the latest replication timing bin, the association between DNA methylation and the rate of C → T mutagenesis at CpG sites was not significant ($r^2 = 0.1577$, p = 0.20119 – Table 5.4). As seen in Table 5.5, there was a significant cumulative increase in the regression slope for the relationship between DNA methylation and the rate of C → T mutagenesis at CpG sites in the *MBD4*-mutant colorectal polyps (Methylation x Replication Timing = 140.2, p = 0.003 – Table 5.5), MSS POL-WT

***MBD4*-Mutant Colorectal Polyps (n = 4):**					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	990.8	-50.2	0.9316	0.9652	< 0.00001
Early	1040.1	-15.1	0.9211	0.9597	< 0.00001
Late	1311.2	21.5	0.8938	0.9478	< 0.00001
Latest	1367.8	34.9	0.9289	0.9638	< 0.00001
MSS POL-WT Colorectal Cancers (n = 1,532):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	44.9	13.52	0.8091	0.8995	0.000068
Early	59.8	32.58	0.656	0.8222	0.00103
Late	105.4	34.05	0.7261	0.8521	0.000432
Latest	126.4	49.2	0.807	0.8983	0.000072
MSI^+ POL-WT Colorectal Cancers (n = 357):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	389	240.5	0.8663	0.9308	0.000011
Early	361	337.4	0.7476	0.8646	0.000284
Late	331	428.3	0.5157	0.7181	0.008531
Latest	224	636.2	0.1577	0.3971	0.20119
MSS *POL-ε* Mutant Colorectal Cancers (n = 16):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	2,519	179	0.7695	0.8772	0.000178
Early	4,144	319	0.7018	0.8377	0.000671
Late	5,139	477	0.6427	0.8017	0.001705
Latest	5,352	667	0.6398	0.7999	0.001786

Table 5.4 – Replication Timing Regression Analysis of *MBD4*-Mutant Colorectal Polyps & Colorectal Cancer: The regression equations of the relationship between DNA methylation and C → T mutation rates at CpG sites in *MBD4*-mutant colorectal polyps, microsatellite stable (MSS) DNA polymerase wild-type (POL-WT) colorectal cancers, microsatellite unstable (MSI^+) POL-WT colorectal cancers and MSS colorectal cancers with *POL-ε* exonuclease domain mutations. Included are the replication timing bin, the regression slope, the regression constant, the methylation-mutation rate r^2 correlation, the Pearson's R correlation measure of this relationship and the p-value associated with the Pearson's R statistic ($p_{(Pearson's\ R)}$).

***MBD4*-Mutant Colorectal Polyps (n = 4):**		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	967.1	< 0.0001
Replication Timing	29.2	0.276
Methylation x Replication Timing Interaction Term	140.2	0.003
Constant	-46	0.358
MSS POL-WT Colorectal Cancers (n = 1,524):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	40.6	0.002
Replication Timing	10.85	0.009
Methylation x Replication Timing Interaction Term	29.02	< 0.0001
Constant	16.06	0.036
MSI^+ POL-WT Colorectal Cancers (n = 357):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	405.2	< 0.0001
Replication Timing	127.8	< 0.0001
Methylation x Replication Timing Interaction Term	-52.7	0.253
Constant	218.9	< 0.0001
MSS *POL-ε* Mutant Colorectal Cancers (n = 16):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	2,852	0.001
Replication Timing	162	0.51
Methylation x Replication Timing Interaction Term	952	0.024
Constant	167	0.716

Table 5.5 – The Effect of Replication Timing on Methylation-Mutation Rate Associations in *MBD4*-Mutant Colorectal Polyps & Colorectal Cancers: Regression equations for the effect of fractional DNA methylation and replication timing on C → T mutation rates at CpG sites in *MBD4*-mutant colorectal polyps, microsatellite stable (MSS) DNA polymerase wild-type (POL-WT) colorectal cancers, microsatellite unstable (MSI^+) POL-WT colorectal cancers and MSS colorectal cancers with *POL-ε* exonuclease domain mutations. Included are the variables of the equation – DNA methylation, replication timing, the interaction term between the two, the regression constant and the p-value of each ($p_{(Coefficient)}$).

CRCs (Methylation x Replication Timing = 29.02, $p < 0.0001$ – Table 5.5) and MSS *POL-ε* mutant CRCs (Methylation x Replication Timing = 952, $p = 0.024$ – Table 5.5). This suggests that highly-methylated CpG sites in late-replicating regions of the genome appear to be at greater risk of C → T mutagenesis in these polyps and cancers that lowly-methylated CpG sites in early-replicating regions of the genome. In the MSI^+ POL-WT CRCs, there was no significant cumulative change in the regression slope between the replication timing bins (Methylation x Replication Timing = -52.7, $p = 0.253$ – Table 5.5), indicating that the rate of C → T mutagenesis at CpG sites is less influenced by replication timing in these cancers – as previously shown in Chapter IV of this thesis.

While the above data indicate that there may be some similarities between the *MBD4*-mutant colorectal polyps, MSI^+ POL-WT CRCs and MSS CRCs with *POL-ε* EDMs (e.g. increased numbers of C → T mutations at CpG sites compared to MSS POL-WT CRCs and a significant association between DNA methylation and the rate of C → T mutagenesis at CpG sites), there is little in this data that may indicate the mechanism(s) underlying C → T mutagenesis at CpG sites in these samples. Previous data by Tomkova *et al.* suggested that C → T mutagenesis in *POL-ε* mutant CRCs is primarily driven by unrepaired *POL-ε* replication errors propagating into C → T mutations in the next round of DNA replication, whilst also

suggesting that this may also be the cause of C → T mutations at CpG sites in MSI^+ cancers (524). As discussed in section 5.1.2, *POL-ε* is thought to be responsible for the synthesis of the leading strand during DNA replication, whereas *POL-δ* replicates the lagging strand (489). As a consequence of this, it has previously been reported that cancers with pathogenic mutations in the exonuclease domain of *POL-ε* present with an increased number of mutations on the template for leading strand synthesis (485,524). As presented in Figure 5.1b, the mutation signature SBS10b – which is associated with defective *POL-ε* DNA "proofreading" – is characterised by a significant enrichment of C → T mutations in the context T**C**G. Of the sixteen *POL-ε* mutant cancers described in section 5.3.1, a median of 71.6% (range of 61.32% –76.82%) of C → T mutations at CpG sites were in the context T**C**G. Therefore, it is plausible that unrepaired errors made by *POL-ε* during the replication of the leading strand were the cause of the majority of the C → T mutations at CpG sites in these *POL-ε* mutant CRCs. However, in MSI^+ POL-WT CRCs, these mutations only accounted for a median of 10.9% of C → T mutations at CpG sites (range of 3.95% – 16.94%).

In these MSS CRCs with *POL-ε* EDMs, there are two possible mechanisms by which an erroneous DNA mismatch produced by *POL-ε* during DNA replication may be propagated into a C → T mutation at a CpG site in the next round of replication (see Figure 5.9 and Figure 5.10). These include the generation of a C:A DNA mismatch (Figure 5.9) or a T:G mismatch (Figure 5.10) which are then propagated into C → T or G → A mutations in the next round of DNA replication. As seen in Figure 5.9, if C:A mismatches were the cause of C → T mutagenesis at CpG sites in these *POL-ε* mutant CRCs, it would be expected that C → T mutations would be more common in left-replicating regions of the genome than G → A mutations. Furthermore, G → A mutations would be expected to be more common than C → T mutations in right-replicating DNA. Conversely, if T:G DNA mismatches were the driving force underpinning C → T mutagenesis at CpG sites, G → A mutations would be expected to be more common than C → T mutations in left-replicating DNA and C → T mutations would be more common than G → A mutations in right-replicating DNA (see Figure 5.10).

In order to identify which of the above mechanisms may drive C → T mutagenesis at CpG sites in these *POL-ε* mutant CRCs, replication strand data was obtained from the study by Haradhvala *et al.,* who divided the genome into 20,000 base-pair regions and classified regions as left-replicating or right-replicating based on replication timing data obtained from lymphoblastoid cell lines (520). In total, 9,638,264 CpG sites could be assigned to one of these states. The C → T mutations at CpG sites of the *POL-ε* mutant cancers, either categorised as C → T or G → A mutations, could then be assigned to either left-replicating or right-replicating regions of the genome. As seen in Figure 5.11a, the ratio of C → T / G → A mutations in indicated an excess of C → T mutations in left-replicating DNA and an excess of G → A mutations in right-replicating DNA ($p = 3.1 \times 10^{-5}$). Therefore, this data indicates that C → T mutations at CpG sites in these cancers were potentially driven by unrepaired C:A DNA mismatches produced during replication of the leading strand which are propagated into mutations in the next round of DNA replication.

From this data, it could be concluded that C → T mutations in left-replicating DNA and G → A mutations in right-replicating DNA correspond to C → T mutations on the leading strand template, whereas C → T mutations in right-replicating DNA and G → A mutations in left-

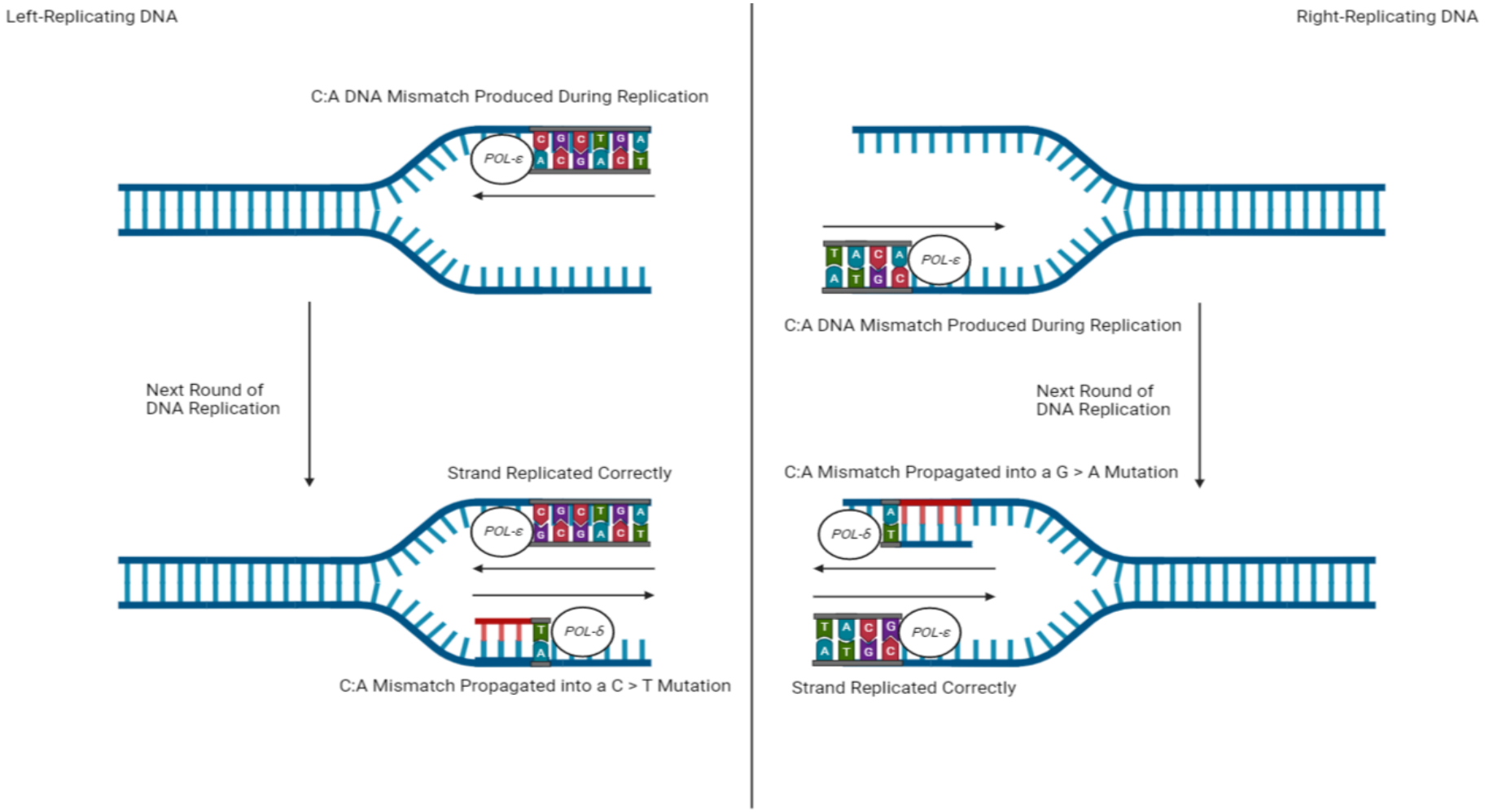


Figure 5.9 – C → T Mutagenesis at CpG Sites Driven by C:A Mismatches Erroneously Produced During DNA Replication: The mechanism by which C:A DNA mismatches produced by DNA polymerase-ε (*POL-ε*) during replication of the leading strand could propagate into C → T mutations at CpG sites in the next round of DNA replication. In left-replicating DNA (left), the C:A mismatch is propagated into a C → T mutation by DNA polymerase-δ (*POL-δ*). In right-replicating DNA (right), the C:A mismatch produced by *POL-ε* is propagated by *POL-δ* into a G → A mutation. Created with BioRender.com (https://app.biorender.com/).

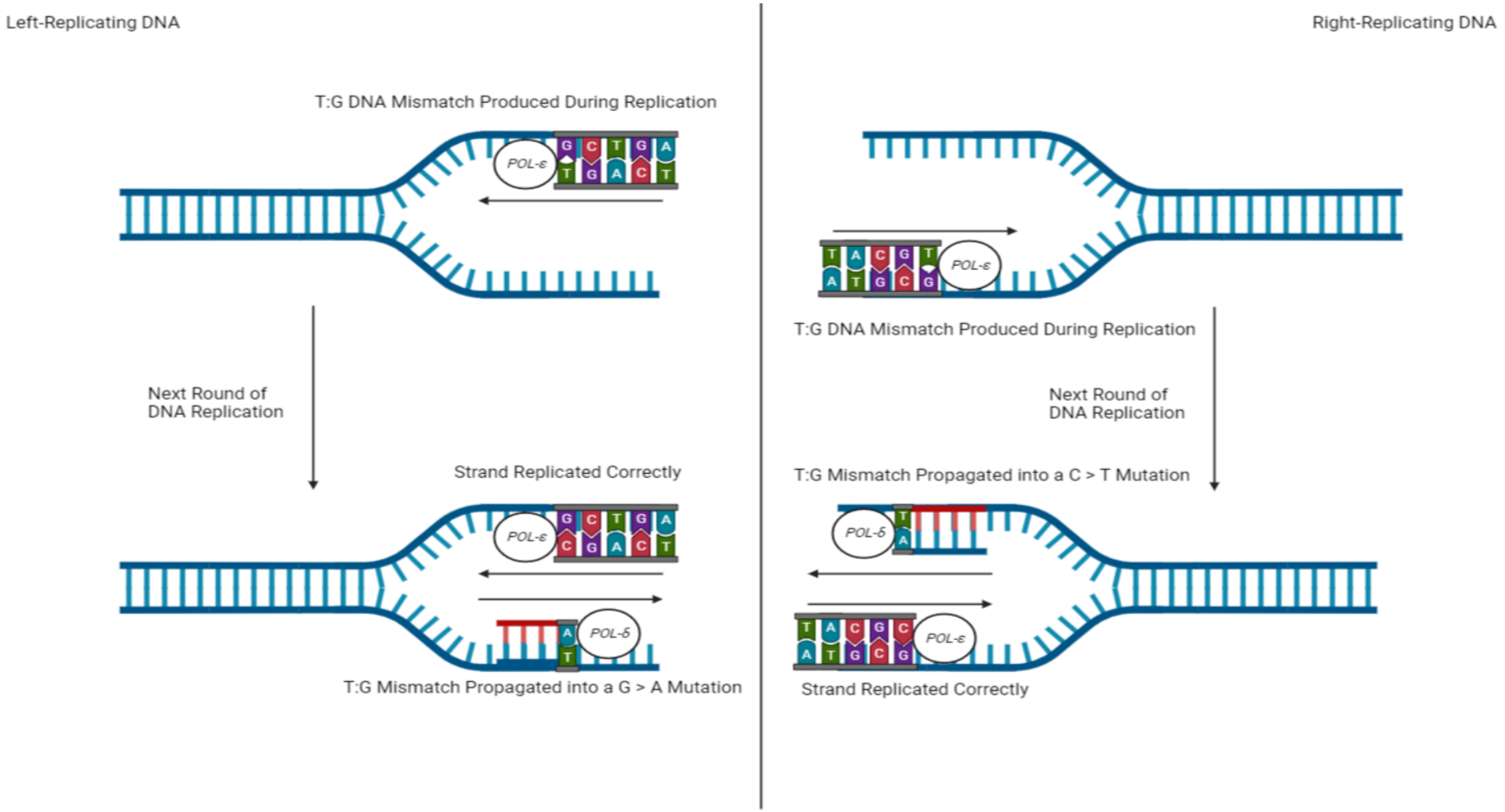


Figure 5.10 – C → T Mutagenesis at CpG Sites Driven by T:G Mismatches Erroneously Produced During DNA Replication: The mechanism by which T:G DNA mismatches produced by DNA polymerase-ε (*POL-ε*) during replication of the leading strand could propagate into C → T mutations at CpG sites in the next round of DNA replication. In left-replicating DNA (left), the T:G mismatch is propagated into a G → A mutation by DNA polymerase-δ (*POL-δ*). In right-replicating DNA (right), the T:G mismatch produced by *POL-ε* is propagated by *POL-δ* into a C → T mutation. Created with BioRender.com (https://app.biorender.com/).

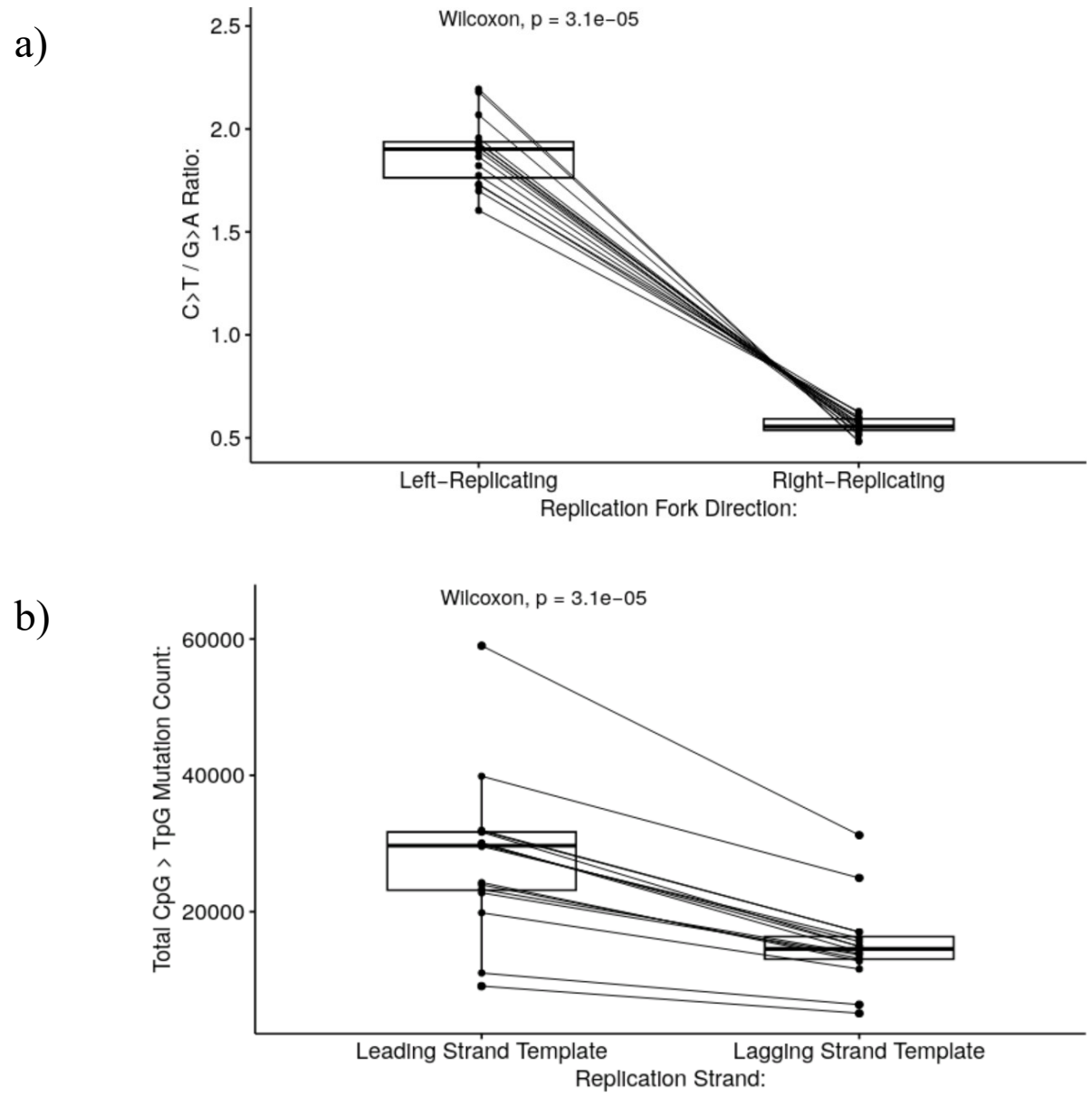


Figure 5.11 – Replication Strand Asymmetries of MSS Colorectal Cancers with Pathogenic DNA Polymerase-ε Exonuclease Domain Mutations: The ratio of C → T / G → A mutations in left-replicating and right-replicating regions of the genome in microsatellite stable colorectal cancers with pathogenic DNA polymerase-ε exonuclease domain mutations (a). Also shown are the number of C → T mutations at CpG sites of the leading strand template and lagging strand templates of these cancers (b).

replicating DNA are more likely to be on the lagging strand template. By this mechanism, cancers with *POL-ε* EDMs presented with significantly more mutations on the template for the leading strand than the lagging strand ($p = 3.1 \times 10^{-5}$, Figure 5.11b), which is in agreement with what has been previously reported in the literature (485,518,520,524).

Using this convention for assigning C → T mutations at CpG sites to the template for leading or lagging strand synthesis, a $\log_2$(Leading/Lagging) ratio of mutations could be calculated for each individual cancer. As seen in Figure 5.12, *MBD4*-mutant colorectal polyps and MSS POL-WT CRCs presented with a $\log_2$(Leading/Lagging) ratio of close to zero – perhaps indicating that these C → T mutations at CpG sites were driven by spontaneous deamination

of 5-mC, given that the spontaneous deamination of 5-mC is theoretically equally likely to occur on either replication strand. While there was no significant difference in the log_2(Leading/Lagging) ratio of mutations between the *MBD4*-mutant polyps and MSS POL-WT CRCs ($p = 0.64$, Figure 5.12), a number of MSS POL-WT CRCs appeared to show an excess of mutations on the lagging strand template. Conversely, MSI^+ POL-WT CRCs and MSS *POL-ε* mutant CRCs both presented with a log_2(Leading/Lagging) ratio that was significantly greater than their MSS POL-WT countrparts ($p < 2.2 \times 10^{-16}$ and $p = 6 \times 10^{-12}$ respectively, Figure 5.12) – indicating that these cancers have a significant excess of C → T mutations at CpG sites on the leading strand template. This data suggests that unrepaired *POL-ε* replication errors may be the driving force underpinning C → T mutagenesis in these cancers, which is plausible given the role of both *POL-ε* "proofreading" and the MMR pathway in the repair of DNA mismatches caused by replication errors (see section 5.1). However, the log_2(Leading/Lagging) ratio of mutations was signifcantly greater in the MSS *POL-ε* mutant CRCs than the MSI^+ POL-WT cancers ($p = 1.4 \times 10^{-11}$, Figure 5.12).

Overall, the data presented above indicates that MSI^+ POL-WT CRCs and MSS CRCs with *POL-ε* EDMs present with a greater number of C → T mutations at CpG sites than their MSS POL-WT counterparts. In addition to this, these cancers also present with a steeper regression slope between DNA methylation and the rate of C → T mutagenesis at CpG sites than the MSS POL-WT cancers, perhaps indicating that highly-methylated CpG sites are at greatest risk of C → T mutagenesis in these cancers. Interestingly, it appears than both MSI^+ POL-WT CRCs and MSS *POL-ε* mutant cancers show a significant enrichment of C → T mutations at CpG sites on the template for leading strand synthesis compared to MSS POL-WT cancers, suggesting that DNA replication errors may be the driving force underpinning C → T mutagenesis at CpG sites in these cancers. As discussed in section 5.1, the hyper-mutated nature of cancers with *POL-ε* EDMs is thought to be a result of abolished DNA "proofreading" mechanisms and a consequent accumulation of replication errors which are propagated into mutations. However, previous studies of MSI^+ POL-WT cancers suggest that C → T mutagenesis at CpG sites may be driven by different mechanisms depending on the MMR gene mutated, with MutS-deficient cancers thought to accumulate mutations via unrepaired spontaneous deaminations of 5-mC and MutL-deficient cancers thought to accumulate mutations via unrepaired DNA replication errors (see section 5.1.3.2). Therefore, in order to further investigate this hypothesis, the MSI^+ POL-WT cancers should be categorised as either MutS-deficient or MutL-deficient and the above analyses repeated, with particular emphasis placed on the replication strand analysis to identify the likely mechanism(s) underpinning C → T mutagenesis at CpG sites in these cancers.

5.3.2 – Characteristics of MutS-Deficient & MutL-Deficient MSI^+ POL-WT CRCs

Previous studies by Fang *et al.* and Sanders *et al.* have suggested that MutS-deficient cancers present with more C → T mutations at CpG sites than their MutL-deficient counterparts (518,519). In order to investigate this in whole-genome sequencing data of CRCs from the

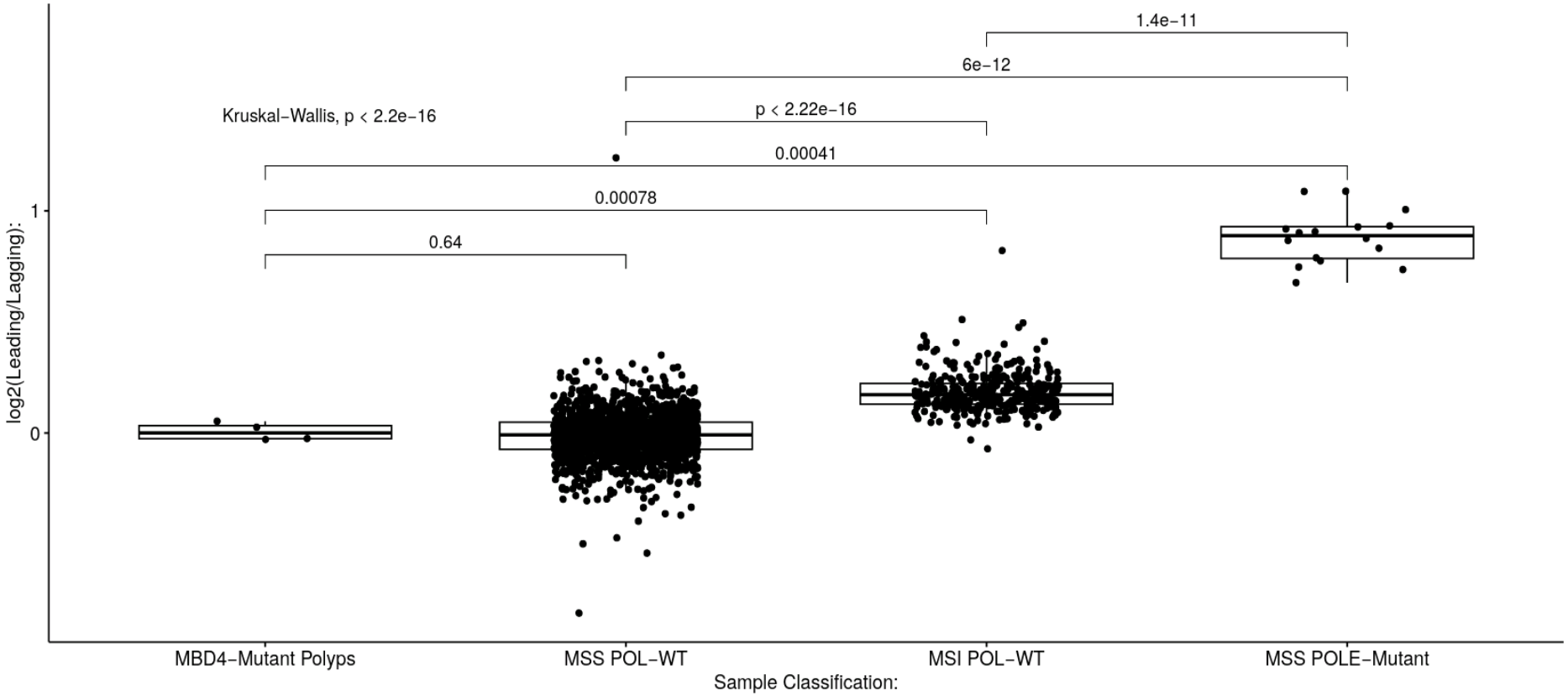


Figure 5.12 – Replication Strand Analysis of *MBD4*-Mutant Colorectal Polyps and Colorectal Cancers: The $\log_2$(Leading/Lagging) ratio of C → T mutations at CpG sites in colorectal polyps derived from a patient with a germline biallelic *MBD4* truncation, microsatellite stable (MSS) DNA polymerase wild-type (POL-WT) colorectal cancers, microsatellite unstable (MSI^+) POL-WT colorectal cancers and MSS colorectal cancers with *POL-ε* exonuclease domain mutations (*POLE*-mutant).

100KGP, MSI^+ POL-WT CRCs were categorised as MutS-deficient if they presented with germline pathogenic mutations or somatic pathogenic mutations that were likely biallelic (as determined by χ^2 analysis) in either *MSH2* or *MSH6.* A total of six CRCs presented with germline *MSH2* mutations, which were categorised as LS-associated MutS-deficient CRCs (see Table 5.6). In addition to these LS samples, twenty-eight cancers presented with pathogenic somatic mutations in *MSH2*, resulting in a Bonferroni-corrected $p_{(\chi^2)}$ threshold of 0.00179 (0.05/28). Of these, three cancers presented a $p_{(\chi^2)}$ below this threshold and a $p_{(\chi^2)} \geq 0.05$ when the null hypothesis assumed that these mutations were homozygous – indicating that these *MSH2* mutations were likely biallelic (see Table 5.7). Similarly, nine MSI^+ POL-WT CRCs presented with germline *MSH6* mutations (see Table 5.6) and 128 CRCs presented with pathogenic *MSH6* mutations, resulting in a $p_{(\chi^2)}$ threshold of 0.000391 (0.05/128). Of these, four MSI^+ cancers presented with a $p_{(\chi^2)}$ below this threshold and a $p_{(\chi^2)} \geq 0.05$ when the null hypothesis assumes that these mutations were biallelic (see Table 5.7). In addition to these cancers, four cancers presented with multiple somatic *MSH6* truncations – indicating that these cancers may also have biallelic loss of *MSH6* (see Table 5.7). This resulted in a total of twenty-six MSI^+ POL-WT CRCs being classified as MutS-deficient, fifteen LS-associated and eleven somatic. A summary of these cancers is provided in Figure 5.13.

In addition to these cancers, six MSI^+ POL-WT cancers presented with germline pathogenic mutations in *MLH1*, resulting in their classification as LS-associated MutL-deficient cancers (see Table 5.6). A further thirty cancers had somatic pathogenic *MLH1* mutations, resulting in a Bonferroni-corrected $p_{(\chi^2)}$ threshold of 0.00167 (0.05/30). Four *MLH1*-mutant CRCs had a $p_{(\chi^2)}$ below this threshold and a $p_{(\chi^2)} \geq 0.05$ when the null hypothesis assumed the mutation was biallelic, indicating that the *MLH1* mutation harboured by the cancer was likely to be biallelic (see Table 5.7). One MSI^+ POL-WT CRC presented with multiple *MLH1* mutations, indicating that this cancer may also have pathogenic mutations in both alleles of the *MLH1* gene (see Table 5.7). Cancers with pathogenic mutations in *PMS2* were excluded from this analysis (see Figure 5.13) due to the presence of a number of pseudogenes at the *PMS2* locus which have been shown to be highly homologous with the *PMS2* gene – thus making genuine pathogenic *PMS2* mutations difficult to to identify (525).

As well as these *PMS2*-mutant cancers, the 105 CRCs with monoallelic somatic MutS mtations, the twelve CRCs with monoallelic somatic MutL mutations and the ten cancers with monoallelic somatic mutations in both MutS and MutL genes were excluded from downstream analyses as they could not be confidently classified as MutS-deficient or MutL-deficient (see Figure 5.13). This left a group of 189 MSI^+ CRCs with no mutations in any of the above MMR genes, indicating that the MSI^+ phenotype was likely driven by *MLH1* promoter hyper-methylation in these cancers (see Figure 5.13).

Following the classification of these cancers as MutS-deficient or MutL-deficient, the characteristics of these cancers were then compared. As seen in Figure 5.14a, there were no significant differences in the SNV burdens of most of the groups ($p_{(Kruskal\text{-}Wallis)} = 0.18$). Cancers with somatic biallelic MutS mutations presented with significantly more SNVs than cancers with presumed *MLH1* promoter hyper-methylation ($p = 0.043$, Figure 5.14a) – which represented the only significant difference between any of the groups. As seen in Figure 5.14b, the LS-associated MutS-deficient and MutL-deficient CRCs were significantly

Cancer #:	Germline Mutation:	Somatic Mutation:	Tumour Purity (%):	Cancer Classification:
1	*MSH2*^{L330P}	NA	23	Lynch Syndrome MutS-Deficient
2	*MSH2*^{Q76*}	NA	25	Lynch Syndrome MutS-Deficient
3	*MSH2*^{A776fs*}	NA	48	Lynch Syndrome MutS-Deficient
4	*MSH2*^{C333Y}	NA	72	Lynch Syndrome MutS-Deficient
5	*MSH2*^{G162R}	*MSH2*^{Q337*}	62	Lynch Syndrome MutS-Deficient
6	*MSH2*^{R711*}	*MSH2*^{R680*}	26	Lynch Syndrome MutS-Deficient
7	*MSH6*^{L1211P}	NA	41	Lynch Syndrome MutS-Deficient
8	*MSH6*^{T783fs*}	NA	40	Lynch Syndrome MutS-Deficient
9	*MSH6*^{G32fs*}	*MSH6*^{E270fs*}	40	Lynch Syndrome MutS-Deficient
10	*MSH6*^{A1055P}	*MSH6*^{Q33*}	54	Lynch Syndrome MutS-Deficient
11	*MSH6*^{R870fs*}	*MSH6*^{W44*}	72	Lynch Syndrome MutS-Deficient
12	*MSH6*^{K88*}	*MSH6*^{N71fs*}	72	Lynch Syndrome MutS-Deficient
13	*MSH6*^{E1109*}	*MSH6*^{T783fs*}	28	Lynch Syndrome MutS-Deficient
14	*MSH6*^{I589fs*}	*MSH6*^{T783fs*}	76	Lynch Syndrome MutS-Deficient
15	*MSH6*^{G686D}	*MSH6*^{T783fs*}	27	Lynch Syndrome MutS-Deficient
16	*MLH1*^{T117M}	NA	74	Lynch Syndrome MutL-Deficient
17	*MLH1*^{W469*}	NA	40	Lynch Syndrome MutL-Deficient
18	*MLH1*^{T117M}	NA	54	Lynch Syndrome MutL-Deficient
19	*MLH1*^{V7fs*}	NA	62	Lynch Syndrome MutL-Deficient
20	*MLH1*^{P579fs*}	NA	42	Lynch Syndrome MutL-Deficient
21	*MLH1*^{L573fs*}	NA	57	Lynch Syndrome MutL-Deficient

Table 5.6 – Characterisation of Lynch Syndrome MutS-Deficient & MutL-Deficient Colorectal Cancers: The details of the twenty-one microsatellite unstable (MSI^+) DNA polymerase wild-type colorectal cancers with germline pathogenic mutations in the DNA mismatch repair (MMR) genes *MSH2, MSH6* or *MLH1*. Included are the germline MMR mutation, any additional somatic MMR mutations, the estimated tumour purity and the final classification of the cancer – either Lynch Syndrome MutS-deficient or MutL-deficient. NA = Not Applicable.

Cancer #:	MMR Mutation:	Tumour Purity (%):	Alternate Reads:	Expected Reads if Heterozygous:	$P_{(\chi^2)}$:	Expected Reads if Homozygous:	$P_{(\chi^2)}$:	Cancer Classification:
1	*MSH2*^{T788Lfs*24}	40	62	18	< 0.00001	37	NA	Somatic Biallelic MutS-Deficient
2	*MSH2*^{E853*}	71	74	41	< 0.00001	82	0.0872	Somatic Biallelic MutS-Deficient
3	*MSH2*^{Y570*}	35	50	25	< 0.00001	50	0.9443	Somatic Biallelic MutS-Deficient
4	*MSH6*^{T783fs*}	53	63	34	< 0.00001	68	0.3914	Somatic Biallelic MutS-Deficient
5	*MSH6*^{T783fs*}	23	31	12	< 0.00001	24	0.1122	Somatic Biallelic MutS-Deficient
6	*MSH6*^{T783fs*}	38	42	19	< 0.00001	39	0.5087	Somatic Biallelic MutS-Deficient
7	*MSH6*^{T783fs*}	34	34	18	0.0000272	36	0.7261	Somatic Biallelic MutS-Deficient
8	*MSH6*^{I115fs*} + *MSH6*^{V299fs*}	26	NA	NA	NA	NA	NA	Somatic Biallelic MutS-Deficient
9	*MSH6*^{R193*} + *MSH6*^{D911fs*}	44	NA	NA	NA	NA	NA	Somatic Biallelic MutS-Deficient
10	*MSH6*^{F116fs*} + *MSH6*^{T783fs*}	25	NA	NA	NA	NA	NA	Somatic Biallelic MutS-Deficient
11	*MSH6*^{R249*} + *MSH6*^{T783fs*}	38	NA	NA	NA	NA	NA	Somatic Biallelic MutS-Deficient
12	*MLH1*^{V437fs*}	54	62	33	< 0.00001	66	0.4239	Somatic Biallelic MutL-Deficient
13	*MLH1*^{R354fs*}	62	95	42	< 0.00001	84	NA	Somatic Biallelic MutL-Deficient
14	*MLH1*^{N38fs*}	37	47	23	< 0.00001	46	0.835	Somatic Biallelic MutL-Deficient
15	*MLH1*^{E89*}	74	90	43	< 0.00001	86	0.3786	Somatic Biallelic MutL-Deficient
16	*MLH1*^{E228*} + *MLH1*^{E648*}	36	NA	NA	NA	NA	NA	Somatic Biallelic MutL-Deficient

Table 5.7 – Characterisation of Somatic Biallelic MutS-Deficient & MutL-Deficient Colorectal Cancers: The details of the sixteen microsatellite unstable (MSI^+) DNA polymerase wild-type colorectal cancers with likely somatic biallelic mutations in the DNA mismatch repair (MMR) genes *MSH2, MSH6* or *MLH1*. Included are the somatic MMR mutation(s) harboured by the cancer, the estimated tumour purity, the observed number of reads associated with the mutation, the expected number of reads if the mutation was heterozygous, the p-value from the subsequent χ^2 analysis ($p_{(\chi^2)}$), the expected number of reads if the mutation was homozygouos and the associated ($p_{(\chi^2)}$). NA = Not Applicable.

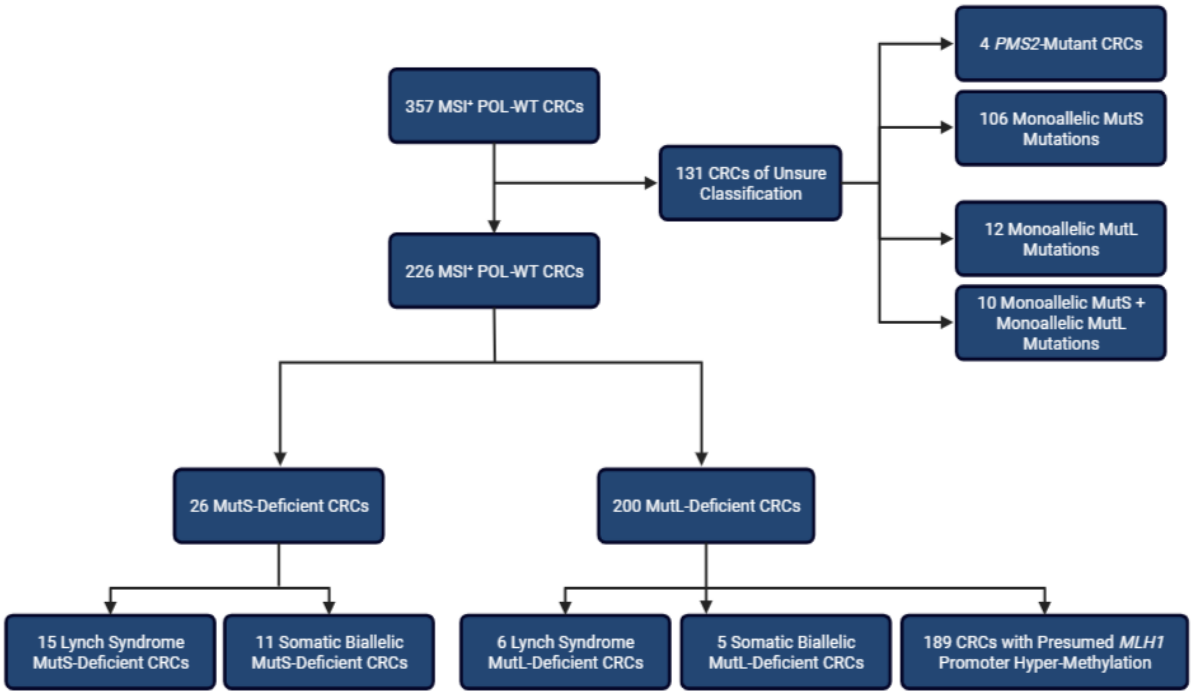
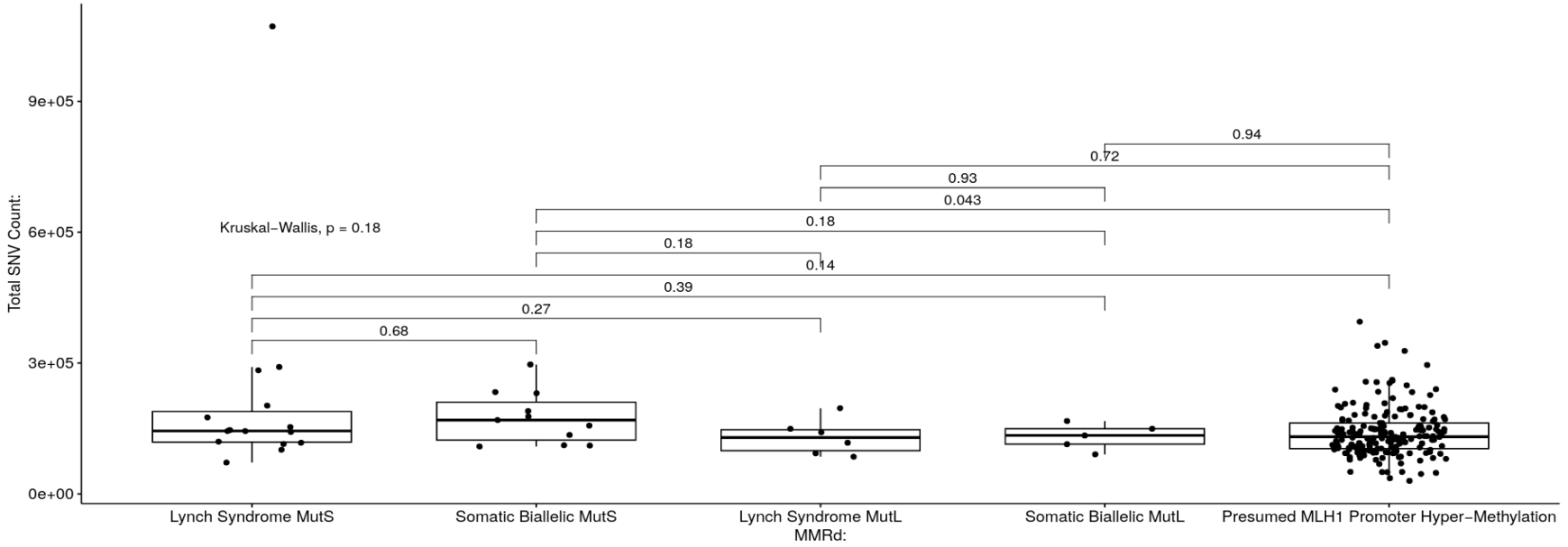


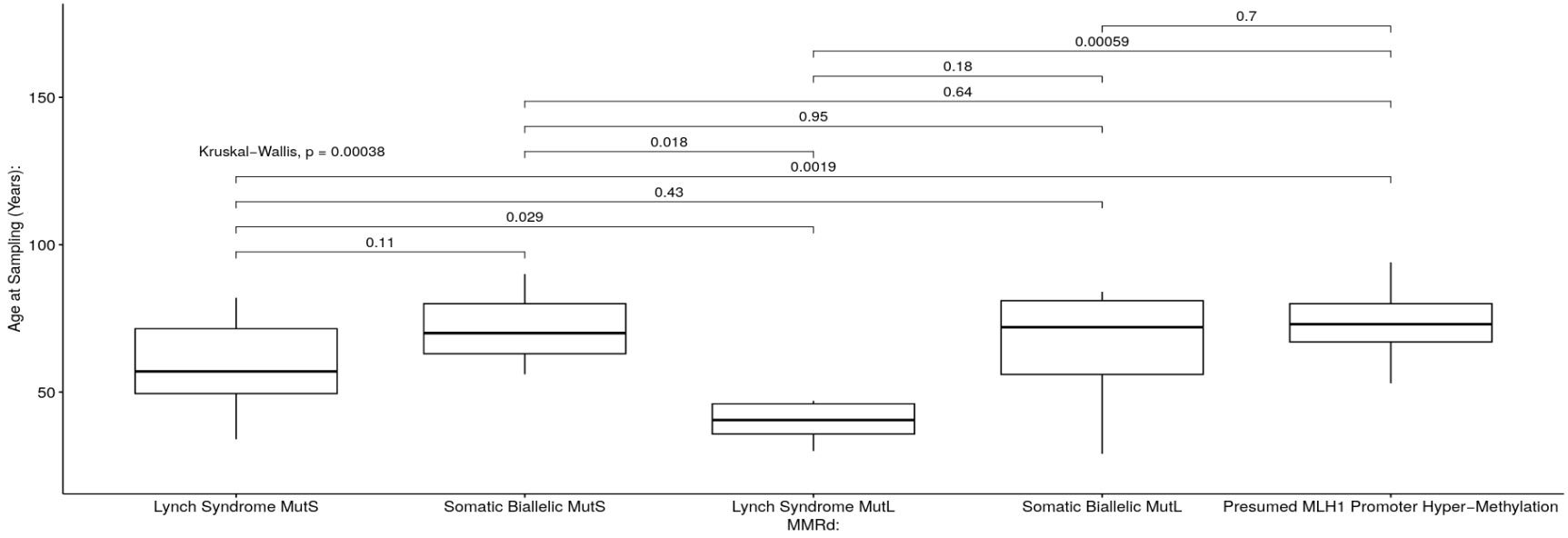
Figure 5.13 – Classification of Mismatch Repair Deficiencies of Colorectal Cancers with Microsatellite Instability: A flow chart detailing the classification of 357 microsatellite unstable (MSI^+) colorectal cancers (CRCs) with no mutations in the exonuclease domain of DNA polymerases δ or ε (POL-WT). Of the original 357 cancers, 131 were excluded from classification of mismatch repair deficiencies. Subsequently, of the 226 remaining cancers, 26 were characterised as MutS-deficient and 200 as MutL-deficient. Created with BioRender.com (https://app.biorender.com/).

younger than the cancers with presumed *MLH1* promoter hyper-methylation ($p = 0.0019$ and $p = 0.00059$ respectively, Figure 5.14b), while there were no significant differences in the ages of the somatic MutS-deficient cancers, somatic MutL-deficient CRCs and the cancers with presumed *MLH1* promoter hyper-methylation. Interestingly, both the LS-associated and somatic MutS-deficient CRCs presented with significantly more C → T mutations at CpG sites than any of the MutL-deficient cancers (see Figure 5.14c), while the LS-associated MutL-deficient CRCs, the somatic MutL-deficient CRCs and the CRCs with presumed *MLH1* promoter hyper-methylation presented with no difference in the number of C → T mutations at CpG sites (see Figure 5.14c). As seen in Figure 5.14d, C → T mutations at CpG sites represented a significantly greater proportion of the total SNV burden of LS-associated MutS-deficient CRCs than in any of the MutL-deficient groups of cancer, while in the somatic MutS-deficient CRCs this proportion was only significantly greater than the cancers with presumed *MLH1* promoter hyper-methylation ($p = 0.016$, Figure 5.14d). Similarly, to what was seen in Figure 5.14c, the proportion of the total SNV burden comprised of C → T mutations at CpG sites was not significantly different in the LS-associated MutL-deficient cancers, somatic MutL-deficient cancers or the cancers with presumed *MLH1* promoter hyper-methylation (see Figure 5.14d). Overall, this data indicates that MutS-deficient CRCs have a greater number of C → T mutations at CpG sites than their MutL-deficient counterparts that cannot be explained by patient age – which is in agreement with previous studies (518,519).

a)


Total SNV Count:
9e+05
6e+05
3e+05
0e+00
Kruskal–Wallis, p = 0.18
0.94
0.72
0.93
0.043
0.18
0.18
0.14
0.39
0.27
0.68
Lynch Syndrome MutS
Somatic Biallelic MutS
Lynch Syndrome MutL
Somatic Biallelic MutL
Presumed MLH1 Promoter Hyper–Methylation
MMRd:

b)


Age at Sampling (Years):
150
100
50
Kruskal–Wallis, p = 0.00038
0.7
0.00059
0.18
0.64
0.95
0.018
0.0019
0.43
0.029
0.11
Lynch Syndrome MutS
Somatic Biallelic MutS
Lynch Syndrome MutL
Somatic Biallelic MutL
Presumed MLH1 Promoter Hyper–Methylation
MMRd:

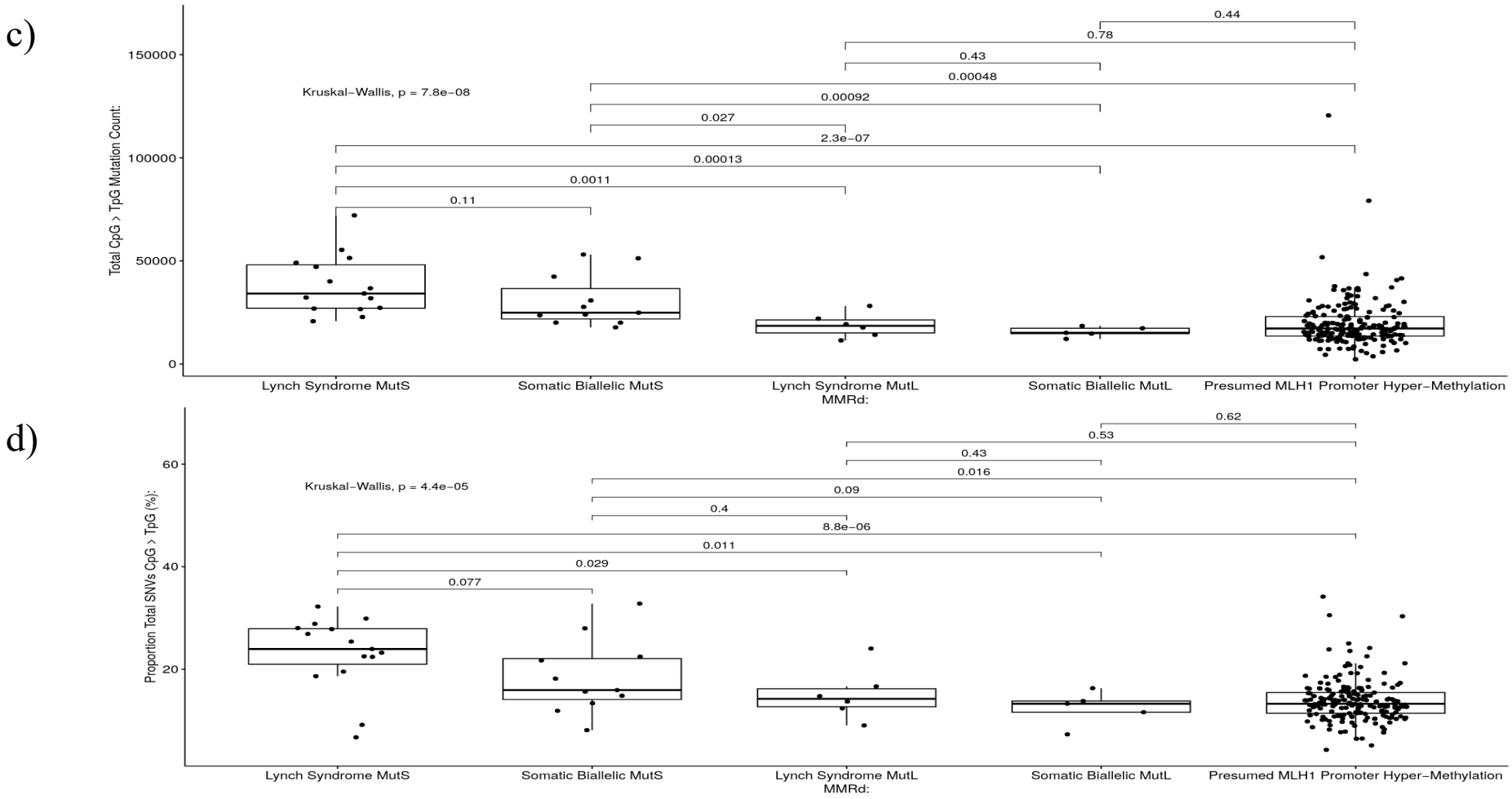
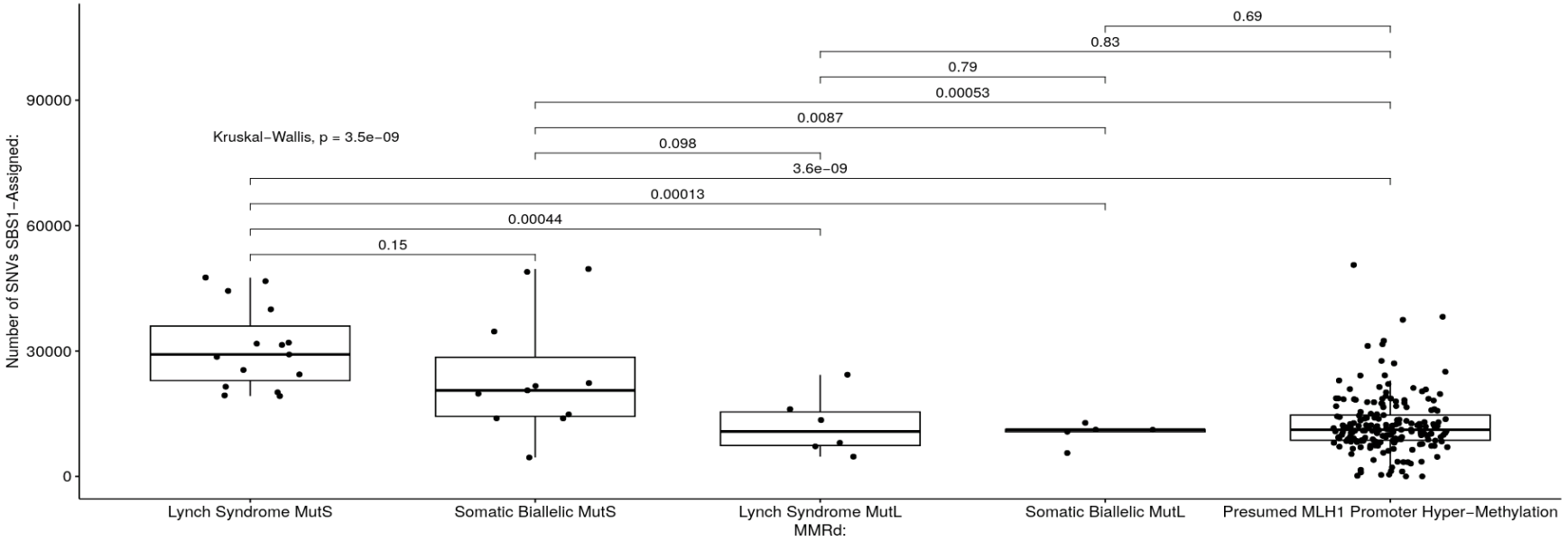


Figure 5.14 – The Mutation Profile of MutS-Deficient & MutL-Deficient MSI$^+$ Colorectal Cancers: Characteristics of microsatellite unstable DNA polymerase wild-type colorectal cancers extracted from the 100,000 Genomes Project. Cancers were classified as Lynch Syndrome MutS-deficient, somatic MutS-deficient, Lynch Syndrome MutL-deficient, somatic MutL-deficient and cancers with presumed *MLH1* promoter hyper-methylation. Included are the total single-nucleotide variation (SNV) burdens of each cancer (a), the ages of each patient (b), the number of C → T mutations at CpG sites in each cancer (c) and the proportion of the total SNV burden comprised of this mutation type (d).

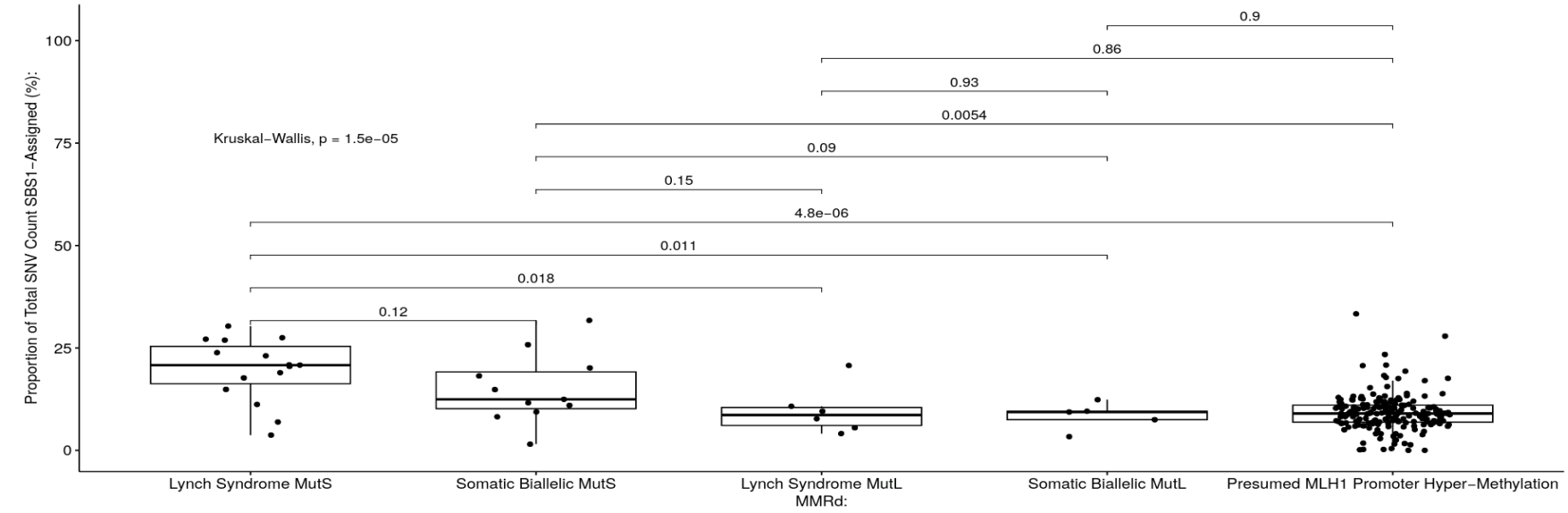
Following mutation signature extraction from these MSI^+ POL-WT CRCs, LS-assocated MutS-deficient cancers presented with a greater number of SNVs attributed to SBS1 than the LS-associated MutL-deficient cancers ($p = 0.00044$, Figure 5.15a), somatic MutL-deficient cancers ($p = 0.00013$, Figure 5.15a) and CRCs with presumed *MLH1* promoter hyper-methylation ($p = 3.6 \times 10^{-9}$, Figure 5.15a). Somatic MutS-deficient CRCs had a greater number of SNVs attributed to SBS1 than all of the MutL-deficient groups, but this difference was only significant compared to the somatic MutL-deficient cancers and the cancers with presumed *MLH1* promoter hyper-methylation (Figure 5.15a). There were no significant differences in the number of SNVs attributed to SBS1 in any of the MutL-deficient groups (Figure 5.15a). Both the LS-associated and somatic MutS-deficient had a significantly greater proportion of SNVs attributed to SBS1 than the cancers with presumed *MLH1* promoter hyper-methylation ($p = 4.8 \times 10^{-6}$ and $p = 0.0054$ respectively, Figure 5.15b), while there were no significant differences in this proportion between any of the MutL-deficient groups (Figure 5.15b). Interestingly, there were no differences in the number ($p_{(Kruskal\text{-}Wallis)} = 0.92$, Figure 5.15c) or proportion ($p_{(Kruskal\text{-}Wallis)} = 0.91$, Figure 5.15d) of SNVs attributed to SBS15, although this may have been a consequence of the low number of cancers that presented with this signature. Overall, this data suggests that the increase in C → T mutagenesis at CpG sites in MutS-deficient CRCs (see Figure 5.14) correlates with an increase in the prevalence of the mutation signature SBS1, indicating that the suggestions of previous studies that these cancers accumulate C → T mutations at CpG sites via unrepaired spontaneous deaminations of 5-mC may be accurate (518,519).

Next, C → T mutations at CpG sites in these MSI^+ POL-WT CRCs were binned according to their DNA methylation status in the normal sigmoid colon (see section 4.2.5 for details). As seen in Figure 5.16, there were significant positive correlations between DNA methylation and the rate of C → T mutagenesis at CpG sites in LS-assoiated MutS-deficient CRCs ($r^2 = 0.9159$, $p < 0.00001$), somatic MutS-deficient CRCs ($r^2 = 0.8332$, $p = 0.000034$), LS-associated MutL-deficient CRCs ($r^2 = 0.9267$, $p < 0.00001$), somatic MutL-deficient CRCs ($r^2 = 0.8105$, $p = 0.000065$) and cancers with presumed *MLH1* promoter hyper-methylation ($r^2 = 0.6599$, $p = 0.001326$). When the regression slopes of the relationship between DNA methylation and the rate of C → T mutagenesis at CpG sites were compared, the slope of the LS-associated MutS-deficient cancers ($\alpha = 1{,}244$) was significantly steeper than the slope of the LS-associated MutL-deficient CRCs ($\alpha = 651.2$, $p < 0.0001$), somatic MutL-deficient CRCs ($\alpha = 425.5$, $p < 0.0001$) and cancers with presumed *MLH1* promoter hyper-methylation ($\alpha = 313.4$, $p < 0.0001$). Interestingly, the regression constant of LS-associated MutS-deficient CRCs was not significiantly different when compared to any of the MutL-deficient groups – suggesting that the rate of C → T mutagenesis at CpG sites in these cancers was the same when DNA methylation was zero. The regression slope of the somatic MutS-deficient CRCs ($\alpha = 726$) was significantly steeper than the slopes of the somatic MutL-deficient CRCs ($p = 0.022$) and CRCs with presumed *MLH1* promoter hyper-methylation ($p = 0.004$). However, the regression constants of the somatic MutS-deficient CRCs was not significantly different to either of these groups of MutL-deficient cancer. Overall, this data suggests that highly-methylated CpG sites are most at risk of C → T mutagenesis in these MSI^+ cancers, with MutS-deficient cancers showing steeper regression slopes than MutL-deficient cancers, perhaps indicating that the association between DNA methylation and the rate of C → T mutagenesis at CpG sites was stronger in these cancers.

a)


Number of SNVs SBS1−Assigned:
90000
60000
30000
0
Kruskal−Wallis, p = 3.5e−09
0.69
0.83
0.79
0.00053
0.0087
0.098
3.6e−09
0.00013
0.00044
0.15
Lynch Syndrome MutS
Somatic Biallelic MutS
Lynch Syndrome MutL
Somatic Biallelic MutL
Presumed MLH1 Promoter Hyper−Methylation
MMRd:

b)


Proportion of Total SNV Count SBS1−Assigned (%):
100
75
50
25
0
Kruskal−Wallis, p = 1.5e−05
0.9
0.86
0.93
0.0054
0.09
0.15
4.8e−06
0.011
0.018
0.12
Lynch Syndrome MutS
Somatic Biallelic MutS
Lynch Syndrome MutL
Somatic Biallelic MutL
Presumed MLH1 Promoter Hyper−Methylation
MMRd:

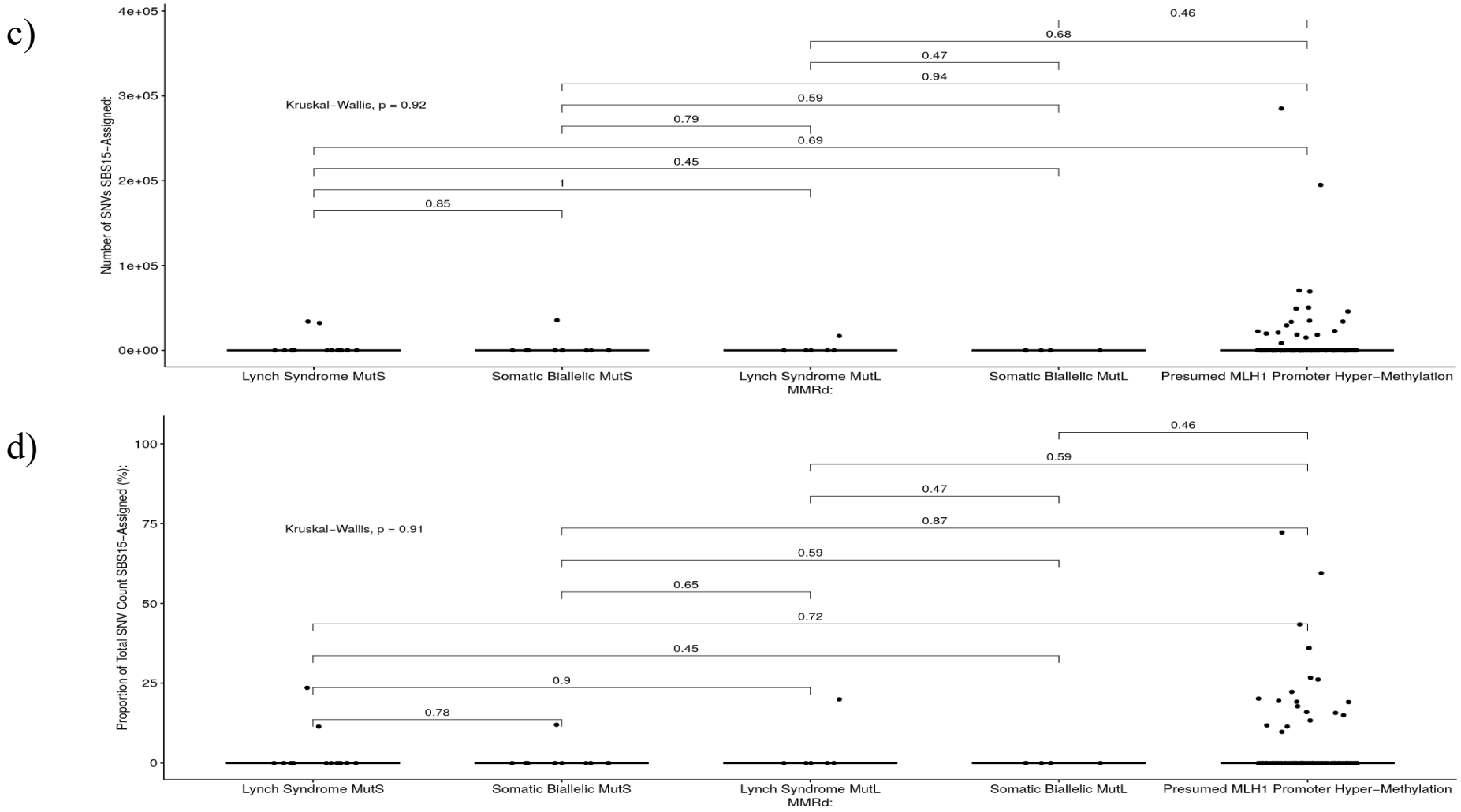


Figure 5.15 – Mutation Signature Distribution of MutS-Deficient & MutL-Deficient MSI$^+$ Colorectal Cancers: The number and proportion of single-nucleotide variations (SNVs) attributed to specific mutation signatures in microsatellite unstable DNA polymerase wild-type colorectal cancers. Cancers were classified as Lynch Syndrome MutS-deficient, somatic MutS-deficient, Lynch Syndrome MutL-deficient, somatic MutL-deficient and cancers with presumed *MLH1* promoter hyper-methylation. Included are the number (a) and proportion (b) of SNVs attributed to SBS1 and the number (c) and proportion (d) of SNVs attributed to SBS15.

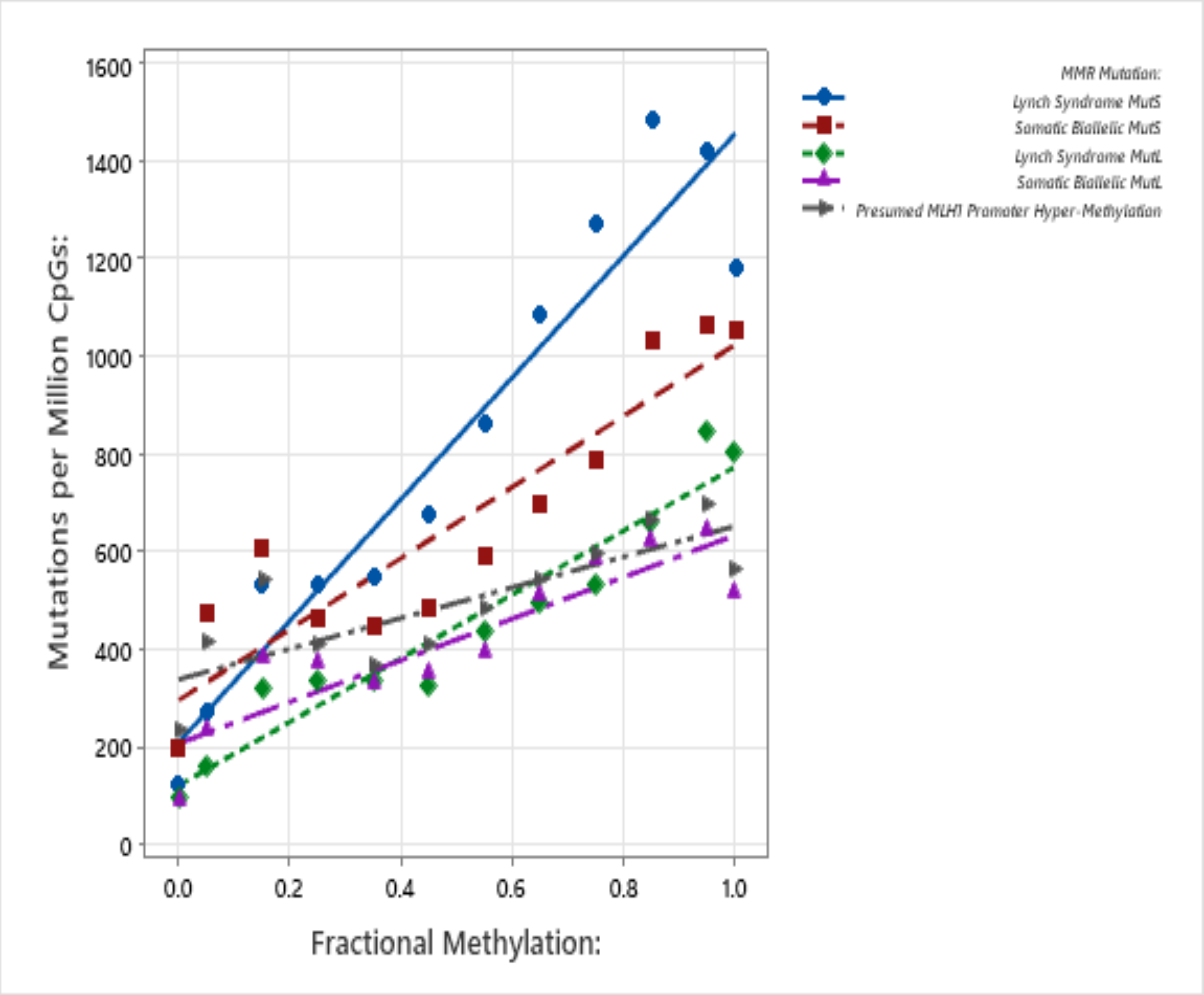


Figure 5.16 – The Effect of DNA Methylation on C → T Mutagenesis at CpG Sites in MutS-Deficient & MutL-Deficient MSI$^+$ Colorectal Cancers: The association between the C → T mutation rate at CpG sites and DNA methylation in whole-genome sequencing data obtained from microsatellite unstable colorectal cancers of the 100,000 Genomes Project. Cancers were classified as Lynch Syndrome MutS-deficient, somatic MutS-deficient, Lynch Syndrome MutL-deficient, somatic MutL-deficient and cancers with presumed *MLH1* promoter hyper-methylation.

Interestingly, the regression slope of the CRCs with presumed *MLH1* promoter hyper-methylation was significantly less steep than LS-associated MutL-deficient CRCs ($p = 0.001$), while the regression constant for these cancers was significantly greater than LS-associated MutL-deficient CRCs ($p = 0.001$) and somatic MutL-deficient cancers ($p = 0.034$), indicating that at the point that methylation was zero, the rate of C → T mutagenesis was higher in the cancers with presumed *MLH1* promoter hyper-methylation.

Following this, C → T mutations at CpG sites in each of these groups of MSI$^+$ POL-WT CRCs were binned according to their replication timing. LS-associated MutS-deficient CRCs presented with significant correlations between DNA methylation and the rate of C → T mutagenesis at CpG sites in all four replication timing bins (Figure 5.17a and Table 5.8). The same positive corrlations were observed in somatic MutS-deficient cancers (Figure 5.17b, Table 5.8), LS-associated MutL-deficient CRCs (Figure 5.17c, Table 5.8) and somatic MutL-deficient CRCs (Figure 5.17d, Table 5.8). In the CRCs with presumed *MLH1* promoter hyper-methylation (Figure 5.17e), there were significant correlations between DNA

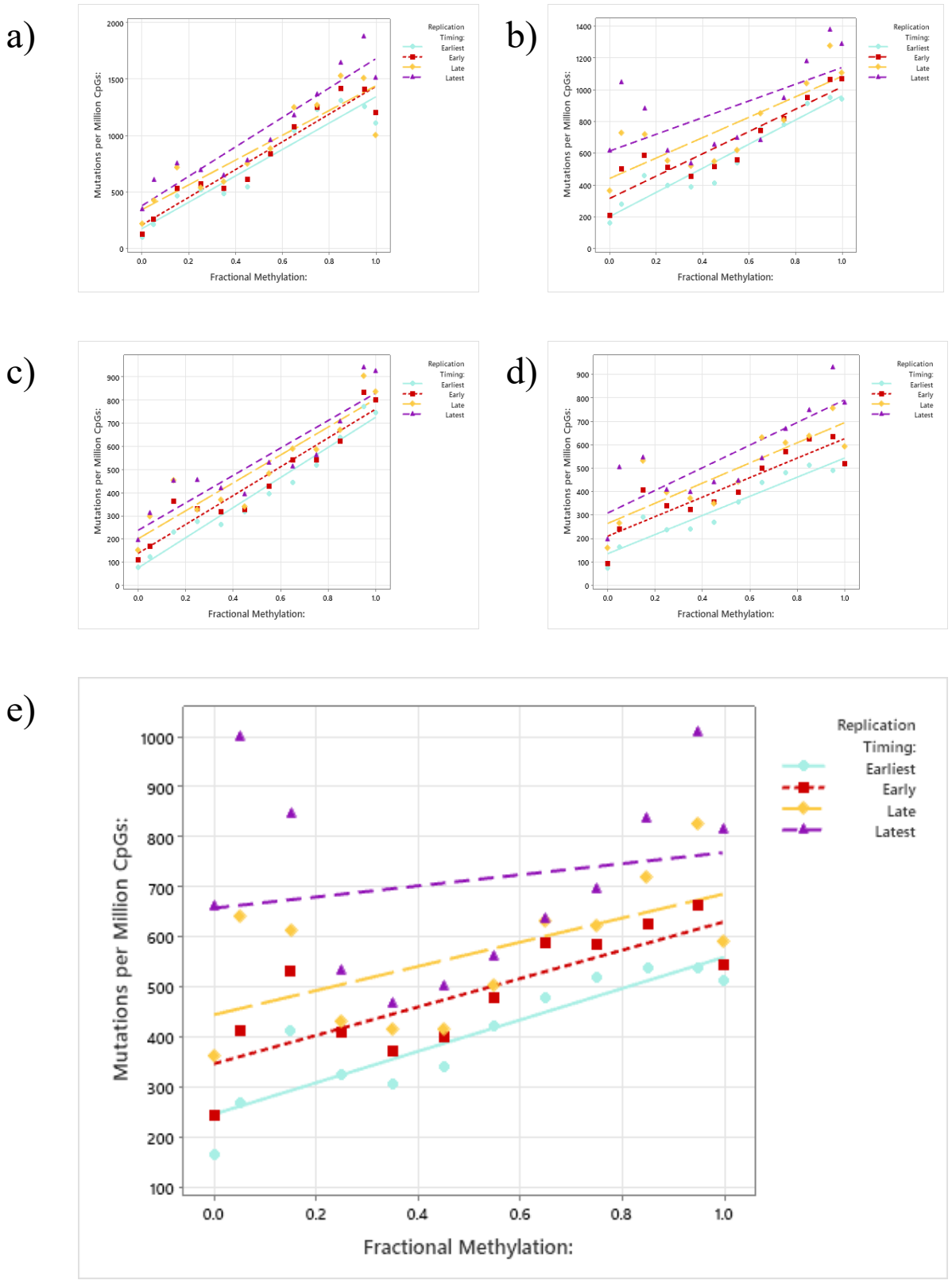


Figure 5.17 – The Effect of Replication Timing on C → T Mutagenesis at CpG Sites in MutS-Deficient & MutL-Deficient MSI$^+$ Colorectal Cancers: The association between DNA methylation and the rate of C → T mutagenesis at CpG sites in the earliest (blue), early (red), late (yellow) and latest (purple) replicating regions of the genome in microsatellite unstable DNA polymerase wild-type colorectal cancers. Cancers were classified as Lynch Syndrome MutS-deficient (a), somatic MutS-deficient (b), Lynch Syndrome MutL-deficient (c), somatic MutL-deficient (d) and cancers with presumed *MLH1* promoter hyper-methylation (e).

Lynch Syndrome MutS-Deficient (n = 15):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	1,165	176.7	0.9122	0.9551	< 0.00001
Early	1,225	206.1	0.9223	0.9604	< 0.00001
Late	1,095	342.2	0.7922	0.8901	0.000105
Latest	1,300	377.2	0.8908	0.9438	< 0.00001
Somatic MutS-Deficient (n = 11):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	760	202.3	0.9235	0.961	< 0.00001
Early	702	315	0.8454	0.9195	0.000023
Late	643	441.1	0.678	0.8234	0.000997
Latest	526	614.4	0.3984	0.6312	0.027725
Lynch Syndrome MutL-Deficient (n = 6):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	650.4	75.7	0.9625	0.9811	< 0.00001
Early	623.2	137.4	0.9113	0.9546	< 0.00001
Late	603.2	200.4	0.8579	0.9262	0.000015
Latest	592.7	237	0.8302	0.9112	0.000037
Somatic MutL-Deficient (n = 5):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	408.2	135.4	0.907	0.9524	< 0.00001
Early	416.6	209.3	0.8061	0.8978	0.000074
Late	430.7	263.6	0.7166	0.8465	0.000516
Latest	483.6	308.1	0.6967	0.8347	0.000731
Presumed *MLH1* Promoter Hyper-Methylation (n =189):					
Replication Timing Bin:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Earliest	313	246.7	0.809	0.8994	0.000068
Early	283	346.6	0.6404	0.8002	0.001774
Late	241	444.8	0.3554	0.5962	0.040754
Latest	111	658.4	0.0424	0.2059	0.520864

Table 5.8 – Replication Timing Regression Analysis of MutS-Deficient & MutL-Deficient MSI$^+$ Colorectal Cancer: The regression equations of the relationship between DNA methylation and C → T mutation rates at CpG sites in microsatellite unstable DNA polymerase wild-type colorectal cancers. Cancers were classified as Lynch Syndrome MutS-deficient, somatic MutS-deficient, Lynch Syndrome MutL-deficient, somatic MutL-deficient and cancers with presumed *MLH1* promoter hyper-methylation. Included are the replication timing bin, the regression slope, the regression constant, the methylation-mutation rate r^2 correlation, the Pearson's R correlation measure of this relationship and the p-value associated with the Pearson's R statistic ($p_{(Pearson's\ R)}$).

methylation and the rate of C → T mutagenesis at CpG sites in three of the four replication timing bins, whereas this correlation was non-significant in the latest replication timing bin (r^2 = 0.0424, p = 0.520864 – Table 5.8), possibly as a consequence of the unusually high rate of C → T mutagenesis of CpG sites with a DNA methylation of between zero and thirty percent. As seen in Table 5.9, the DNA replication timing bin had no effect on the regression slope between DNA methylation and the rate of C → T mutagenesis at CpG sites in LS-associated MutS-deficient CRCs (Methylation x Replication Timing = 27.4, p = 0.654), somatic MutS-deficient CRCs (Methylation x Replication Timing = -76.2, p = 0.2), LS-associated MutL-deficient CRCs (Methylation x Replication Timing = -19.3, p = 0.512), somatic MutL-deficient CRCs (Methylation x Replication Timing = 24, p = 0.466) and CRCs with presumed *MLH1* promoter hyper-methylation (Methylation x Replication Timing = -65, p = 0.165). Overall, this data indicates that, in these cancers, replication timing has no effect on the rate of C → T mutagenesis at CpG sites – consistant with what has been reported for MSI^+ cancers throughout this thesis.

While it is apparent that MutS-deficient MSI^+ CRCs present with more C → T mutations at CpG sites, a greater representation of SBS1 and a steeper regression slope between DNA methylation and the rate of C → T mutagenesis at CpG sites than their MutL-deficient counterparts, the potential mechanism(s) underlying this has not yet been investigated. Should the theory presented by Fang *et al.* (that MutS-deficient cancers accumulate C → T mutations at CpG sites via unrepaired spontaneous deaminations of 5-mC while MutL-deficient cancers accumulate mutations via unrepaired DNA replication errors) be accurate, the $\log_2$(Leading/Lagging) ratio of mutations in the MutS-deficient cancers should be closer to zero and resemble the *MBD4*-mutant colorectal polyps compared to the MutL-deficient CRCs, which should have a $\log_2$(Leading/Lagging) ratio of greater than zero. Surprisingly, there were no differences in the $\log_2$(Leading/Lagging) ratios between any of the MutS-deficient or MutL-deficient cancers ($p_{(Kruskal\text{-}Wallis)}$ = 0.27, Figure 5.18) – possibly indicating that these cancers accumulate C → T mutations at CpG sites via the same mechanism. As seen in Figure 5.18, the $\log_2$(Leading/Lagging) ratio of these MSI^+ cancers was greater than zero, indicating that there was an excess of C → T mutations at CpG sites on the template for leading strand synthesis – suggesting that these mutations may be a consequence of unrepaired DNA replication errors and not spontaneous deaminations of 5-mC. This goes against what has been hypothesised in section 5.1.4 and the data presented previously by Fang *et al.,* who suggested that MutS is involved in the repair of spontaneous deaminations of 5-mC (518). The suggestion that MSI^+ CRCs accumulate C → T mutations at CpG sites via unrepaired DNA replication errors and the previous data suggesting that highly-methylated CpG sites are at greatest risk of mutation in both these cancers and CRCs with *POL-ε* EDMs (see Figure 5.7) implies that DNA methylation may impact the likelihood of a replication error occurring at a CpG site. Therefore, the next stage of analysis should be to combine the replication strand and fractional DNA methylation data used previously to investigate this hypothesis.

5.3.3 – The Relationship Between DNA Methylation & *POL-ε* Replication Errors

Lynch Syndrome MutS-Deficient (n = 15):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	1,155	< 0.0001
Replication Timing	73.8	0.049
Methylation x Replication Timing Interaction Term	27.4	0.654
Constant	164.9	0.02
Somatic MutS-Deficient (n = 11):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	772	< 0.0001
Replication Timing	136.2	< 0.0001
Methylation x Replication Timing Interaction Term	-76.2	0.2
Constant	188.9	0.006
Lynch Syndrome MutL-Deficient (n = 6):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	646.4	< 0.0001
Replication Timing	54.7	0.003
Methylation x Replication Timing Interaction Term	-19.3	0.512
Constant	80.6	0.018
Somatic MutL-Deficient (n = 5):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	398.7	< 0.0001
Replication Timing	57.2	0.006
Methylation x Replication Timing Interaction Term	24	0.466
Constant	143.2	< 0.0001
Presumed *MLH1* Promoter Hyper-Methylation (n = 189):		
Regression Variable:	**Coefficient:**	**$P_{(Coefficient)}$:**
DNA Methylation	334.8	< 0.0001
Replication Timing	133.3	< 0.0001
Methylation x Replication Timing Interaction Term	-65	0.165
Constant	224.2	< 0.0001

Table 5.9 – The Effect of Replication Timing on Methylation-Mutation Rate Associations in MutS-Deficient & MutL-Deficient MSI^+ Colorectal Cancer: Regression equations for the effect of fractional DNA methylation and replication timing on C → T mutation rates at CpG sites in microsatellite unstable DNA polymerase wild-type colorectal cancers. Cancers were classified as Lynch Syndrome MutS-deficient, somatic MutS-deficient, Lynch Syndrome MutL-deficient, somatic MutL-deficient and cancers with presumed *MLH1* promoter hyper-methylation. Included are the variables of the equation – DNA methylation, replication timing, the interaction term between the two, the regression constant and the p-value of each ($p_{(Coefficient)}$).

The results presented in the previous section are contrary to the hypothesis set out in section 5.1.4, where it was hypothesised that MutS-deficient CRCs would present with no asymmetry of mutations on the template for either the leading or lagging strand – similar to the *MBD4*-mutant colorectal polyps. Instead, an excess of mutations on the leading strand template was seen, indicating that C → T mutagenesis at CpG sites in these cancers may be a consequence of unrepaired C:A *POL-ε* replication errors that propagate into mutations in the next round of DNA replication and not via unrepaired spontaneous deaminations of 5-mC. In Chapter IV of this thesis, it was shown that more highly-methylated CpG sites were at greater risk of spontaneous deamination. In addition to this, there was a significant positive correlation between DNA methylation and the rate of C → T mutagenesis at CpG sites in both *POL-ε* mutant and MSI^+ POL-WT CRCs, potentially indicating that highly-methylated CpG sites may also be at greater risk of *POL-ε* replication error producing a C:A DNA mismatch. In order to investigate this, C → T mutations at CpG sites of either the leading or lagging strand

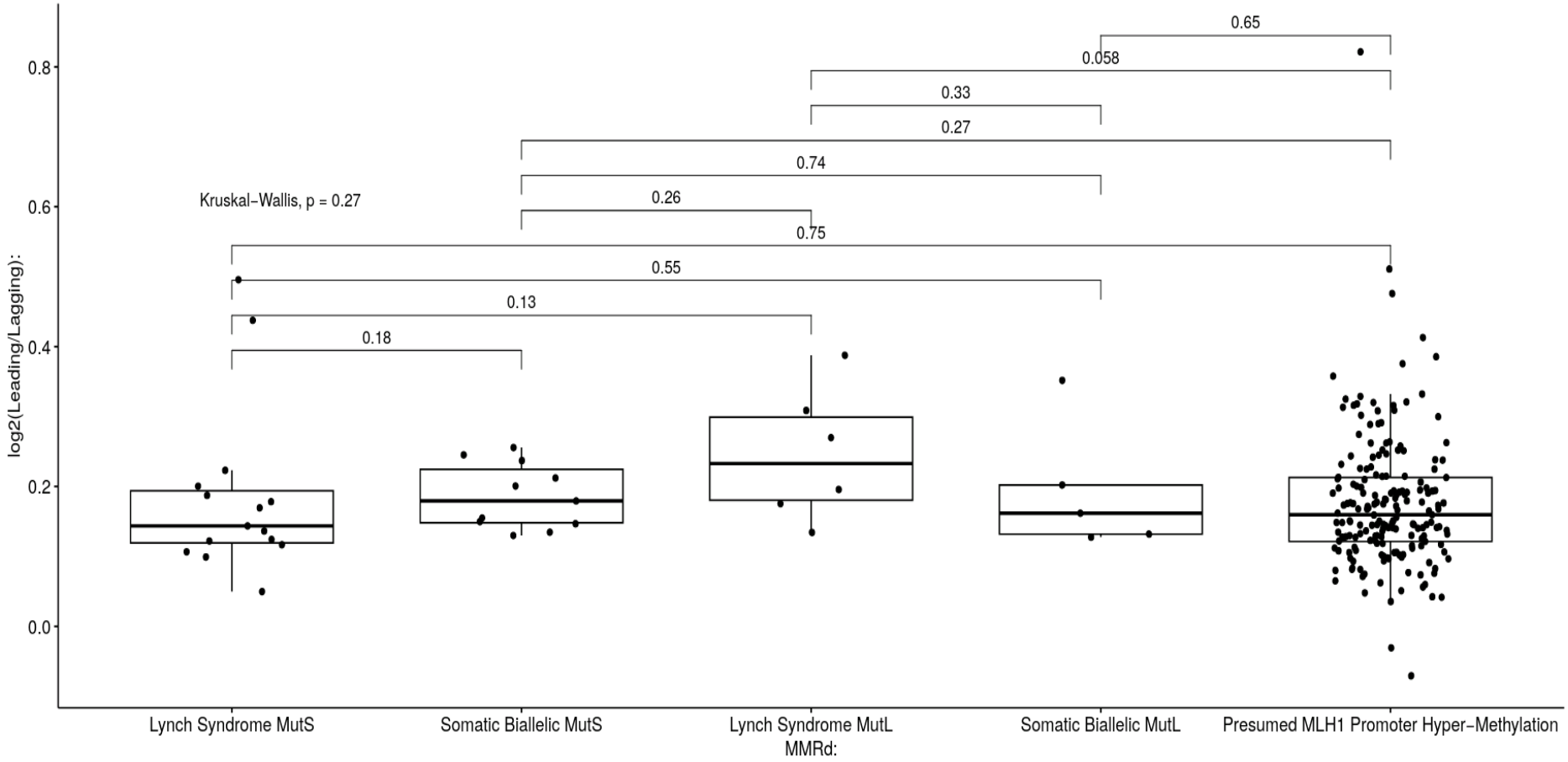


Figure 5.18 – Replication Strand Analysis of MutS-Deficient & MutL-Deficient MSI$^+$ Colorectal Cancers: The $\log_2$(Leading/Lagging) ratio of C → T mutations at CpG sites in microsatellite unstable DNA polymerase wild-type colorectal cancers. Cancers were classified as Lynch Syndrome MutS-deficient, somatic MutS-deficient, Lynch Syndrome MutL-deficient, somatic MutL-deficient and cancers with presumed *MLH1* promoter hyper-methylation

template identified previously were binned according to their fractional DNA methylation in the normal sigmoid colon.

In cancers where C → T mutagenesis at CpG sites was assumed to be a result of unrepaired spontaneous deaminations, there was no difference in the association between DNA methylation and the C → T mutation rate at CpG sites on the leading or lagging strand templates. In the *MBD4*-mutant colorectal polyps (Figure 5.19a), there were significant positive correlations between DNA methylation and the rate of C → T mutagenesis at CpG sites on both the leading (r^2 = 0.9297, p < 0.00001 – Table 5.10) and lagging (r^2 = 0.9172, p < 0.00001 – Table 5.10) strand templates. However the regression slopes of the leading (α = 608.9) and lagging (α = 599.3) strand templates were not significantly different (p = 0.894). In MSS POL-WT CRCs (Figure 5.19b), there were also significant correlations between DNA methylation and the C → T mutation rate at CpG sites on both the leading (r^2 = 0.7523, p = 0.000257 – Table 5.10) and lagging (r^2 = 0.7479, p = 0.000282 – Table 5.10) replication strand templates. Similarly to the *MBD4*-mutant colorectal polyps, the regression slopes of the leading (α = 42.35) and lagging (α = 42.62) replication strand templates were not significantly different (p = 0.98). Conversely, in cancers with *POL-ε* EDMs – which were thought to accumulate C → T mutations at CpG sites via unrepaired C:A mismatches caused by *POL-ε* replication error – the association between DNA methylation and the rate of C → T mutagenesis at CpG sites was stronger on the leading strand template. As seen in Figure 5.19c, there were significant correlations between DNA methylation and the rate of C → T mutagenesis at CpG sites on both the leading (r^2 = 0.6842, p = 0.000901 – Table 5.10) and lagging (r^2 = 0.6911, p = 0.000805 – Table 5.10) replication strand templates in *POL-ε* mutant CRCs. However, the regression slope of the leading strand template (α = 3,079) was significantly higher than the slope of the lagging strand template (α = 1,492, p = 0.031), suggesting that more highly-methylated CpG sites were at greater risk of C:A DNA mismatches caused by *POL-ε* replication errors.

When this analysis was performed on MSI^+ POL-WT CRCs (Figure 5.20), significant positive correlations between DNA methylation and the rate of C → T mutagenesis at CpG sites were observed on both the leading and lagging strand templates (Table 5.10). In LS-associated MutS-deficient CRCs (Figure 5.20a), the regression slope of the leading strand (α = 689.4) was steeper than that of the lagging strand template (α = 643), but this difference was not significant (p = 0.599). In somatic MutS-deficient CRCs (Figure 5.20b), the regression slope of the leading strand (α = 423.2) was again steeper than the slope of the lagging strand (α = 313.8), but this difference was non-significant (p = 0.265). LS-associated MutL-deficient CRCs (Figure 5.20c) had a similarly steeper slope on the leading strand (α = 340.1) than lagging strand (α = 309.7), but this difference was again non-significant (p = 0.594). CRCs with somatic MutL mutations (Figure 5.20d) presented with a non-significant difference in the slopes of the leading strand (α = 232.9) and lagging strand (α = 223.9, p = 0.862). Finally, CRCs with presumed *MLH1* promoter hyper-methylation (Figure 5.20e) showed a similar steeper slope on the leading strand (α = 149.2) than the lagging strand (α = 138.4) – but this difference was not significant (p = 0.862).

Overall, this data suggests that highly-methylated CpG sites were more at risk of the erroneous mis-incorporation of adenine opposite a template cytosine in MSS *POL-ε* mutant CRCs, suggesting that increasing DNA methylation consequently increases the likelihood of

***MBD4*-Mutant Colorectal Polyps (n = 4):**					
Replication Strand:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Leading Strand Template	608.9	-5.9	0.9297	0.9642	< 0.00001
Lagging Strand Template	599.3	-5.4	0.9172	0.9577	< 0.00001
MSS POL-WT CRC (n = 1,532):					
Replication Strand:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Leading Strand Template	42.35	15.52	0.7523	0.8674	0.000257
Lagging Strand Template	42.62	15.56	0.7479	0.8648	0.000282
MSS *POL-ε* Mutant CRC (n = 16):					
Replication Strand:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Leading Strand Template	3,079	197	0.6842	0.8272	0.000901
Lagging Strand Template	1,492	132	0.6911	0.8313	0.000805
MSI^+ Lynch Syndrome MutS-Deficient (n = 15):					
Replication Strand:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Leading Strand Template	689.4	128.2	0.922	0.9602	< 0.00001
Lagging Strand Template	643	102.1	0.9214	0.9599	< 0.00001
MSI^+ Somatic MutS-Deficient (n = 11):					
Replication Strand:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Leading Strand Template	423.2	176.8	0.7531	0.8678	0.000254
Lagging Strand Template	313.8	171.7	0.7538	0.8682	0.00025
MSI^+ Lynch Syndrome MutL-Deficient (n = 6):					
Replication Strand:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Leading Strand Template	340.1	84.4	0.8313	0.9118	0.000036
Lagging Strand Template	309.7	62.5	0.921	0.9597	< 0.00001
MSI^+ Somatic MutL-Deficient (n = 5):					
Replication Strand:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Leading Strand Template	232.9	125.3	0.7822	0.8844	0.000133
Lagging Strand Template	223.9	102.6	0.8175	0.9042	0.000054
MSI^+ Presumed *MLH1* Promoter Hyper-Methylation (n = 189):					
Replication Strand:	**Regression Slope:**	**Regression Constant:**	**r^2:**	**Pearson's R:**	**$P_{(Pearson's\ R)}$:**
Leading Strand Template	149.2	205.8	0.5126	0.716	0.008819
Lagging Strand Template	138.4	183.8	0.5302	0.7281	0.007255

Table 5.10 – Regression Analysis of Replication Strand in *MBD4*-Mutant Colorectal Polyps & Colorectal Cancer: The regression equations of the relationship between DNA methylation and C → T mutation rates at CpG sites in colorectal polyps derived from a patient with a germline biallelic *MBD4* truncation, microsatellite stable (MSS) and microsatellite unstable (MSI^+) DNA polymerase wild-type (POL-WT) colorectal cancers and colorectal cancers with pathogenic DNA polymerase-ε (*POL-ε*) exonuclease domain mutations. MSI^+ colorectal cancers were classified as Lynch Syndrome MutS-deficient, somatic MutS-deficient, Lynch Syndrome MutL-deficient, somatic MutL-deficient and cancers with presumed *MLH1* promoter hyper-methylation Included are the DNA replication strand, the regression slope, the regression constant, the methylation-mutation rate r^2 correlation, the Pearson's R correlation measure of this relationship and the p-value associated with the Pearson's R statistic ($p_{(Pearson's\ R)}$).

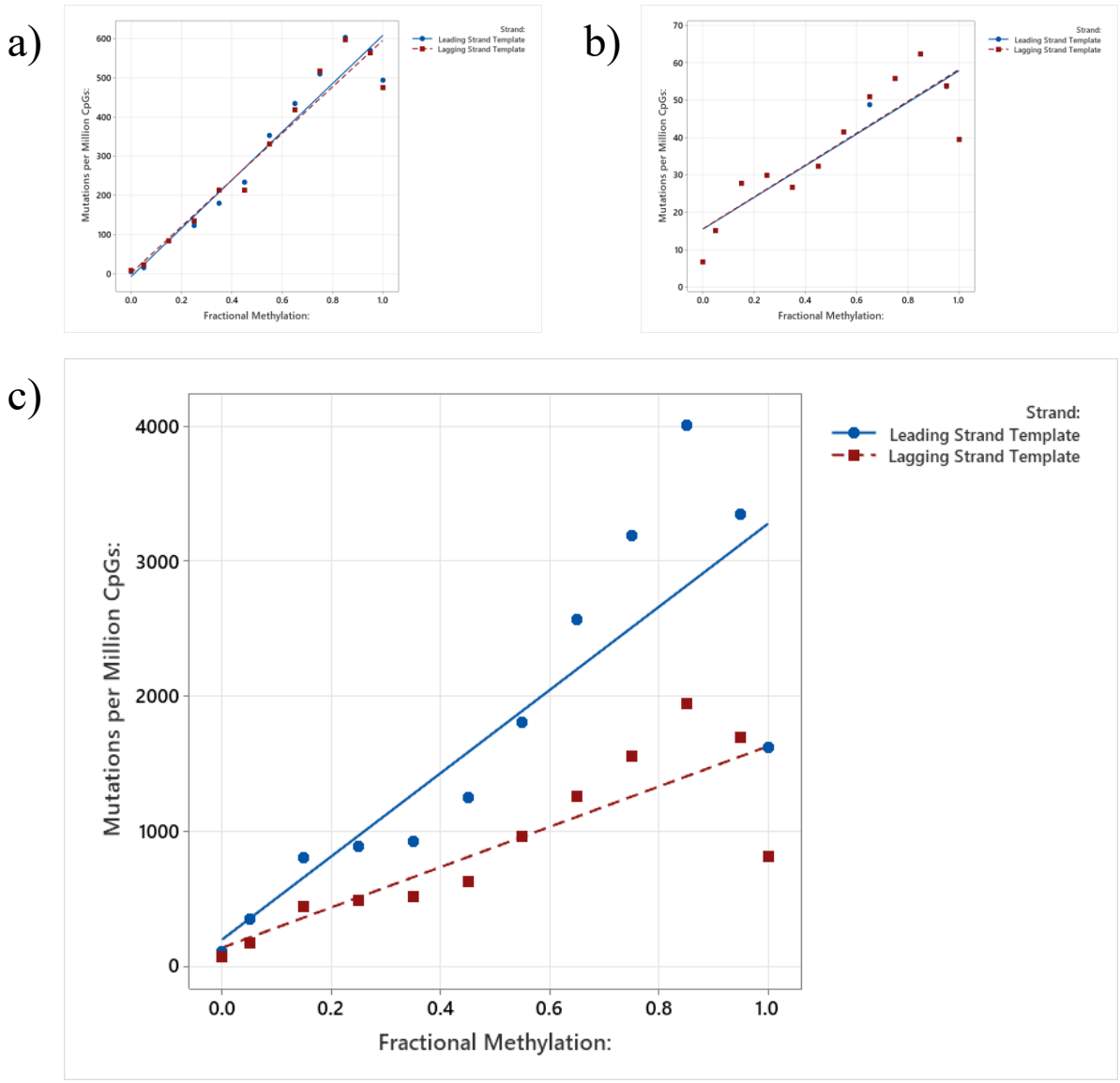


Figure 5.19 – Replication Strand-Specific Associations Between DNA Methylation & C → T Mutagenesis at CpG Sites of *MBD4*-Mutant Colorectal Polyps & Colorectal Cancers: The relationship between fractional DNA methylation from the normal sigmoid colon and the rate of C → T mutagenesis at CpG sites on either the leading (blue) or lagging (red) template strand for DNA replication. Included are *MBD4*-mutant colorectal polyps (a), microsatellite stable (MSS) DNA polymerase (POL) wild-type colorectal cancers (CRCs) (b) and MSS CRCs with pathogenic *POL-ε* exonuclease domain mutations (c).

DNA replication erorrs occurring. The same may be true in MSI^+ POL-WT cancers, which presented with a $\log_2$(Leading/Lagging) ratio greater than zero and a consistently steeper (albeit non-significant) regression slope on the leading strand compared to the lagging strand.

5.3.4 – Further Characterisation of CRCs with Presumed *MLH1* Promoter Hyper-Methylation

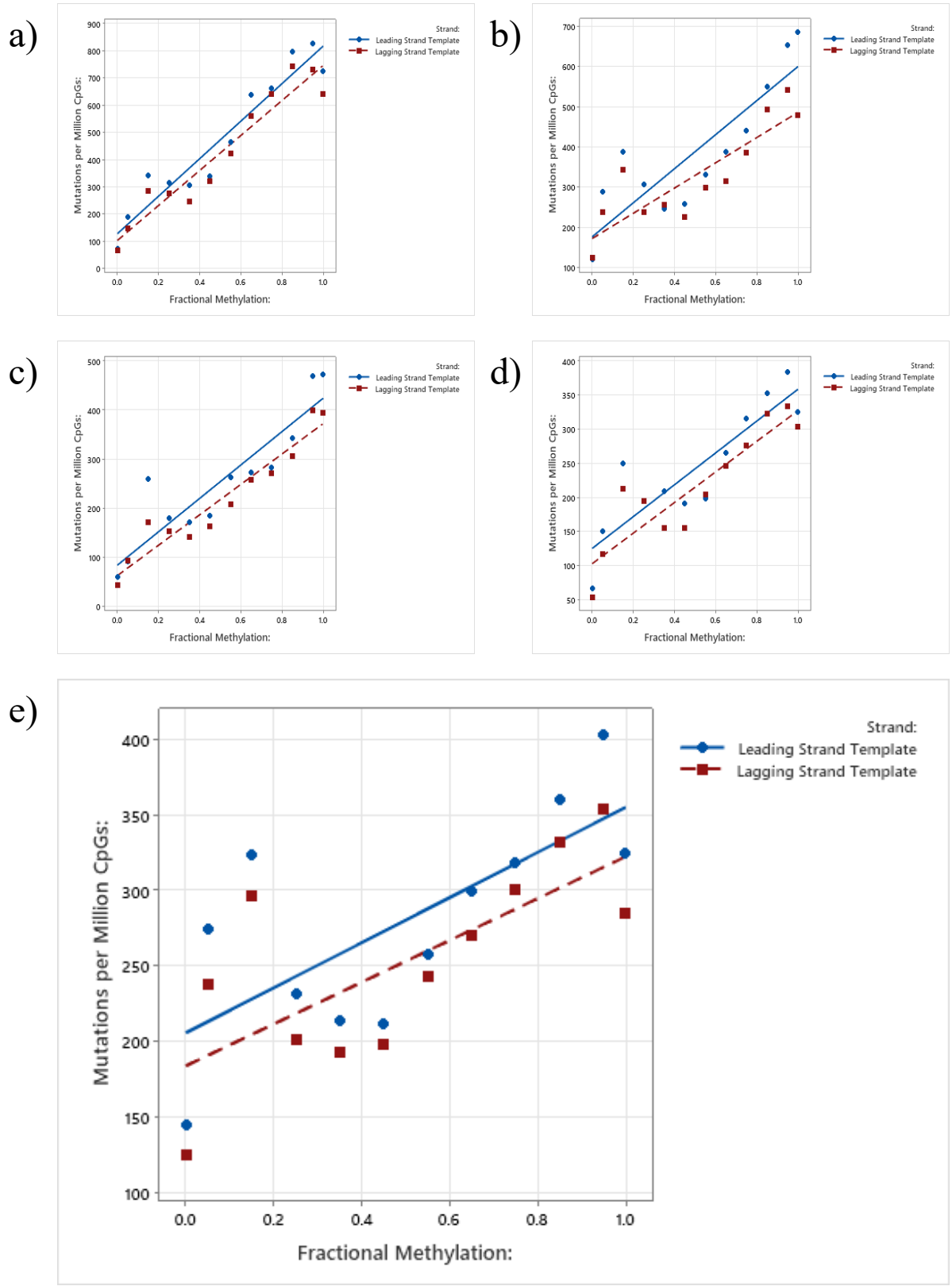


Figure 5.20 – Replication Strand-Specific Associations Between DNA Methylation & C → T Mutagenesis at CpG Sites of MutS-Deficient & MutL-Deficient MSI$^+$ Colorectal Cancers: The relationship between fractional DNA methylation from the normal sigmoid colon and the rate of C → T mutagenesis at CpG sites on either the leading (blue) or lagging (red) template strand for DNA replication. Cancers were classified as Lynch Syndrome MutS-deficient (a), somatic MutS-deficient (b), Lynch Syndrome MutL-deficient (c), somatic MutL-deficient (d) and cancers with presumed *MLH1* promoter hyper-methylation (e).

The data presented above suggest that MSI^+ POL-WT CRCs may accumulate C → T mutations at CpG sites via unrepaired DNA replication errors, with more highly-methylated CpG sites possibly being at greatest risk of mutation. However, the group of MSI^+ cancers with presumed *MLH1* promoter hyper-methylation appear to present with some unique characteristics compared to other MSI^+ POL-WT CRCs. As seen in Figure 5.16 and Figure 5.17, CRCs with presumed *MLH1* promoter hyper-methylation appear to present with a greater C → T mutation rate at lowly-methylated (0 – 30% methylation) CpG sites compared to other MutL-deficient cancers, particularly in the latest-replicating region of the genome. As discussed in section 1.4.4, the study by Samowitz *et al.* suggested that *MLH1* promoter hyper-methylation represents one of the characteristics of $CIMP^+$ cancer. As discussed in Chapter I and Chapter III of this thesis, $CIMP^+$ cancers are characterised by substantial hyper-methylation of CpG sites throughout the genome, possibly driving tumorigenesis via the epigenetic silencing of tumour suppressor genes. Throughout this thesis, it has been suggested that highly-methylated CpG sites are more at risk of C → T mutagenesis than lowly-methylated CpG sites – whether tha be via unrepaired spontaneous deaminations of 5-mC or DNA replication errors. Therefore, it could be that if these cancers with presumed *MLH1* promoter hyper-methylation were $CIMP^+$, the consequent hyper-methylation of CpG sites that are lowly-methylated in the normal sigmoid colon may increase their risk of C → T mutagenesis – potentially explaining the increase in mutations in the 0 – 30% methylation region in these cancers.

In order to determine if these CRCs were likely to be $CIMP^+$ in the absence of DNA methylation array data, CRCs with presumed *MLH1* promoter hyper-methylation were searched for other characteristics of $CIMP^+$ disease. As discussed in Chapter III of this thesis, $CIMP^+$ CRCs appear to show preferential hyper-methylation of bivalent promoter regions. Therefore, C → T mutations at CpG sites in the MSI^+ POL-WT CRCs described in section 5.3.2 were mapped to the bivalent promoter regions described by Court & Arnaud (329). As seen in Table 5.2, a total of 855,851 CpG sites mapped to these regions, with more than half of these CpG sites having a fractional methylation of 0 – 30% in the normal sigmoid colon. As seen in Figure 5.21a, the number of C → T mutations at CpG sites within bivalent promoter regions was significantly higher in CRCs with *MLH1* promoter hyper-methylation than LS-associated MutL-deficient CRCs ($p = 0.0048$, Figure 5.21a) and somatic MutL-deficient CRCs ($p = 0.017$, Figure 5.21a). This increase made the number of C → T mutations at CpG sites in bivalent promoter regions in these CRCs with presumed *MLH1* promoter hyper-methylation comparable with the number of mutations seen in LS-associated MutS-deficient CRCs ($p = 0.41$, Figure 5.21a) and somatic MutS-deficient CRCs ($p = 0.46$, Figure 5.21a). The number of C → T mutations at CpG sites in bivalent promoter regions was not significantly different in LS-associated MutL-deficient CRCs and somatic MutL-deficient cancers ($p = 0.66$, Figure 5.21a). Interestingly, the proportion of C → T mutations at CpG sites mapping to bivalent promoter regions (Figure 5.21b) was significantly higher in CRCs with presumed *MLH1* promoter hyper-methylation than LS-associated MutS-deficient CRCs ($p = 8.4 \times 10^{-8}$, Figure 5.21b), somatic MutS-deficient CRCs ($p = 0.012$, Figure 5.21b), LS-associated MutL-deficient cancers ($p = 0.0012$, Figure 5.21b) and somatic MutL-deficient CRCs ($p = 0.015$, Figure 5.21b). Therefore, if the above hypothesis is correct and CIMP-mediated DNA hyper-methylation at bivalent promoters consequently increases the number of C → T mutations, it appears that the CRCs with presumed *MLH1* promoter hyper-methylation may be $CIMP^+$ CRCs.

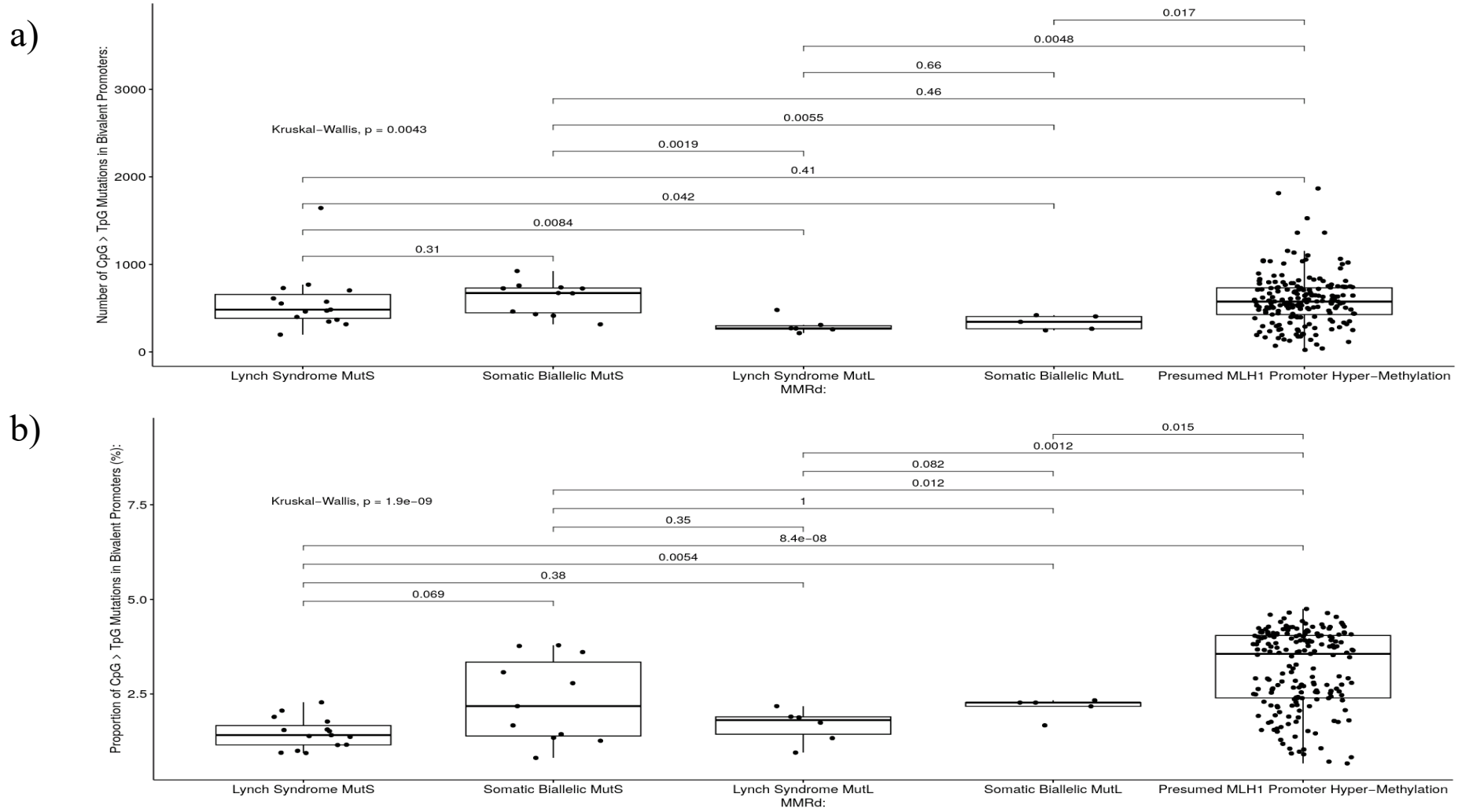


Figure 5.21 – C → T Mutagenesis at CpG Sites within Bivalent Promoters in MutS-Deficient & MutL-Deficient MSI$^+$ Colorectal Cancers: The number (a) and proportion (b) of C → T mutations at CpG sites mapping to bivalent promoter regions of the genome in microsatellite unstable DNA polymerase wild-type colorectal cancers. Cancers were classified as Lynch Syndrome MutS-deficient, somatic MutS-deficient, Lynch Syndrome MutL-deficient, somatic MutL-deficient and cancers with presumed *MLH1* promoter hyper-methylation.

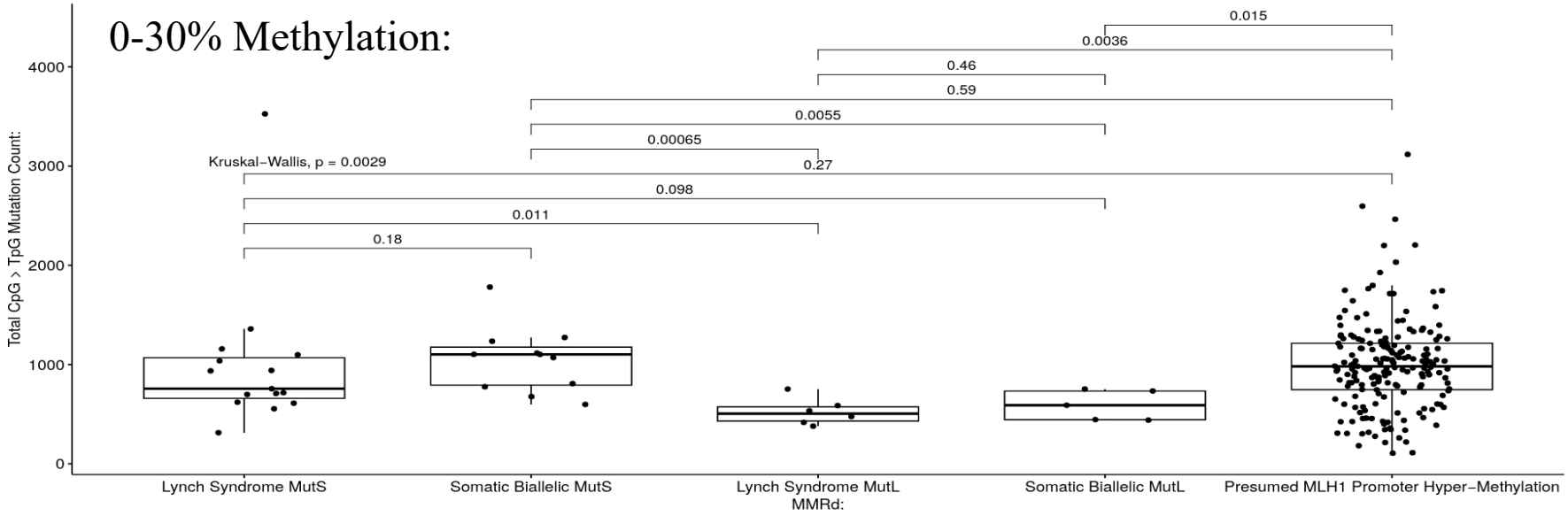
Following this suggestion that the CRCs with presumed *MLH1* promoter hyper-methylation were CIMP$^+$ CRCs, the number of C → T mutations in lowly-methylated (0 – 30%) regions of the genome in the normal sigmoid colon were compared to the rest of the genome (31 – 100% methylation) in these cancers, in order to identify if this hypothesised CIMP-mediated hyper-methylation of lowly-methylated CpG sites drives an increase in the rate of C → T mutagenesis. As seen in Figure 5.22a, the number of C → T mutations in these lowly-methylated regions of the genome was significantly greater in the CRCs with presumed *MLH1* promoter hyper-methylation than in LS-associated MutL-deficient CRCs ($p = 0.0036$, Figure 5.22a) and somatic MutL-deficient cancers ($p = 0.015$, Figure 5.22a). Furthermore, the number of C → T mutations at lowly-methylated CpG sites was comparable to the number seen in LS-associated MutS-deficient CRCs ($p = 0.27$, Figure 5.22a) and somatic MutS-deficient CRCs ($p = 0.59$, Figure 5.22a). Interestingly, the proportion of the total C → T mutation burden at these lowly-methylated CpG sites in CRCs with presumed *MLH1* promoter hyper-methylation was significantly greater than that of LS-associated MutS-deficient CRCs ($p = 3.4 \times 10^{-8}$, Figure 5.22b), somatic MutS-deficient cancers ($p = 0.0053$, Figure 5.22b), LS-associated MutL-deficient cancers ($p = 0.00053$, Figure 5.22b) and somatic MutL-deficient CRCs ($p = 0.012$, Figure 5.22b). In comparison, the number of C → T mutations at CpG sites with a DNA methylation of between 31 – 100% in the normal sigmoid colon in these CRCs with presumed *MLH1* promoter hyper-methylation was significantly lower than the number of mutations in LS-associated MutS-deficient CRCs ($p = 1.3 \times 10^{-7}$, Figure 5.22c) and somatic MutS-deficient CRCs ($p = 0.00041$, Figure 5.22c) – while not being significantly different to the number of mutations in LS-associated MutL-deficient CRCs ($p = 0.65$, Figure 5.22c) and somatic MutL-deficient cancers ($p = 0.54$, Figure 5.22c).

In summary, it appears that MSI$^+$ CRCs with presumed *MLH1* promoter hyper-methylation may show some characteristics of CIMP$^+$ disease and potentially show an increased rate of C → T mutagenesis at lowly-methylated CpG sites as a consequence of CIMP-mediated DNA hyper-methylation increasing the likelihood of C → T mutagenesis occurring. However, from this data it was not possible to suggest by which mechanism – unrepaired spontaneous deaminations or replication errors – may drive this increase in C → T mutagenesis at these lowly-methylated regions.

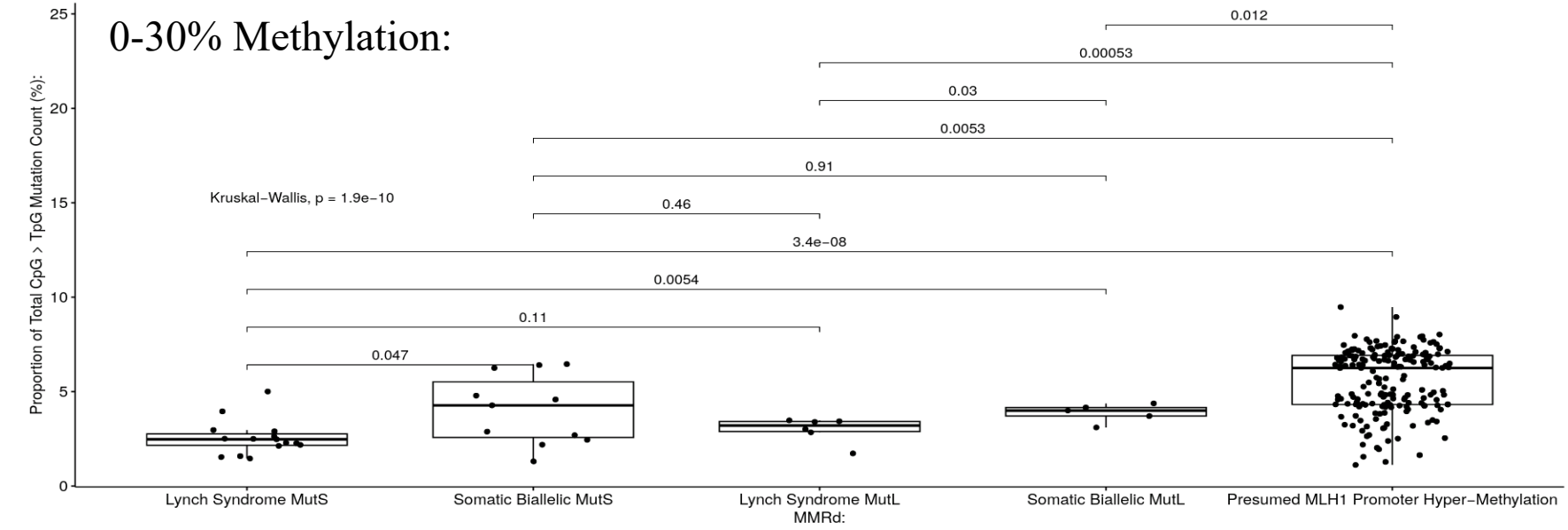
5.4 – Discussion

C → T mutations at CpG sites represent the most common SNV in human cancer (312,426,434). Chapter IV of this thesis explored one of the mechanisms by which these mutations can occur in CRC, via the spontaneous deamination of 5-mC to thymine. While this represents one mechanism of C → T mutagenesis at CpG sites, the significantly elevated number of these mutations in MSI$^+$ POL-WT CRCs and cancers with *POL-ε* EDMs compared to MSS POL-WT cancers suggests that there may be an alternative mechanism by which C → T mutagenesis at CpG sites is driven. The study by Tomkova *et al.* suggested that if the number of G**C**G → G**T**G mutations in MSI$^+$ and *POL-ε* mutant cancers were produced by unrepaired spontaneous deamination alone, it would take twenty-eight and ninety-seven years respectively (524). This calculation assumes that every spontaneous deamination is propagated into a mutation and none are repaired, therefore the true estimate may be

a)


0-30% Methylation:
Total CpG > TpG Mutation Count:
Kruskal–Wallis, p = 0.0029
0.18
0.011
0.098
0.27
0.00065
0.0055
0.59
0.46
0.0036
0.015
Lynch Syndrome MutS
Somatic Biallelic MutS
Lynch Syndrome MutL
Somatic Biallelic MutL
Presumed MLH1 Promoter Hyper–Methylation
MMRd:

b)


0-30% Methylation:
Proportion of Total CpG > TpG Mutation Count (%):
Kruskal–Wallis, p = 1.9e–10
0.047
0.11
0.0054
3.4e–08
0.46
0.91
0.0053
0.03
0.00053
0.012
Lynch Syndrome MutS
Somatic Biallelic MutS
Lynch Syndrome MutL
Somatic Biallelic MutL
Presumed MLH1 Promoter Hyper–Methylation
MMRd:

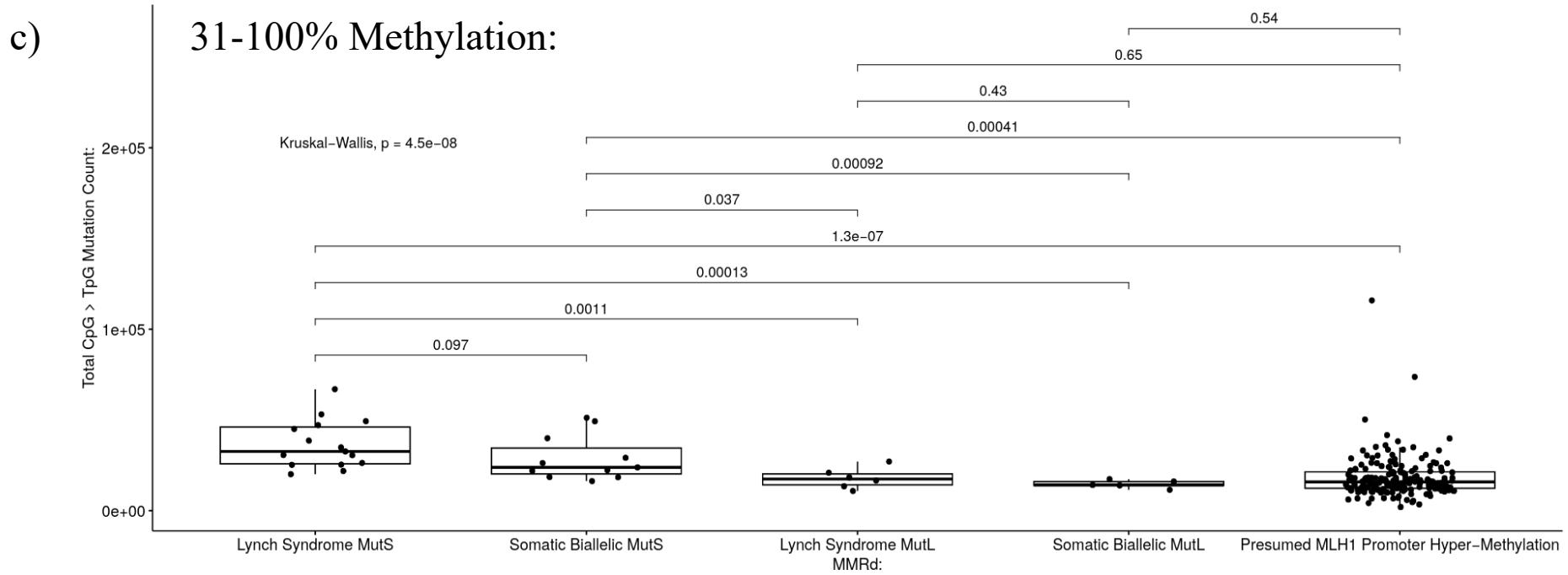


Figure 5.22 – The Number of C → T Mutations at Lowly-Methylated CpG Sites in MutS-Deficient & MutL-Deficient MSI$^+$ Colorectal Cancers: The number (a) and proportion (b) of C → T mutations mapping to lowly-methylated (0 – 30% methylation) CpG sites and the number of C → T mutations at more highly-methylated (31 – 100% methylation) regions in microsatellite unstable DNA polymerase wild-type colorectal cancers. Cancers were classified as Lynch Syndrome MutS-deficient, somatic MutS-deficient, Lynch Syndrome MutL-deficient, somatic MutL-deficient and cancers with presumed *MLH1* promoter hyper-methylation.

considerably longer than this, further indicating that another mechanism may be involved in C → T mutagenesis at CpG sites in these cancers. Mutation signatures associated with *POL-ε* EDMs and MSI^+ cancers (see Figure 5.1) have also been suggested to present with an excess of C → T mutations at CpG sites on the template for leading strand synthesis – indicating that unrepaired *POL-ε* replication errors may be the alternative mechanism underpinning C → T mutagenesis at CpG sites in these cancers (485). This claim is supported by previous evidence that *POL-ε* has difficulty replicating 5-mC, suggesting that highly-methylated regions of the genome are more prone to replication-induced DNA mismatches – which are normally repaired by a combination of *POL-ε* "proofreading" and the MMR pathway (524).

Unsurprisingly, MSI^+ POL-WT cancers and CRCs with *POL-ε* EDMs presented with significantly more SNVs and C → T mutations at CpG sites than MSS POL-WT CRCs. *POL-ε* mutant cancers were significantly younger than their MSS POL-WT counterparts, indicating that the increase in C → T mutagenesis at CpG sites was not a consequence of older age. While MSI^+ cancers were significantly older than their MSS POL-WT counterparts, the substantial increase in C → T mutagenesis at CpG sites in these cancers was unlikely to be a consequence of this, given the estimate by Tomkova *et al.* on the timescale that would hypothetically be required in these cancers to accumulate that number of deaminations (524). Following the identficiation of the increased SNV burden and number of C → T mutations at CpG sites in MSI^+ POL-WT CRCs and CRCs with *POL-ε* EDMs compared to MSS POL-WT cancers, mutation signature analysis also revealed a significant increase in the number (but not proportion) of SNVs attributed to the mutation signature SBS1 in these cancers compared to their MSS POL-WT counterparts. Unsurprisingly, MSI^+ POL-WT CRCs were the only group to present with the mutation signature SBS15 and *POL-ε* mutant CRCs were the only cancers to present with SBS10b. In addition to this, MSI^+ cancers and CRCs with *POL-ε* EDMs presented with a significantly steeper regression slope for the relationship between DNA methylation and the rate of C → T mutagenesis at CpG sites than their MSS POL-WT counterparts. This find has been previously reported by Tomkova *et al.* and the study by Poulos *et al.* (453,524).

However, the data presented in this chapter was not able to accurately assess the additive effect of defective *POL-ε* "proofreading" and MMR in the context of CRC. The only primary, treatment-naïve MSI^+ *POL-ε* mutant CRC with a PCR-free library preparation in the 100KGP dataset harboured a $POL\text{-}\varepsilon^{P286L}$ mutation, whereas the EDMs present in MSS CRCs were $POL\text{-}\varepsilon^{P286R}$, $POL\text{-}\varepsilon^{V411L}$, $POL\text{-}\varepsilon^{A456P}$ and $POL\text{-}\varepsilon^{S459F}$ – making comparisons difficult. In order to accurately assess the additive effect of MSI and *POL-ε* EDMs, the same analysis could be extended to endometrial cancers of the 100KGP, where MSI^+ *POL-ε* mutant cancers are more common. As seen in Table 5.1, an additional MSI^+ CRC with a *POL-δ* EDM ($POL\text{-}\delta^{G321D}$) was also identified in the 100KGP dataset. However this cancer presented with no mutation signatures associated with *POL-δ* EDMs (e.g. SBS10c, SBS10d and SBS20), indicating that $POL\text{-}\delta^{G321D}$ was not a pathogenic EDM, which would explain why it has not been previously reported in the literature. Therefore, it was not possible to explore the role of unrepaired *POL-δ* replication errors in C → T mutagenesis at CpG sites in this chapter, although this may be possible in the context of endometrial cancer. The mutation signatures listed above associated with *POL-δ* EDMs do not present with an enrichment of C → T mutations at CpG sites, indicating that *POL-δ* mediated DNA "proofreading" may not be responsible for the repair of these mutations in any case. The study by Andrianova *et al.* found no enrichment of

C → T mutations of any type in cancers with *POL-δ* EDMs, indicating that these cancers do not display elevated C → T mutagenesis at CpG sites driven by *POL-δ* replication error (526).

One interesting aspect of the binning of C → T mutations at CpG sites based on their fractional methylation in the normal sigmoid colon – performed in this chapter and in Chapter IV of this thesis – was the seemingly bimodal distribution of the association between DNA methylation and the rate of C → T mutagenesis at CpG sites. The initial positive correlation between these two variables often peaks at around 20% methylation – before falling, rising again from around 40% methylation, peaking again at around 90% methylation before a final fall. This suggests that the relationship between DNA methylation and the rate of C → T mutagenesis at CpG sites is far from linear, despite the significant r^2 between these two variables in the CRCs analysed throughout this thesis. There are a number of reasons that may, at least in part, explain the shape of the curve observed. As seen in Table 4.1, the number of CpG sites in each DNA methylation bin is not equal, with the majority of CpG sites having a DNA methylation of between ninety-one and ninety-nine percent. Interestingly, at the points where the curve starts to fall, the number of CpG sites in the bin is lower. For example, when a fall in the rate of C → T mutagenesis at CpG sites occurs at 100% methylation compared to previous bins, the number of CpG sites in this bin is only 746,947 – compared to millions in the previous bins. This suggests that the shape of the curve may be a technicial consequence of the whole-genome bisulphite sequencing analysis. However, as discussed in section 4.4, it may also be that the use of the normal sigmoid colon for fractional methylation reference data may not have been the most representative of the DNA methylation profiles of CRCs. Therefore, the shape of the curve may be a consequence of the changes in DNA methylation in each cancer compared to the reference data from the normal colon. As also discussed in section 4.4, the ideal analysis would have included DNA methylation analysis of each CRC studied in this chapter – but this was not feasible.

Subsequent analysis identified the likely mechanism by which DNA mismatches erroneously produced by *POL-ε* during the replication of the leading strand template may drive C → T mutagenesis at CpG sites. Previous studies by Tomkova *et al.* have suggested that an erroneous C:A DNA mismatch produced by *POL-ε* could be propagated into a C → T mutation in the next round of DNA replication (524). Following this, the $\log_2$(Leading/Lagging) ratio of C → T mutations at CpG sites revealed no significant differences between the *MBD4*-mutant colorectal polyps and MSS POL-WT CRCs – indicating that spontaneous deaminations of 5-mC may be the driving force underpinning C → T mutagenesis at CpG sites in these groups. The identification of a comparable number of C → T mutations at CpG sites on both the leading and lagging streand templates in the *MBD4*-mutant colorectal polyps supports what has previously been described by Fang *et al.,* who reported the same phenomenon in *MBD4*-mutant cancers (518). This suggests that these mutations were likely a consequence of unrepaired spontaneous deaminations of 5-mC, a process which has been suggested to be independent of DNA replication – implying that deaminations are theoretically equally likely to occur on both the leading and lagging strand templates (518,519,524). Interestingly, despite not having a significantly different $\log_2$(Leading/Lagging) ratio of mutations to these *MBD4*-mutant polyps, a number of MSS POL-WT CRCs appeared to show a greater number of C → T mutations on the lagging strand template. As discussed above, the lack of C → T mutations at CpG sites in SBS10c,

SBS10d and SBS20 suggests that this excess of mutations on the lagging strand template was unlikely to be a consequence of unrepaired *POL-δ* replication errors, therefore suggesting that the C → T mutations at CpG sites in these cancers were also a consequence of unrepaired deaminations. Previous studies by Tomkova *et al.* and Fang *et al.* also report a $\log_2$(Leading/Lagging) ratio of C → T mutations at CpG sites of close to zero in MSS cancers, with a small excess of mutations on the lagging strand template (485,518). This is unexpected given what has been described in *MBD4*-mutant samples and possibly implicates DNA replication in the spontaneous deamination of 5-mC. During DNA replication, the leading strand is synthesised continuously and travels in the same direction as the DNA helicase which unwinds the DNA double helix during replication, whereas the lagging strand is synthesised discontinuously in the opposite direction to the helicase in short stretches known as Okazaki fragments (527). While the rate of elongation of both strands of the replication fork are comparable, it is possible that the discontinuous nature of lagging strand synthesis exposes the single-stranded DNA of the template strand for longer than the continuously-replicated single-stranded DNA of the leading strand template. As discussed in Chapter IV of this thesis, single-stranded DNA is more vulnerable to spontaneous deamination than double-stranded DNA, meaning that the exposed single-stranded DNA of the lagging strand template may be more at risk of spontaneous deamination than the single-stranded DNA on the leading strand template. Bhagwat *et al.* investigated the rate of cytosine to uracil deamination following the expression of a single-stranded DNA cytosine deaminase in bacterial cells with defective uracil repair mechanisms (438). The study found an excess of unrepaired cytosine deaminations on the lagging strand template following the expression of cytosine deaminase in bacteria, with the number of unrepaired deaminations unsurprisingly being highest in repair-deficient bacteria (438). This data, coupled with the data presented in this chapter, imply that the process of DNA replication may, indirectly, influence the spontaneous deamination of 5-mC.

The assignment of C → T mutations at CpG sites to either the leading or lagging replication strand template was achieved using a map of left-replicating and right-replicating regions of the genome from the study by Haradhvala *et al.* (520). This data used replication timing profiles from lymphoblastoid cell lines described by Koren *et al.,* separating 20,000 base-pair regions of the genome into one of these two categories according to the predominant direction in which a region was replicated, thereby reducing noise (520,523). While this method has been used to assign mutations to replication strands in a number of previous studies, there are a number of limitations associated with this method (485,518,520,524). The smoothing of data into regions defined as "predominantly left-replicating" and "predominantly right-replicating" doesn't make use of every origin of replication in a genomic region for fork mapping, instead using only the consensus from a set of replication origins within the same 20,000 base-pair region. A number of previous studies have made use of a number of different sequencing technologies to construct a high-resolution map of human replication origins (528–530). Using these maps, mutations could be accurately be called on the template for leading or lagging strand synthesis instead of using the consensus map from the study by Haradhvala *et al.,* which mapped mutations to "predominantly" left-replicating or right-replicating regions. However, the map of human replication origins remains far from complete, with the above studies often only defining a few thousand replication origins with high-confidence in the entire genome (520,528,530). Therefore, these maps would only encompass a fraction of the ~27,000,000 CpG sites in the genome,

therefore representing a poor reference dataset for replication strand mutation mapping due to this poor coverage. The data from Haradhvala *et al.* allows ~35% of CpG sites to be assigned a direction of replication, thus representing the option that provided the best coverage of the genome to use in this analysis.

In comparison to these MSS POL-WT cancers, the $\log_2$(Leading/Lagging) ratio was significantly greater in MSI^+ POL-WT CRCs and MSS *POL-ε* mutant CRCs, indicating a significant excess of mutations on the leading strand. While this was expected in the *POL-ε* mutant CRCs, given the well-defined role of *POL-ε* in the replication of the leading strand, the excess of mutations on the leading strand in the MSI^+ POL-WT CRCs supports the previous work by Tomkova *et al.,* who also suggested that unrepaired DNA replication errors may contribute to C → T mutagenesis at CpG sites in these cancers . However, previous work by Fang *et al.* suggested that MutS-deficient cancers accumulated C → T mutations at CpG sites via unrepaired spontaneous deaminations of 5-mC, whereas MutL-deficient cancers accumulate these mutations via unrepaired DNA replication errors (518). Therefore, MutS-deficient CRCs should theoretically present with a $\log_2$(Leading/Lagging) ratio of C → T mutations at CpG sites of approximately zero – while MutL-deficient cancers should present with a $\log_2$(Leading//Lagging) ratio of greater than zero, indicating an excess of mutations on the leading strand template.

Therefore, the 357 MSI^+ POL-WT CRCs were categorised as either MutS-deficient or MutL-deficient. These included MSI^+ cancers with germline pathogenic mutations in MMR genes, which were defined as LS-associated, cancers with likely biallelic somatic pathogenic MMR mutations (as determined by χ^2 analysis) and cancers with multiple somatic mutations in the same MMR gene. While these cancers with multiple mutations in the same MMR gene were assumed to have mutations different alleles, phasing of these mutations to confirm this was not possible – thereby representing one of the limitations of the analysis. A total of 189 MSI^+ CRCs presented with no mutations in any MMR gene – leading to the assumption that the MSI^+ phenotype in these cancers was a consequence of *MLH1* promoter hyper-methylation. This represents another limitation of this analysis. In the absence of RNA-sequencing or DNA methylation data from these cancers, cancers could only be assumed to have somatic biallelic MMR mutations or *MLH1* promoter hyper-methylation. Should the appropriate data have been available, a more comprehensive analysis of these cancers could have been performed. However, the exclusion of a number of CRCs where a definitive MutS or MutL classification could not be made, including cancers with likely monoallelic truncations in MMR genes, allowed the analysis of a high-confidence set of MutS-deficient or MutL-deficient cancers. In total, 189 (of the 226 cancers included in the final analysis (83.6%) were assumed to have *MLH1* promoter hyper-methylation, which fits previous estimates that *MLH1* methylation is the cause of MSI in the majority of CRCs (89). The study by Fang *et al.* analysed 316 MSI^+ CRCs, endometrial cancers and stomach adenocarcinomas from the TCGA dataset – using the available RNA-sequencing and DNA methylation array data to reliably define the cancers as MutS-deficient or MutL-deficient (518). However, the analysis performed in this chapter has the advantage of including data from whole-genome sequencing analysis of a single type of cancer, compared to the predominantly whole-exome sequencing data used by Fang *et al.,* who included a number of distinct cancer types in the same analysis (518). The use of whole-exome sequencing data for the analysis of C → T mutagenesis at CpG sites has several limitations, including the fact that exonic DNA is often restricted to

early-replicating regions of the genome and only encompasses a fraction of the number of CpG sites that whole-genome sequencing analysis allows – making whole-exome sequencing data not an accurate representation of the mechanics of C → T mutagenesis at CpG sites across the whole genome.

In line with what has previously been reported by Fang *et al.* and Sanders *et al.,* MutS-deficient CRCs (both LS-associated and somatic) presented with enhanced C → T mutagenesis at CpG sites compared to MutL-deficient cancers that could not be explained by age (518,519). In addition to this, these MutS-deficient cancers showed a greater enrichment of the mutation signature SBS1 compared to their MutL-deficient counterparts, supporting the study by Giner-Calabuig *et al.,* who performed *de novo* mutation signature extraction from 105 MSI^+ cancers and found that 62.5% of cancers in the cluster that most resembled SBS1 were MutS-deficient (531). These MutS-deficient CRCs also presented with a steeper regression slope for the association between DNA methylation and the rate of C → T mutagenesis at CpG sites than the MutL-deficient CRCs. This increase of C → T mutagenesis at CpG sites in MutS-deficient cancers may be a consequence of the different roles of these complexes in the MMR pathway. As discussed in section 5.1.3, the MutS complex is crucial for the initial recognition of a DNA mismatch and recruiting the MutL complex to the mismatch for subsequent repair (532). It is possible that in MutS-deficient cancers, a lack of mismatch recognition could lead to a near-complete abolition of MMR, whereas in MutL-deficient cancers DNA mismatches are still recognised by the MutS complex and may be repaired by the remaining components of the MMR pathway – although this hypothesis would require substantial further investigation.

Surprisingly, when C → T mutations at CpG sites were mapped to a replication strand in these MutS-deficient and MutL-deficient cancers, there was no significant difference in the $\log_2$(Leading/Lagging) ratio between any of the groups. This ratio indicated that there was an excess of mutations on the leading strand template, suggesting that unrepaired DNA replication errors were involved in C → T mutagenesis at CpG sites in both MutS-deficient and MutL-deficient CRCs. While this is contrary to both the hypothesis set out in section 5.1.4 and the analysis of whole-exome sequencing by Fang *et al.,* the analysis of the limited number of whole-genome sequenced cancers from the TCGA dataset by Fang *et al.* also revealed an excess of C → T mutations on the leading strand in both MutS-deficient and MutL-deficient cancers (518). Previous work by Tomkova *et al.* has supported the suggestion that MSI^+ cancers present with an excess of C → T mutations at CpG sites on the leading strand template, although not to the same extent as CRCs with *POL-ε* EDMs (524). This provides a model to suggest that in these MSI^+ cancers, *POL-ε* "proofreading" mechanisms (which are still intact) repair the majority of the C:A mismatches erroneously produced during DNA replication before they are propagated into C → T mutations in the next round of replication, while also implying that MMR is required to repair those mismatches which escape DNA "proofreading". This fits the canonical role of MMR in the context of all DNA mismatches described in section 5.1.3, suggesting that unrepaired DNA replication errors that escape *POL-ε* "proofreading" are the cause of C → T mutagenesis at CpG sites in MSI^+ CRCs.

The suggestion that an adenine residue could be erroneously incorporated opposite a 5-mC by *POL-ε* during DNA replication is plausible given the structural similarities between 5-mC and thymine (see Figure 5.23). Interestingly, when replication strand data was combined with

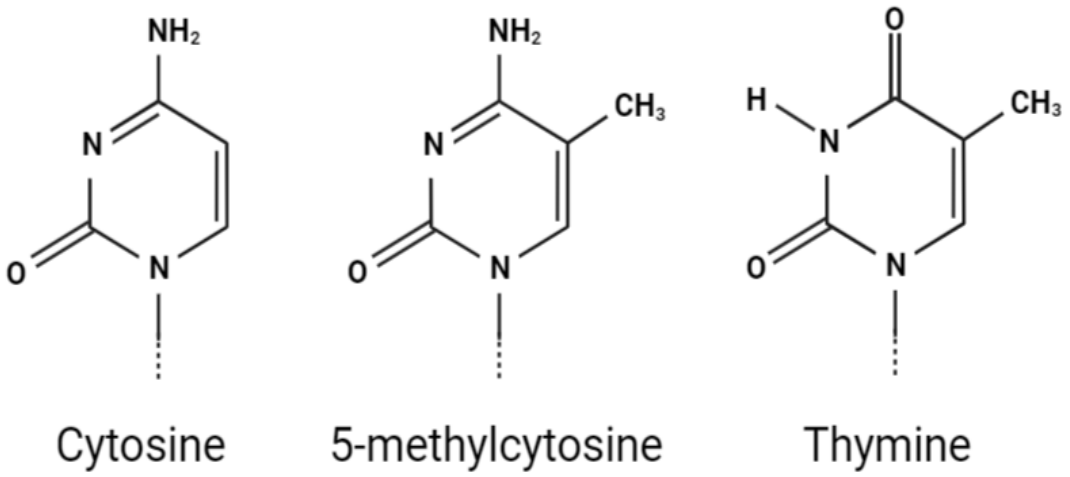


Figure 5.23 – Structural Similarities Between Unmodified Cytosine, 5-methylcytosine & Thymine: The structures of the DNA bases cytosine and thymine, as well as the structure of the modified base 5-methylcytosine. Created with BioRender.com (https://app.biorender.com/).

fractional DNA methylation data, the regression slope of the association between DNA methylation and the rate of C → T mutagenesis at CpG sites was significantly greater on the leading strand template than the lagging strand template in MSS *POL-ε* mutant CRCs, indicating that more highly-methylated CpG sites were more at risk of C:A mispairings during DNA replication. Therefore, this could provide further evidence for the hypothesis that the erroneous C:A mispairings produced by *POL-ε* during replication are a consequence of mis-recognition of a 5-mC on the template strand as a thymine, given their structural similarities. As seen in Figure 5.23, unmodified cytosine bears less resemblance to thymine than 5-mC does, thereby potentially reducing the likelihood that that it is mis-recognised as a thymine by *POL-ε* during DNA replication, consequently resulting in the lower rate of replication errors seen at lowly-methylated CpG sites. This data supports what has been previously published by Tomkova *et al.,* who also found that the rate of *POL-ε* DNA replication error was higher at more highly-methylated CpG sites (524).

Furthermore, the study by Soriano *et al.* examined the equivalent of human *POL-ε*^{P286R} in Schizosaccharomyces pombe (*pol2*^{P287R}) and found the expected increase in T**C**T → T**A**T mutations that characterises SBS10a in these yeast but were unable to detect the T**C**G → T**T**G mutation that characterises SBS10b (499). These organisms lack DNA methylation, further suggesting that C → T mutagenesis at CpG sites in *POL-ε* mutant organisms is, at least in part, influenced by DNA methylation. Recently, Buitrago *et al.* expressed murine DNMTs in yeast in order to explore the role of DNA methylation in the three-dimensional organisation of the genome (533). This model system could be used to examine the role of DNA methylation in C → T mutagenesis at CpG sites in *POL-ε* mutant yeast. If the above hypothesis is true and methylated CpG sites are more at risk of erroneous C:A DNA mismatches, the transformation of *pol2*^{P287R} yeast with these DNMTs should increase the number of C → T mutations at CpG sites, especially in the context T**C**G.

Finally, when analysing the association between DNA methylation and the rate of C → T mutagenesis in MSI^{+} POL-WT CRCs, the r^2 correlation between these variables was lower in

cancers with presumed *MLH1* promoter hyper-methylation. This may be a consequence of the increase in the C → T mutation rate at lowly-methylated CpG sites (0 – 30% methylation in the normal colon reference dataset). As discussed in section 5.3.4, this may be a consequence of CIMP-mediated hyper-methylation of these CpG sites in these cancers, consequently increasing the rate of C → T mutagenesis. Samowitz *et al.* described *MLH1* promoter hyper-methylation as one of the hallmarks of $CIMP^+$ cancer, indicating that these cancers may well be $CIMP^+$, which was further evidenced by the increase in the number of C → T mutations mapping to CpG sites within bivalent promoter regions in these cancers compared to other MSI^+ POL-WT CRCs. As described in Chapter III of this thesis, $CIMP^+$ cancers may show preferential hyper-methylation of these bivalent promoter regions, further indicating that the enrichment of C → T mutations in these regions may be a consequence of CIMP-mediated DNA hyper-methylation. As also discussed in Chapter III of this thesis, $CIMP^+$ disease was also correlated with tumours of the proximal colon and mutations in *BRAF* or *KRAS.* Therefore, it may have been possible to study if these CRCs with presumed *MLH1* promoter hyper-methylation also correlated with these characteristics of $CIMP^+$ disease, which may have added further evidence in the absence of the appropriate DNA methylation data from these cancers. Therefore, this represents one of the limitations of this analysis. In the absence of confirmatory methylation array data, these CRCs could only be assumed to have *MLH1* promoter hyper-methylation and be $CIMP^+$. While the same analysis could have been performed in CRCs from the TCGA dataset, which have the necessary DNA methylation data and CIMP statuses, this analysis would have been hampered by the same drawbacks of whole-exome sequencing data described above. DNA methylation array data is currently being produced for CRCs from the 100KGP dataset, which would allow for a more comprehensive analysis of these CRCs with presumed *MLH1* promoter hyper-methylation.

In conclusion, the data presented in this chapter suggest that unrepaired C:A DNA mismatches produced by *POL-ε* during DNA replication of the leading strand represents an alternative mechanism of C → T mutagenesis at CpG sites to spontaneous deamination of 5-mC. This chapter also presented data suggesting that highly-methylated CpG sites were more at risk of these C:A DNA mismatches than lowly-methylated sites in *POL-ε* mutant CRCs. These unrepaired replication errors also appear to play a significant role in C → T mutagenesis at CpG sites in MSI^+ cancers. Overall, this data provides novel insight into how DNA methylation affects the mechanisms of C → T mutagenesis at CpG sites, in turn providing a deeper understanding of the mechanisms by which *POL-ε* EDMs and MMR deficiency may drive CRC tumorigenesis.

Chapter VI – Conclusions & Future Perspectives

DNA methylation represents a key mechanism by which gene expression can be regulated in cells – while at the same time also representing a mechanism by which aberrant gene expression occurs in cancer (211,534). In this thesis, the mechanisms by which DNA hyper-methylation can drive CRC pathogenesis have been described, either by inducing the silencing of key tumour suppressor genes or by being more directly involved in the mechanisms underpinning C → T mutagenesis at CpG sites within the protein-coding sequence of cancer driver genes. In doing so, this thesis has provided novel insights into the mechanisms by which CIMP may drive colorectal tumorigenesis and how DNA hyper-methylation drives C → T mutagenesis via a combination of unrepaired spontaneous deaminations and DNA replications errors occurring at 5-mC.

Chapter II of this thesis, via comprehensive analysis of GWAS meta-analysis and gene expression data, provided evidence suggesting that reduced *TET2* expression in intestinal tissues predisposed individuals to CRC. While somatic mutations in *TET2* have previously been reported in CRC and a recent CRC landscape paper identified *TET2* as a candidate CRC driver gene, this presents what is (to the best of my knowledge) the first suggestion that inherited variations in *TET2* expression can predispose affected individuals to CRC development. This finding could be explored functionally via genetically-engineered cell line and/or organoid models. This system could be engineered to contain the various risk alleles of the candidate causal variants identified in the chapter. RNA-sequencing data from these cell lines and/or organoid models could then be used to validate the effect of these candidate causal variants on *TET2* expression. These cell lines and/or organoid models could also be used for subsequent Capture Hi-C analysis (see section 2.4), in order to evaluate the long-range effects of these candidate causal variants on distant regions of the genome – potentially revealing additional candidate CRC predisposition genes.

The 4q24 locus of chromosome 4 has previously been explored in the context of breast cancer and the lead variant in European populations around *TET2,* rs7679673, has previously been associated with prostate cancer predisposition (247,273). In light of this, it may also be worth extending the analyses performed in this chapter to breast and prostate cancer GWAS data. Summary statistics for each of these cancer types could be obtained from the Breast Cancer Association Consortium (BCAC) and The Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome (PRACTICAL) (535,536). This GWAS data could be combined with eQTL data from GTEx for the relevant tissues or even larger-scale RNA-sequencing data from genetically-engineered cell lines containing risk alleles for the candidate causal variant(s) to identify if the candidate causal variant(s) reduce the expression of *TET2* in these tissues. Overall, this may represent a new avenue of exploration for the role of *TET2* within solid tumours, which so far has remained limited as much of the previous research into the role of *TET2* in tumorigenesis has been focussed on haematological malignancies.

Chapter III of this thesis characterised *TET2*-mutant and IDH-mutant CRCs, given the well-described effect of 2-HG produced by mutant IDH on *TET2* activity (297,330). The results presented in this chapter implicated mutations in *TET2* and IDH in a novel sub-cluster of $CIMP^{+}$ cancers, characterised by small increases in DNA methylation and enrichment of significantly hyper-methylated probes at bivalent promoter regions. The data presented in this chapter also presented evidence for the epigenetic silencing of candidate tumour suppressor genes in *TET2*-mutant and IDH-mutant cancers, providing a mechanistic insight into how

DNA hyper-methylation may have driven tumorigenesis in these CIMP$^+$ cancers. While interesting, this data only offers an explanation for CIMP-mediated tumorigenesis in a small subset of CIMP$^+$ cancers. There are a number of other candidate genes that may drive DNA hyper-methylation (and therefore CIMP) in cancer. Examples of these candidates include a number of the DNMT proteins, including the *de novo* methyltransferases *DNMT3A* and *DNMT3B* (174,175). The study by Heyn *et al.* identified a gain-of-function mutation in *DNMT3A* which was associated with hyper-methylation of developmental genes in microcephalic dwarfism (537). The majority of *DNMT3A* mutations reported in AML are *DNMT3A*^{R882H}, which is thought to be associated with loss-of-function (538). In light of this, in addition to the lack of gain-of-function mutations reported in the literature in the context of CRC, it seems unlikely that *DNMT3A* represents a candidate CIMP driver gene. However, the study by Nosho *et al.* suggests that expression of *DNMT3B* is correlated with CIMP$^+$ CRC (539). Additionally, the study by Verma *et al.* suggested DNA hyper-methylation at bivalent promoters in triple TET-knockout human embryonic stem cells was mediated by *DNMT3B* (407). It was described in Chapter III of this thesis that hyper-methylation of bivalent promoter regions was a characteristic CIMP$^+$ cancers, meaning that *DNMT3B*-mediated hyper-methylation of these regions may drive CIMP. Therefore, *DNMT3B* may represent a candidate CIMP driver gene which warrants further investigation. RNA-sequencing and methylation array analysis could be used to determine if expression of *DNMT3B* is indeed correlated with CIMP$^+$ CRC, with the necessary data available from both the TCGA-COAD and TCGA-READ domains. This analysis could subsequently be extended to over-expression of *DNMT3B* in both *in vitro* cell line and/or organoid systems and *in vivo* mouse models of colorectal tumorigenesis.

Alternative candidate CIMP driver genes include the other members of the TET family, *TET1* and *TET3*. Given the roles of these proteins in active DNA de-methylation, it is plausible that loss-of-function mutations in these genes may drive the same DNA hyper-methylation seen in *TET2*-mutant cancers. The study by Neri *et al.* implicated *TET1* in the regulation of *Wnt* signalling in colonic cells via de-methylation of the promoters of *Wnt* inhibitors (199). The study by Mo *et al.* identified loss of *TET3* expression in nearly 30% of CRCs, with *TET3* mutations strongly correlating with MSI$^+$ disease (540). Therefore, the analyses performed in this chapter could be extended to include *TET1*-mutant and *TET3*-mutant CRCs from the TCGA cohort, in order to study the effect of *TET1* or *TET3* loss on DNA methylation in CRC. These cancers could then be compared to their *TET2*-mutant counterparts for similarities in the patterns of DNA hyper-methylation. Furthermore, *TET1*-knockout or *TET3*-knockout cell lines, organoids or animal models may provide further insight into the role of these members of the TET family in colorectal tumorigenesis and also provide further opportunities to characterise any DNA hyper-methylation arising as a consequence of *TET1* or *TET3* deficiency. A further candidate CIMP driver gene that warrants similar investigation is *ARID1A,* which plays an important role in chromatin re-modelling (541). Down-regulation of *ARID1A* has been identified in a number of CRCs and has been suggested to be a tumour suppressor gene and driver of CRC tumorigenesis (541,542). Overall, the further investigation of these candidate CIMP driver genes may improve our understanding of the mechanisms of CIMP-mediated tumorigenesis and potentially allow the characterisation of further sub-clusters of CIMP$^+$ cancers.

Chapter IV of this thesis characterised the role of germline *MBD4* mutations in the development of intestinal polyps and a number of cancer types – including breast cancer, sarcoma and uveal melanoma. This chapter identified that germline, but not somatic, mutations in *MBD4* significantly increased the number of C → T mutations at CpG sites compared to their *MBD4*-WT counterparts. These mutations were likely a result of unrepaired spontaneous deaminations of 5-mC, which arise as a consequence of *MBD4* deficiency, which are then propagated into C → T mutations. In addition to this, it was found that highly-methylated CpG sites and CpG sites within late-replicating regions of the genome were more at risk of spontaneous deamination than lowly-methylated CpG sites and CpG sites in early-replicating regions of the genome. Despite previous studies suggesting CpG sites on the transcriptional coding strand were more at risk of spontaneous deamination than the template strand, there was no excess of C → T mutations at CpG sites observed on either transcription strand – which was in agreement with the previous work by Sanders *et al.,* who also saw no excess of C → T mutations at CpG sites on either transcription strand (446,456).

The suggestion that germline loss of *MBD4* drives intestinal polyposis and various types of cancer implies that *MBD4* carries out a novel tumour suppressor function via the repair of spontaneous deaminations of 5-mC. However, as discussed in section 4.4, previous studies of *Mbd4*$^{-/-}$ found no increase in intestinal tumorigenesis compared to control animals (450). This data, coupled with the lifespan of laboratory mice and the hypothesised mechanism by which *MBD4* deficiency may drive tumorigenesis, indicates that mouse models may not represent the best system in which to investigate the role of *MBD4* in tumorigenesis. Given the rarity of individuals with germline pathogenic mutations in *MBD4,* whole-genome sequencing data from individuals with germline deficiencies in *MBD4* aside from those reported in this chapter may be difficult to obtain. While *MBD4*-deficient cell lines and/or organoids could theoretically be genetically-engineered or established from the *MBD4*-deficient individuals described above, long-term culture of these *in vitro* models to allow unrepaired spontaneous deaminations to accumulate would be expensive and unpractical. As briefly described in section 5.4, a recent publication by Buitrago *et al.* presented yeast transfected to express murine DNMT proteins, thereby facilitating DNA methylation within this model organism (533). The rapid population doubling time of yeast could be exploited in order to investigate the rate at which unrepaired spontaneous deaminations of 5-mC accumulate in the methylated yeast described in this study (533). In addition to this, these yeast could subsequently be transfected with murine *Mbd4* in order to investigate the rate at which unrepaired deaminations accumulate in both *Mbd4*-deficient and *Mbd4*-proficient systems. While mapping these unrepaired deaminations to protein-coding genes would be uninformative, the effects of DNA methylation, replication timing and transcription strand on the likelihood of spontaneous deamination could still be investigated to validate what has been seen in human whole-genome sequencing data.

As discussed in section 4.1, *TDG* performs the same role as *MBD4* in the repair of spontaneous deaminations of 5-mC (312,441). Therefore, it may also be that germline pathogenic mutations in *TDG* also increase the rate of C → T mutagenesis at CpG sites as a consequence of unrepaired spontaneous deaminations of 5-mC. If this is true, *TDG* may also represent a novel cancer predisposition gene alongside *MBD4* – potentially driving intestinal polyposis and the same forms of cancer seen in *MBD4*-deficient individuals. If the mechanism by which *TDG* deficiency drives polyposis or tumorigenesis is the same as in

MBD4-deficient individuals, mouse models and *in vitro* models would not be the best system to investigate the consequences of *TDG* deficiency for the same reasons as described above. However, similarly to what has been described above, the yeast models expressing murine DNMTs could also be engineered to express murine *Tdg* to further characterise the role this protein plays in the repair of spontaneous deaminations of 5-mC. This also possibly allows for the co-expression of both *Tdg* and *Mbd4* in these yeast alongside systems which express just one of the two – thereby allowing the contribution of each individual protein in the repair of spontaneous deaminations to be quantified. Overall, the data presented in this chapter suggests that *MBD4* represents a tumour predisposition gene, as well as an intestinal polyposis predisposition gene. Further mechanistic investigation of *MBD4* and *TDG* deficiency in yeast would be important in order to confirm the hypothesised mechanism by which inactivation of these genes may drive disease pathogenesis.

Chapter V of this thesis explored the role of DNA polymerase replication errors in C → T mutagenesis at CpG sites. Mutation signatures associated with *POL-ε* EDMs and DNA MMR deficiencies show an enrichment of C → T mutations at CpG sites much like SBS1, indicating that an alternative mechanism may also be involved in driving C → T mutagenesis at these sites. CRCs with *POL-ε* EDMs and MSI^+ CRCs presented with a much greater number of C → T mutations at CpG sites than their MSS POL-WT counterparts, while MutS-deficient MSI^+ CRCs presented with more C → T mutations at CpG sites than MutL-deficient MSI^+ cancers – as has been previously reported in the literature (518,519). This phenomenon has previously been hypothesised to be a consequence of MutS playing a role in the repair of spontaneous deaminations, whereas MutL has been suggested to be involved in the repair of DNA replication errors (518). However, while the *MBD4*-deficient colorectal polyps presented with no excess of C → T mutations at CpG sites on either DNA replication strand as expected, cancers with *POL-ε* EDMs, MutS-deficient MSI^+ cancers and MutL-deficient MSI^+ cancers all presented with an excess of C → T mutations at CpG sites of the leading strand template – indicating that these mutations were a consequence of unrepaired DNA replication errors. It was hypothesised in this chapter that replication errors on the leading strand template by *POL-ε* result in an adenine being erroneously incorporated opposite a template 5-mC. If this mismatch is not repaired (e.g. in CRCs with a *POL-ε* EDM or MMR deficiency) it is thought to propagate into a C → T mutation in the next round of DNA replication (see Figure 5.9). This mechanism has previously been proposed by Tomkova *et al.* due to the structural similarities between 5-mC and thymine (see Figure 5.23) (524). It was also suggested in this chapter that DNA methylation was correlated with the likelihood of DNA replication error, with highly-methylated CpG sites presenting with more C → T mutations than lowly-methylated sites. In addition to this, MSI^+ CRCs with presumed *MLH1* promoter hyper-methylation showed characteristics of $CIMP^+$ disease – including an enrichment of C → T mutations at bivalent promoter regions. These CRCs with presumed *MLH1* promoter hyper-methylation presented with an increase in C → T mutations at CpG sites with fractional DNA methylation of 0 – 30% in the normal sigmoid colon. It is possible in these cancers that these previously lowly-methylated CpG sites become hyper-methylated as a consequence of CIMP, thereby increasing the likelihood of C → T mutagenesis in these cancers.

The data presented in this chapter suggests that highly-methylated CpG sites, including those that are possibly hyper-methylated in $CIMP^+$ cancers, were more likely to undergo C → T

mutagenesis as a result of DNA replication error. In order to investigate this further, DNA methylation array data from CRCs with presumed *MLH1* promoter hyper-methylation could be generated, allowing the further characterisation of these cancers – including their CIMP status and the DNA methylation status of these CpG sites identified as lowly-methylated in the normal sigmoid colon. Following this, both *in vitro* and *in vivo* methods could be exploited to produce models of *POL-ε* EDMs, MSI and CIMP in order to further investigate the role of DNA methylation on the likelihood of replication errors occurring. For example, intestinal organoids could be derived from MSI^+ CRCs, cancers with *POL-ε* EDMs, MSI^+ CRCs with *POL-ε* EDMs and $CIMP^+$ CRCs. Whole-genome sequencing and methylation array analysis could subsequently be performed on these organoids in order to identify any correlations between DNA methylation and the rate of C → T mutagenesis at CpG sites. Given the current reference data used for replication strand assignment only covers approximately 35% of CpG sites, it may be prudent to produce a comprehensive map of the origins of replication in these organoids via initiation-site sequencing in order to more accurately assign mutations to the leading and lagging strand template strands (528). The number of C → T mutations at CpG sites of the leading and lagging strand templates could then be used to determine if C → T mutations at CpG sites in these organoids were a consequence of unrepaired DNA replication errors.

In addition to these organoids, mouse models of CRC could also be used to study the relationship between DNA methylation and C → T mutagenesis at CpG sites. Animals with conditional *Pol-ε*^{P286R} or *Msh2* alleles can be obtained commercially from the Jackson Laboratory (543) and subsequently crossed with *Vil1-cre* animals (see Chapter III of this thesis) to produce a targeted knock-in of *Pol-ε*^{P286R} and knockout of *Msh2* in intestinal epithelial cells. Similarly to the organoids described above, whole-genome sequencing and methylation array analysis could be performed on intestinal tissues harvested from these animals at a specific age, allowing for more control over the time in which C → T mutations can accumulate at CpG sites compared to organoids derived from CRC patients. In addition to these animals, an *Mlh1*-knockout mouse could also be engineered in order to provide an *in vivo* model of MutL deficiency. An *Mbd4*-deficient mouse is also currently in development, providing a system in which C → T mutagenesis at CpG sites is largely assumed to be a consequence of unrepaired spontaneous deaminations for comparison with the *Pol-ε*^{P286R}, *Msh2*-knockout and *Mlh1*-knockout animals. In addition to crossing the above animals with *Vil1-cre*, germline *Pol-ε*^{P286R}, *Msh2*-knockout, *Mlh1*-knockout and *Mbd4*-knockout animals could be produced via a cross with the universal *Pgk-cre* (544). Organoids could then be derived from a number of tissues with differing rates of cellular turnover (e.g. colon, small intestine, liver, breast etc.) and used for subsequent whole-genome sequencing and methylation analysis. If the rate of C → T mutagenesis at CpG sites is influenced by DNA replication, in theory organoids derived from the tissue with the highest rate of cellular turnover would present with the greatest C → T mutation burden at CpG sites. As discussed in section 5.4, it is possible that DNA replication may also play a role in spontaneous deamination, with single-stranded DNA possibly exposed for longer on the lagging strand template than the leading strand template, thereby making this DNA more vulnerable to spontaneous deamination. If this is true, then it is possible that in the *Mbd4*-knockout mouse the rate of C → T mutagenesis would also be highest in the tissue with the highest rate of DNA replication.

The relationship between DNA methylation and C → T mutagenesis at CpG sites, as well as the mechanism by which these mutations may occur, could also be investigated using the methylated yeast system described above. MMR-deficient yeast strains and yeast with the equivalent *Pol-ε*^{P286R} mutation have previously been described in the literature (499,545). These strains could be transfected with murine DNMTs to produce methylated yeast strains. The number of C → T mutations at CpG sites could then be compared to non-methylated yeast strains to test if the presence of DNA methylation increases the rate of C → T mutagenesis at CpG sites in yeast strains with a *POL-ε* EDM or MMR deficiency. These methylated yeast strains could then be transformed with murine *Mbd4* and *Tdg* to see if the presence of these proteins reduces the rate of C → T mutagenesis at CpG sites. However, given the hypothesised mechanism of how C → T mutagenesis is driven in these systems, it would be expected that the addition of these proteins would do little to reduce the rate of C → T mutagenesis at CpG sites.

The data presented in Chapter III of this thesis suggests that CIMP may drive colorectal tumorigenesis via the epigenetic silencing of key tumour suppressor genes (see above). However, the data presented in Chapter V of this thesis provides an alternative mechanism that may explain CIMP-mediated CRC development. Hyper-methylated CpG sites in CIMP^{+} CRCs may be more at risk of C → T mutagenesis, which may result in pathogenic C → T mutations within the protein-coding sequence of cancer driver genes. The recent CRC landscape paper identified 185 potential CRC driver genes (295). One example of these drivers is *APC,* which has been extensively studied in the context of CRC. Chapter IV of this thesis identified a number of pathogenic mutations in this gene caused by a C → T mutation at a CpG site – including *APC*^{R1450*} (221,295). Therefore, the role of CIMP in driving colorectal tumorigenesis could be investigated in more detail by a comprehensive analysis of both coding sequence mutations and epigenetic alterations in these 185 CRC driver genes. This analysis would confirm if CIMP-mediated CRC pathogenesis is primarily driven by unrepaired replication errors in the coding sequence of driver genes or by aberrant DNA hyper-methylation altering the expression of driver genes. Overall, this chapter provides additional insight into how DNA methylation may drive tumorigenesis, not only via promoter hyper-methylation of tumour suppressor genes but also via influencing the likelihood of *POL-ε* DNA replication errors.

In conclusion, this thesis has provided a comprehensive assessment of the 4q24 locus of chromosome 4 in CRC predisposition and identified down-regulation of *TET2,* a gene involved in the regulation of DNA methylation, as the likely mechanism underpinning this GWAS association. This led to the exploration of the effect of *TET2* down-regulation, either by mutation or inhibition by mutant IDH, on DNA methylation and the development of CIMP^{+} cancer. Upon further investigation of the characteristics of these CIMP^{+} cancers, a possible mechanism underpinning CIMP-mediated tumorigenesis was identified, with distinct patterns of DNA hyper-methylation observed at bivalent promoter regions, resulting in the epigenetic silencing of candidate tumour suppressor genes. In addition to the epigenetic mechanisms by which DNA methylation may drive colorectal tumorigenesis, this thesis also investigated how DNA hyper-methylation may influence the likelihood of C → T mutagenesis at CpG sites. It was suggested that highly-methylated CpG sites were more at risk of C → T mutagenesis than lowly-methylated CpG sites via both spontaneous deamination of 5-mC and the mis-incorporation of adenine opposite a template 5-mC by

POL-ε during DNA replication. Overall, this thesis provides a novel insight into the role of DNA methylation, particularly hyper-methylation, in driving colorectal tumorigenesis – providing mechanistic insights that may subsequently offer both prognostic and clinical benefits in the management of CRC.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;

2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA Cancer J Clin. 2021;

3. McGuigan A, Kelly P, Turkington RC, Jones C, Coleman HG, McCain RS. Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. World Journal of Gastroenterology. 2018.

4. Slack JMW. Molecular Biology of the Cell. In: Principles of Tissue Engineering: Fourth Edition. 2013.

5. Cooper GM. The Cell: A Molecular Approach. 2nd edition. Sinauer Associates. 2000.

6. Aaronson SA. Growth factors and cancer. Science (1979). 1991;

7. Wintersberger U. The selective advantage of cancer cells: A consequence of genome mobilization in the course of the induction of DNA repair processes? (model studies on yeast). Adv Enzyme Regul. 1984;

8. Wang M, Zhao J, Zhang L, Wei F, Lian Y, Wu Y, et al. Role of tumor microenvironment in tumorigenesis. Journal of Cancer. 2017.

9. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011.

10. Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM. An Introduction to Genetic Analysis, 7th edition. Isbn. 2000.

11. Garber JE, Offit K. Hereditary cancer predisposition syndromes. Journal of Clinical Oncology. 2005.

12. Wang Q. Cancer predisposition genes: Molecular mechanisms and clinical impact on personalized cancer care: Examples of Lynch and HBOC syndromes. Acta Pharmacologica Sinica. 2016.

13. Armaou S, Pertesi M, Fostira F, Thodi G, Athanasopoulos PS, Kamakari S, et al. Contribution of BRCA1 germ-line mutations to breast cancer in Greece: a hospital-based study of 987 unselected breast cancer cases. Br J Cancer. 2009;

14. Ramus SJ, Gayther SA. The Contribution of BRCA1 and BRCA2 to Ovarian Cancer. Molecular Oncology. 2009.

15. Su GH, Hruban RH, Bansal RK, Bova GS, Tang DJ, Shekher MC, et al. Germline and somatic mutations of the STK11/LKB1 Peutz-Jeghers gene in pancreatic and biliary cancers. American Journal of Pathology. 1999;

16. Gardner E. A genetic and clinical study of intestinal polyposis a predisposing factor for carcinoma of the colon and rectum. Am J Hum Gen. 1951;3:167–76.

17. Crispens M. Endometrial and ovarian cancer in lynch syndrome. Clin Colon Rectal Surg. 2012;

18. Scott RJ, Sobol HH. Prognostic implications of cancer susceptibility genes: any news? Recent results in cancer research. Fortschritte der Krebsforschung. Progrès dans les recherches sur le cancer. 1999.

19. Ravindran A, He R, Jawad MD, Ketterling RP, Chen D, Oliveira JL, et al. Frequency of Acquired Genetic Mutations and Their Prognostic Impact on Patients with Incidental Finding of Isolated 20q- in Bone Marrow without Morphologic Evidence of a Myeloid Neoplasm. Blood. 2018;

20. Lee A, Moon BI, Kim TH. BRCA1/BRCA2 pathogenic variant breast cancer: Treatment and prevention strategies. Annals of Laboratory Medicine. 2020.

21. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-One.

22. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut. 2017;

23. Baena R, Salinas P. Diet and colorectal cancer. Maturitas. 2015.

24. Boyle P. ABC of colorectal cancer: Epidemiology. BMJ. 2002;

25. Haggar FA, Boushey RP. Colorectal cancer epidemiology: Incidence, mortality, survival, and risk factors. Clin Colon Rectal Surg. 2009;

26. Rawla P, Sunkara T, Barsouk A. Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors. Przeglad Gastroenterologiczny. 2019.

27. Steele SR, Park GE, Johnson EK, Martin MJ, Stojadinovic A, Maykel JA, et al. The impact of age on colorectal cancer incidence, treatment, and outcomes in an equal-access health care system. Dis Colon Rectum. 2014;57(3).

28. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;

29. Sharma R, Abbasi-Kangevari M, Abd-Rabu R, Abidi H, Abu-Gharbieh E, Acuna JM, et al. Global, regional, and national burden of colorectal cancer and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Gastroenterol Hepatol. 2022;7(7).

30. Lee HP, Gourley L, Duffy SW, Estève J, Lee J, Day NE. Colorectal cancer and diet in an asian population—A case-control study among Singapore Chinese. Int J Cancer. 1989;

31. Macquart-Moulin G, Riboli E, Cornée J, Charnay B, Berthezene P, Day N. Case-control study on colorectal cancer and diet in marseilles. Int J Cancer. 1986;

32. Manousos O, Day NE, Trichopoulos D, Gerovassilis F, Tzonou A, Polychronopoulou A. Diet and colorectal cancer: A case-control study in Greece. Int J Cancer. 1983;

33. Pietinen P, Malila N, Virtanen M, Hartman TJ, Tangrea JA, Albanes D, et al. Diet and risk of colorectal cancer in a cohort of Finnish men. Cancer Causes and Control. 1999;

34. Whittemore AS, Wu-williams AH, Lee M, Shu Z, Gallagher RP, Deng-ao J, et al. Diet, physical activity, and colorectal cancer among Chinese in North America and China. J Natl Cancer Inst. 1990;

35. Armstrong B, Doll R. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. Int J Cancer. 1975;

36. Ferguson LR. Meat and cancer. Meat Science. 2010.

37. Shin A, Shrubsole MJ, Ness RM, Wu H, Sinha R, Smalley WE, et al. Meat and meat-mutagen intake, doneness preference and the risk of colorectal polyps: The Tennessee colorectal polyp study. Int J Cancer. 2007;

38. Sugimura T. Nutrition and dietary carcinogens. Carcinogenesis. 2000.

39. Pöschl G, Stickel F, Wang XD, Seitz HK. Alcohol and cancer: genetic and nutritional aspects. Proceedings of the Nutrition Society. 2004;

40. Rossi M, Anwar MJ, Usman A, Keshavarzian A, Bishehsari F. Colorectal cancer and alcohol consumption—populations to molecules. Cancers. 2018.

41. Van Duijnhoven FJB, Bueno-De-Mesquita HB, Ferrari P, Jenab M, Boshuizen HC, Ros MM, et al. Fruit, vegetables, and colorectal cancer risk: The European Prospective Investigation into Cancer and Nutrition. American Journal of Clinical Nutrition. 2009;

42. Tsoi KKF, Pau CYY, Wu WKK, Chan FKL, Griffiths S, Sung JJY. Cigarette Smoking and the Risk of Colorectal Cancer: A Meta-analysis of Prospective Cohort Studies. Clinical Gastroenterology and Hepatology. 2009;

43. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990.

44. Aghabozorgi AS, Bahreyni A, Soleimani A, Bahrami A, Khazaei M, Ferns GA, et al. Role of adenomatous polyposis coli (APC) gene mutations in the pathogenesis of colorectal cancer; current status and perspectives. Biochimie. 2019.

45. Muto T, Bussey HJR, Morson BC. The evolution of cancer of the colon and rectum. Cancer. 1975;

46. Leslie A, Carey FA, Pratt NR, Steele RJC. The colorectal adenoma-carcinoma sequence. British Journal of Surgery. 2002.

47. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM BJ. Genetic Alterations During Colorectal-tumor Development. N Engl J Med. 1988;

48. Pędziwiatr M, Mizera M, Witowski J, Major P, Torbicz G, Gajewska N, et al. Primary tumor resection in stage IV unresectable colorectal cancer: what has changed? Medical Oncology. 2017.

49. Kwong LN, Dove WF. APC and Its Modifiers in Colon Cancer. In 2010.

50. Mah AT, Yan KS, Kuo CJ. Wnt pathway regulation of intestinal stem cells. Journal of Physiology. 2016.

51. Perochon J, Carroll LR, Cordero JB. Wnt signalling in intestinal stem cells: Lessons from mice and flies. Genes. 2018.

52. Schneikert J, Behrens J. The canonical Wnt signalling pathway and its APC partner in colon cancer development. Gut. 2007.

53. Huelsken J, Behrens J. The Wnt signalling pathway. J Cell Sci. 2002;

54. Polakis P. Wnt signaling and cancer. Genes and Development. 2000.

55. Benchabane H, Ahmed Y. The Adenomatous Polyposis Coli Tumor Suppressor and Wnt Signaling in the Regulation of Apoptosis. 2010;1:75–84.

56. Vanuytsel T, Senger S, Fasano A, Shea-Donohue T. Major signaling pathways in intestinal stem cells. Biochimica et Biophysica Acta - General Subjects. 2013.

57. Valenta T, Hausmann G, Basler K. The many faces and functions of Î 2-catenin. EMBO Journal. 2012.

58. Fevr T, Robine S, Louvard D, Huelsken J. Wnt/ -Catenin Is Essential for Intestinal Homeostasis and Maintenance of Intestinal Stem Cells. Mol Cell Biol. 2007;

59. Lüchtenborg M, Weijenberg MP, Roemen GMJM, de Bruïne AP, van den Brandt PA, Lentjes MHFM, et al. APC mutations in sporadic coloretal carcinomas from The Netherlands Cohort Study. Carcinogenesis. 2004.

60. Wong SCC, Lo ESF, Chan AKC, Lee KC, Hsiao WL. Nuclear β catenin as a potential prognostic and diagnostic marker in patients with colorectal cancer from Hong Kong. Journal of Clinical Pathology - Molecular Pathology. 2003;

61. Tan C, Du X. KRAS mutation testing in metastatic colorectal cancer. World Journal of Gastroenterology. 2012.

62. Liu X, Yan S, Zhou T, Terada Y, Erikson RL. The MAP kinase pathway is required for entry into mitosis and cell survival. Oncogene. 2004;

63. Morrison DK. MAP kinase pathways. Cold Spring Harb Perspect Biol. 2012;

64. Sanz-Garcia E, Argiles G, Elez E, Tabernero J. BRAF mutant colorectal cancer: Prognosis, treatment, and new perspectives. Annals of Oncology. 2017.

65. Barras D. BRAF Mutation in Colorectal Cancer: An Update . Biomark Cancer. 2015;

66. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, et al. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. New England Journal of Medicine. 2011;

67. Salovaara R, Roth S, Loukola A, Launonen V, Sistonen P, Avizienyte E, et al. Frequent loss of SMAD4/DPC4 protein in colorectal cancers. Gut. 2002;51(1):56–9.

68. Iacopetta B. TP53 mutation in colorectal cancer. Hum Mutat. 2003;21(3):271–6.

69. Willett CG, Chang DT, Czito BG, Meyer J, Wo J. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012. (5). International Journal of Radiation Oncology Biology Physics. 2013.

70. Frey P, Devisme A, Rose K, Schrempp M, Freihen V, Andrieux G, et al. SMAD4 mutations do not preclude epithelial–mesenchymal transition in colorectal cancer. Oncogene. 2022;

71. Zhao M, Mishra L, Deng CX. The role of TGF-β/SMAD4 signaling in cancer. Int J Biol Sci. 2018;

72. Neel JC, Humbert L, Lebrun JJ. The Dual Role of TGFβ in Human Cancer: From Tumor Suppression to Cancer Metastasis. ISRN Mol Biol. 2012;

73. Chernov M V., Stark GR, Agarwal ML, Taylor WR, Chernova OB. The p53 Network. Journal of Biological Chemistry. 2002;273(1):1–4.

74. Lane DP. p53, guardian of the genome. Nature. 1992.

75. Lin D, Fiscella M, O'Connor PM, Jackman J, Chen M, Luo LL, et al. Constitutive expression of B-myb can bypass p53-induced Waf1/Cip1-mediated G1 arrest. Proceedings of the National Academy of Sciences. 1994;91(21):10079–83.

76. Ljungman M, Zhang F, Chen F, Rainbow AJ, McKay BC. Inhibition of RNA polymerase II as a trigger for the p53 response. Oncogene. 1999;

77. Williams AB, Schumacher B. p53 in the DNA-damage-repair process. Cold Spring Harb Perspect Med. 2016;

78. Chen J. The cell-cycle arrest and apoptotic functions of p53 in tumor initiation and progression. Cold Spring Harb Perspect Med. 2016;

79. Luo Y, Hurwitz J, Massagué J. Cell-cycle inhibition by independent CDK and PCNA binding domains in p21cip1. Nature. 1995.

80. Amaral JD, Xavier JM, Steer CJ, Rodrigues CM. The role of p53 in apoptosis. Discovery medicine. 2010.

81. Kim JM. Involvement of the Fas/Fas Ligand System in p53-Mediated Granulosa Cell Apoptosis during Follicular Development and Atresia. Endocrinology. 1999;

82. Li XL, Zhou J, Chen ZR, Chng WJ. P53 mutations in colorectal cancer- Molecular pathogenesis and pharmacological reactivation. World J Gastroenterol. 2015;

83. López I, Oliveira LL, Tucci P, Álvarez-Valín F, A. Coudry R, Marín M. Different mutation profiles associated to P53 accumulation in colorectal cancer. Gene. 2012;

84. Riihimaki M, Hemminki A, Sundquist J, Hemminki K. Patterns of metastasis in colon and rectal cancer. Sci Rep. 2016;

85. van der Geest LGM, Lam-Boer J, Koopman M, Verhoef C, Elferink MAG, de Wilt JHW. Nationwide trends in incidence, treatment and survival of colorectal cancer patients with synchronous metastases. Clin Exp Metastasis. 2015;

86. Simon K. Colorectal cancer development and advances in screening. Clinical Interventions in Aging. 2016.

87. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000.

88. Zhu L, Li X, Yuan Y, Dong C, Yang M. APC Promoter Methylation in Gastrointestinal Cancer. Vol. 11, Frontiers in Oncology. 2021.

89. Li X, Yao X, Wang Y, Hu F, Wang F, Jiang L, et al. MLH1 Promoter Methylation Frequency in Colorectal Cancer Patients and Related Clinicopathological and Molecular Features. PLoS ONE. 2013.

90. Satorres C, García-Campos M, Bustamante-Balén M. Molecular Features of the Serrated Pathway to Colorectal Cancer: Current Knowledge and Future Directions. Gut Liver. 2021;15(1).

91. De Palma FDE, D'argenio V, Pol J, Kroemer G, Maiuri MC, Salvatore F. The molecular hallmarks of the serrated pathway in colorectal cancer. Vol. 11, Cancers. 2019.

92. Jass JR, Smith M. Sialic acid and epithelial differentiation in colorectal polyps and cancer — a morphological, mucin and lectin histochemical study. Pathology. 1992;24(4).

93. Murcia O, Juárez M, Hernández-Illán E, Egoavil C, Giner-Calabuig M, Rodríguez-Soler M, et al. Serrated colorectal cancer: Molecular classification, prognosis, and response to chemotherapy. Vol. 22, World Journal of Gastroenterology. 2016.

94. Haque T, Greene KG, Crockett SD. Serrated neoplasia of the colon: What do we really know? Curr Gastroenterol Rep. 2014;16(4).

95. Boland CR, Goel A. Microsatellite Instability in Colorectal Cancer. Gastroenterology. 2010;

96. Benatti P, Gafà R, Barana D, Marino M, Scarselli A, Pedroni M, et al. Microsatellite instability and colorectal cancer prognosis. Clinical Cancer Research. 2005;

97. Itatani Y, Kawada K, Sakai Y. Transforming growth factor-β signaling pathway in colorectal cancer and its tumor microenvironment. Vol. 20, International Journal of Molecular Sciences. 2019.

98. Murakami T, Mitomi H, Saito T, Takahashi M, Sakamoto N, Fukui N, et al. Distinct WNT/β-catenin signaling activation in the serrated neoplasia pathway and the adenoma-carcinoma sequence of the colorectum. Modern Pathology. 2015;28(1).

99. Issa JP. CpG island methylator phenotype in cancer. Nature Reviews Cancer. 2004.

100. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JPJ. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci U S A. 1999;

101. Hughes LAE, Melotte V, Schrijver J De, Maat M De, Smit VTHBM, Bovee JVMG, et al. The CpG island methylator phenotype: What's in a name? Vol. 73, Cancer Research. 2013.

102. Pino MS, Chung DC. The Chromosomal Instability Pathway in Colon Cancer. Gastroenterology. 2010;

103. Árnadóttir SS, Jeppesen M, Lamy P, Bramsen JB, Nordentoft I, Knudsen M, et al. Characterization of genetic intratumor heterogeneity in colorectal cancer and matching patient-derived spheroid cultures. Mol Oncol. 2018;

104. Giaretti W, Macciocu B, Geido E, Hermsen MAJA, Postma C, Baak JPA, et al. Intratumor heterogeneity of k-ras and p53 mutations among human colorectal adenomas containing early cancer. Analytical Cellular Pathology. 2000;

105. Stanta G, Bonin S. Overview on Clinical Relevance of Intra-Tumor Heterogeneity. Front Med (Lausanne). 2018;

106. O'Connor JPB, Rose CJ, Waterton JC, Carano RAD, Parker GJM, Jackson A. Imaging intratumor heterogeneity: Role in therapy response, resistance, and clinical outcome. Clinical Cancer Research. 2015.

107. Greaves M. Evolutionary determinants of cancer. Cancer Discovery. 2015.

108. Losi L, Baisse B, Bouzourene H, Benhattar J. Evolution of intratumoral genetic heterogeneity during colorectal cancer progression. Carcinogenesis. 2005;

109. Lugli A, Roberts DE, Zlobec I, Blank A, Dawson H. Tumor Heterogeneity in Primary Colorectal Cancer and Corresponding Metastases. Does the Apple Fall Far From the Tree? Front Med (Lausanne). 2018;5(August):1–8.

110. Saito T, Niida A, Uchi R, Hirata H, Komatsu H, Sakimura S, et al. A temporal shift of the evolutionary principle shaping intratumor heterogeneity in colorectal cancer. Nat Commun. 2018;

111. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A big bang model of human colorectal tumor growth. Nat Genet. 2015;

112. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Swanton C. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. New England Journal of Medicine. 2012;

113. Cross WCH, Graham TA, Wright NA. New paradigms in clonal evolution: punctuated equilibrium in cancer. Journal of Pathology. 2016.

114. Joung JG, Oh BY, Hong HK, Al-Khalidi H, Al-Alem F, Lee HO, et al. Tumor heterogeneity predicts metastatic potential in colorectal cancer. Clinical Cancer Research. 2017;

115. van Ginkel J, Tomlinson IPM, Soriano I. The Evolutionary Landscape of Colorectal Tumorigenesis: Recent Paradigms, Models, and Hypotheses. Gas. 2023;1–6.

116. Stoffel EM, Yurgelun MB. Genetic predisposition to colorectal cancer: Implications for treatment and prevention. Seminars in Oncology. 2016.

117. Wang J, Carvajal-Carmona LG, Chu JH, Zauber AG, Kubo M, Matsuda K, et al. Germline variants and advanced colorectal adenomas: Adenoma prevention with celecoxib trial genome-wide association study. Clinical Cancer Research. 2013;

118. Leoz ML, Carballal S, Moreira L, Ocaña T, Balaguer F. The genetic basis of familial adenomatous polyposis and its implications for clinical practice and risk management. Application of Clinical Genetics. 2015.

119. Galiatsatos P, Foulkes WD. Familial adenomatous polyposis. American Journal of Gastroenterology. 2006.

120. Dolan S. Familial adenomatous polyposis: Development, presentation, and treatment strategies. Clin J Oncol Nurs. 2019;

121. Dinarvand P, Davaro EP, Doan J V., Ising ME, Evans NR, Phillips NJ, et al. Familial adenomatous polyposis syndrome an update and review of extraintestinal manifestations. Archives of Pathology and Laboratory Medicine. 2019.

122. Vasen HFA, Möslein G, Alonso A, Aretz S, Bernstein I, Bertario L, et al. Guidelines for the clinical management of familial adenomatous polyposis (FAP). In: Gut. 2008.

123. Lynch HT, Shaw MW, Magnuson CW, Larsen AL, Krush AJ. Hereditary Factors in Cancer: Study of Two Large Midwestern Kindreds. Arch Intern Med. 1966;

124. Hampel H, Frankel WL, Martin E, Arnold M, Khanduja K, Kuebler P, et al. Screening for the Lynch Syndrome (Hereditary Nonpolyposis Colorectal Cancer). New England Journal of Medicine. 2005;

125. Haraldsdottir S, Rafnar T, Frankel WL, Einarsdottir S, Sigurdsson A, Hampel H, et al. Comprehensive population-wide analysis of Lynch syndrome in Iceland reveals founder mutations in MSH6 and PMS2. Nat Commun. 2017;

126. Giardiello FM, Allen JI, Axilbund JE, Boland CR, Burke CA, Burt RW, et al. Guidelines on genetic evaluation and management of Lynch syndrome: A consensus statement by the U.S. Multi-Society Task Force on Colorectal Cancer. Gastrointest Endosc. 2014;

127. Li GM. Mechanisms and functions of DNA mismatch repair. Vol. 18, Cell Research. 2008.

128. Sameer AS. Colorectal Cancer: Molecular Mutations and Polymorphisms. Front Oncol. 2013;

129. Liu T, Wahlberg S, Burek E, Lindblom P, Rubio C, Lindblom A. Microsatellite instability as a predictor of a mutation in a DNA mismatch repair gene in familial colorectal cancer. Genes Chromosomes Cancer. 2000;

130. Lipton LR, Johnson V, Cummings C, Fisher S, Risby P, Eftekhar Sadat AT, et al. Refining the Amsterdam criteria and Bethesda guidelines: Testing algorithms for the prediction of mismatch repair mutation status in the familial cancer clinic. Journal of Clinical Oncology. 2004;

131. Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Rüschoff J, et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst. 2004;

132. Vasen HFA, Möslein G, Alonso A, Bernstein I, Bertario L, Blanco I, et al. Guidelines for the clinical management of Lynch syndrome (hereditary non-polyposis cancer). Journal of Medical Genetics. 2007.

133. Idos GE, Valle L. Gene Reviews (Internet) [Internet]. 2021. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1211/#hnpcc.Clinical_Characteristics

134. Edelstein DL, Axilbund J, Baxter M, Hylind LM, Romans K, Griffin CA, et al. Rapid Development of Colorectal Neoplasia in Patients With Lynch Syndrome. Clinical Gastroenterology and Hepatology. 2011;

135. Lynch HT, Lynch J. Genetics, natural history, surveillance, management, and gene mapping in the Lynch syndrome. Pathologie-biologie. 1995.

136. Smyrk TC, Watson P, Kaul K, Lynch HT. Tumor-infiltrating lymphocytes are a marker for microsatellite instability in colorectal carcinoma. Cancer. 2001;

137. Llosa NJ, Cruise M, Tam A, Wicks EC, Hechenbleikner EM, Taube JM, et al. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. Cancer Discov. 2015;

138. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. New England Journal of Medicine. 2015;

139. Overman MJ, Lonardi S, Wong KYM, Lenz HJ, Gelsomino F, Aglietta M, et al. Durable clinical benefit with nivolumab plus ipilimumab in DNA mismatch repair-deficient/microsatellite instability-high metastatic colorectal cancer. Journal of Clinical Oncology. 2018;

140. Beggs AD, Latchford AR, Vasen HFA, Moslein G, Alonso A, Aretz S, et al. Peutz - Jeghers syndrome: A systematic review and recommendations for management. Gut. 2010.

141. Peutz JLA. Very remarkable case of familial polyposis of mucous membrane of intestinal tract and nasopharynx accompanied by peculiar pigmentations of skin and mucous membrane. Nederl Maandschr Geneesk. 1921;

142. BRUWER A, BARGEN JA, KIERLAND RR. Surface pigmentation and generalized intestinal polyposis; (Peutz-Jeghers syndrome). Proc Staff Meet Mayo Clin. 1954;

143. Jeghers H, McKusick VA, Katz KH. Generalized intestinal polyposis and melanin spots of the oral mucosa, lips and digits: a syndrome of diagnostic significance. New England Journal of Medicine. 1949;

144. Giardiello FM, Trimbath JD. Peutz-Jeghers Syndrome and Management Recommendations. Clinical Gastroenterology and Hepatology. 2006.

145. Utsunomiya J, Gocho H, Miyanaga T, Hamaguchi E, Kashimure A. Peutz-Jeghers syndrome: its natural course and management. Johns Hopkins Medical Journal. 1975;

146. Zyla RE, Hahn E, Hodgson A. Gene of the month: STK11. Journal of clinical pathology. 2021.

147. Alessi DR, Sakamoto K, Bayascas JR. LKB1-dependent signaling pathways. Annual Review of Biochemistry. 2006.

148. Slattery ML, Herrick JS, Lundgreen A, Fitzpatrick FA, Curtin K, Wolff RK. Genetic variation in a metabolic signaling pathway and colon and rectal cancer risk: mTOR, PTEN, STK11, RPKAA1, PRKAG2, TSC1, TSC2, PI3K and Akt1. Carcinogenesis. 2010;

149. Zhou W, Marcus AI, Vertino PM. Dysregulation of mTOR activity through LKB1 inactivation. Chin J Cancer. 2013;

150. Shaw RJ, Bardeesy N, Manning BD, Lopez L, Kosmatka M, DePinho RA, et al. The LKB1 tumor suppressor negatively regulates mTOR signaling. Cancer Cell. 2004;

151. Tsang F, Mallappa S, Teh W, Clark S. PB.46: Peutz-Jeghers Syndrome and carcinoma of the breast: call for new breast imaging surveillance guidelines. Breast Cancer Research. 2013;

152. Fostira F, Fountzilas E, Papadopoulou K, Karaiskos T, Mpatsi O, Pastelli N, et al. Lung cancer as a predominant feature in a patient with Peutz-Jeghers syndrome. Thorac Cancer. 2022;13(12):1862–5.

153. Korsse SE, Harinck F, van Lier MGF, Biermann K, Offerhaus GJA, Krak N, et al. Pancreatic cancer risk in Peutz-Jeghers syndrome patients: A large cohort study and implications for surveillance. J Med Genet. 2013;

154. Giardiello FM, Brensinger JD, Tersmette AC, Goodman SN, Petersen GM, Booker S V., et al. Very high risk of cancer in familial Peutz-Jeghers syndrome. Gastroenterology. 2000;

155. Church JM. Polymerase proofreading-associated polyposis: A new, dominantly inherited syndrome of hereditary colorectal cancer predisposition. Dis Colon Rectum. 2014;

156. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat Genet. 2013;

157. Briggs S, Tomlinson I. Germline and somatic polymerase ε and δ mutations define a new class of hypermutated colorectal and endometrial cancers. Journal of Pathology. 2013;

158. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, et al. Recurrent R-spondin fusions in colon cancer. Nature. 2012;

159. Palles C, Martin L, Domingo E, Chegwidden L, McGuire J, Cuthill V, et al. The clinical features of polymerase proof-reading associated polyposis (PPAP) and recommendations for patient management. Fam Cancer. 2022;

160. Mccoll I, Busxey HC, Veale AC, Morson BC. JUVENILE POLYPOSIS COLI. Proc R Soc Med. 1964;57:896–7.

161. Gao XH, Li J, Zhao ZY, Xu XD, Du YQ, Yan HL, et al. Juvenile polyposis syndrome might be misdiagnosed as familial adenomatous polyposis: A case report and literature review. BMC Gastroenterol. 2020;

162. Guillén-Ponce C, Serrano R, Sánchez-Heras AB, Teulé A, Chirivella I, Martín T, et al. Clinical guideline seom: hereditary colorectal cancer. Clinical and Translational Oncology. 2015;

163. Pérez-Castilla A, Peñailillo P, Oksenberg D. Juvenile polyposis syndrome: A case report. Int J Surg Case Rep. 2019;

164. Lieberman S, Walsh T, Schechter M, Adar T, Goldin E, Beeri R, et al. Features of Patients With Hereditary Mixed Polyposis Syndrome Caused by Duplication of GREM1 and Implications for Screening and Surveillance. Gastroenterology. 2017;

165. Aretz S. [The differential diagnosis and surveillance of hereditary gastrointestinal polyposis syndromes]. Dtsch Arztebl Int. 2010;

166. Toboeva MK, Shelygin YA, Frolov SA, Kuzminov MA, Tsukanov AS. MutYH-associated polyposis. Ter Arkh. 2019;

167. Tronick E, Hunter RG. Waddington, Dynamic systems, and epigenetics. Front Behav Neurosci. 2016;10(JUN).

168. Noble D. Conrad Waddington and the origin of epigenetics. Vol. 218, Journal of Experimental Biology. 2015.

169. Waddington CH. The epigenotype. 1942. Int J Epidemiol. 2012;41(1).

170. Wu CT, Morris JR. Genes, genetics, and epigenetics: A correspondence. Vol. 293, Science. 2001.

171. Moosavi A, Ardekani AM. Role of epigenetics in biology and human diseases. Vol. 20, Iranian Biomedical Journal. 2016.

172. Al Aboud NM, Jialal I. Genetics, Epigenetic Mechanism. StatPearls. 2018.

173. Harvey L, Arnold B, S LZ, Paul M, David B, James D. Molecular Cell Biology. 4th edition. Molecular Cell Biology. 4th edition. 2000.

174. Bommarito PA, Fry RC. The role of DNA methylation in gene regulation. In: Toxicoepigenetics: Core Principles and Applications. 2018.

175. Gujar H, Weisenberger DJ, Liang G. The roles of human DNA methyltransferases and their isoforms in shaping the epigenome. Genes. 2019.

176. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;

177. Robertson KD, Uzvolgyi E, Liang G, Talmadge C, Sumegi J, Gonzales FA, et al. The human DNA methyltransferases (DNMTs) 1, 3a and 3b: Coordinate mRNA expression in normal tissues and overexpression in tumors. Nucleic Acids Res. 1999;

178. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992;

179. Miller JL, Grant PA. The role of DNA methylation and histone modifications in transcriptional regulation in humans. Subcell Biochem. 2013;

180. Smith ZD, Meissner A. DNA methylation: Roles in mammalian development. Nature Reviews Genetics. 2013.

181. Wu G, Schöler HR. Role of Oct4 in the early embryo development. Vol. 3, Cell Regeneration. 2014.

182. Feldman N, Gerson A, Fang J, Li E, Zhang Y, Shinkai Y, et al. G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. Nat Cell Biol. 2006;8(2).

183. Yeo S, Jeong S, Kim J, Han JS, Han YM, Kang YK. Characterization of DNA methylation change in stem cell marker genes during differentiation of human embryonic stem cells. Biochem Biophys Res Commun. 2007;359(3).

184. Hackett JA, Reddington JP, Nestor CE, Dunican DS, Branco MR, Reichmann J, et al. Promoter DNA methylation couples genome-defence mechanisms to epigenetic reprogramming in the mouse germline. Development. 2012;139(19):3623–32.

185. Martin Caballero I, Hansen J, Leaford D, Pollard S, Hendrich BD. The methyl-CpG binding proteins Mecp2, Mbd2 and Kaiso are dispensable for mouse embryogenesis, but playa redundant function in neural differentiation. PLoS One. 2009;4(1).

186. Jones PL, Veenstra GJC, Wade PA, Vermaak D, Kass SU, Landsberger N, et al. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet. 1998;19(2).

187. Nan X, Ng HH, Johnson CA, Laherty CD, Turner BM, Eisenman RN, et al. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature. 1998;393(6683).

188. Schmolka N, Karemaker ID, Cardoso da Silva R, Recchia DC, Spegg V, Bhaskaran J, et al. Dissecting the roles of MBD2 isoforms and domains in regulating NuRD complex function during cellular differentiation. Nat Commun. 2023;14(1).

189. Hellman A, Chess A. Gene body-specific methylation on the active X chromosome. Science (1979). 2007;

190. Kulis M, Queirós AC, Beekman R, Martín-Subero JI. Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. Biochimica et Biophysica Acta - Gene Regulatory Mechanisms. 2013.

191. Du Q, Bert SA, Armstrong NJ, Caldon CE, Song JZ, Nair SS, et al. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. Nat Commun. 2019;

192. Rivera-Mulia JC, Buckley Q, Sasaki T, Zimmerman J, Didier RA, Nazor K, et al. Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. Genome Res. 2015;

193. Skvortsova K, Stirzaker C, Taberlay P. The DNA methylation landscape in cancer. Essays in Biochemistry. 2019.

194. Ehrlich M, Lacey M. DNA hypomethylation and hemimethylation in cancer. Advances in Experimental Medicine and Biology. 2013.

195. Ashktorab H, Brim H. DNA Methylation and Colorectal Cancer.

196. Esteller M, Silva JM, Dominguez G, Bonilla F, Matias-Guiu X, Lerma E, et al. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. J Natl Cancer Inst. 2000;

197. Liang TJ, Wang HX, Zheng YY, Cao YQ, Wu X, Zhou X, et al. APC hypermethylation for early diagnosis of colorectal cancer: A meta-analysis and literature review. Oncotarget. 2017;

198. Kaneko E, Sato N, Sugawara T, Noto A, Takahashi K, Makino K, et al. MLH1 promoter hypermethylation predicts poorer prognosis in mismatch repair deficiency endometrial carcinomas. J Gynecol Oncol. 2021;

199. Neri F, Dettori D, Incarnato D, Krepelova A, Rapelli S, Maldotti M, et al. TET1 is a tumour suppressor that inhibits colon cancer growth by derepressing inhibitors of the WNT pathway. Oncogene. 2015;

200. He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science (1979). 2011;

201. Rasmussen SL, Krarup HB, Sunesen KG, Pedersen IS, Madsen PH, Ussing OT. Hypermethylated DNA as a biomarker for colorectal cancer: A systematic review. Colorectal Disease. 2016.

202. Pack SC, Kim HR, Lim SW, Kim HY, Ko JY, Lee KS, et al. Usefulness of plasma epigenetic changes of five major genes involved in the pathogenesis of colorectal cancer. Int J Colorectal Dis. 2013;

203. Leung WK, To KF, Man EPS, Chan MWY, Bai AHC, Hui AJ, et al. Quantitative detection of promoter hypermethylation in multiple genes in the serum of patients with colorectal cancer. American Journal of Gastroenterology. 2005;

204. Yang X, Han H, DeCarvalho DD, Lay FD, Jones PA, Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer Cell. 2014;

205. Mossman D, Kim KT, Scott RJ. Demethylation by 5-aza-2'-deoxycytidine in colorectal cancer cells targets genomic DNA whilst promoter CpG island methylation persists. BMC Cancer. 2010;

206. Christman JK. 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation: Mechanistic studies and their implications for cancer therapy. Oncogene. 2002.

207. Juttermann R, Li E, Jaenisch R. Toxicity of 5-aza-2'-deoxycytidine to mammalian cells is mediated primarily by covalent trapping of DNA methyltransferase rather than DNA demethylation. Proc Natl Acad Sci U S A. 1994;

208. Lavelle D, Saunthararajah Y, Desimone J. DNA methylation and mechanism of action of 5-azacytidine. Vol. 111, Blood. 2008.

209. Momparler RL. Pharmacology of 5-aza-2'-deoxycytidine (decitabine). Semin Hematol. 2005;42(SUPPL. 2).

210. Miller BF, Sánchez-Vega F, Elnitski L. The emergence of pan-cancer CIMP and its elusive interpretation. Vol. 6, Biomolecules. 2016.

211. Jeong MB, Kim JH, Kang GH. Epigenetic alterations in colorectal cancer: The CpG island methylator phenotype. Histology and Histopathology. 2013.

212. Samowitz WS, Albertsen H, Herrick J, Levin TR, Sweeney C, Murtaugh MA, et al. Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. Gastroenterology. 2005;

213. Shen L, Toyota M, Kondo Y, Lin E, Zhang L, Guo Y, et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Proc Natl Acad Sci U S A. 2007;

214. Paweł K, Małgorzata SM. CpG Island Methylator Phenotype—A Hope for the Future or a Road to Nowhere? Vol. 23, International Journal of Molecular Sciences. 2022.

215. Juo YY, Johnston FM, Zhang DY, Juo HH, Wang H, Pappou EP, et al. Prognostic value of CpG island methylator phenotype among colorectal cancer patients: A systematic review and meta-analysis. Annals of Oncology. 2014.

216. Wang Y, Long Y, Xu Y, Guan Z, Lian P, Peng J, et al. Prognostic and predictive value of CpG island methylator phenotype in patients with locally advanced nonmetastatic sporadic colorectal cancer. Gastroenterol Res Pract. 2014;

217. Rhee YY, Kim KJ, Kang GH. CpG Island methylator phenotype-high colorectal cancers and their prognostic implications and relationships with the serrated Neoplasia pathway. Gut and Liver. 2017.

218. Advani SM, Advani PS, Brown DW, Desantis SM, Korphaisarn K, Vonville HM, et al. Global differences in the prevalence of the CpG island methylator phenotype of colorectal cancer. BMC Cancer. 2019;19(1).

219. Fang M, Ou J, Hutchinson L, Green MR. The BRAF Oncoprotein Functions through the Transcriptional Repressor MAFG to Mediate the CpG Island Methylator Phenotype. Mol Cell. 2014;

220. Serra RW, Fang M, Park SM, Hutchinson L, Green MR. A KRAS-directed transcriptional silencing pathway that mediates the CpG island methylator phenotype. Elife. 2014;2014(3).

221. Palles C, West HD, Chew E, Galavotti S, Flensburg C, Grolleman JE, et al. Germline MBD4 deficiency causes a multi-tumor predisposition syndrome. The American Journal of Human Genetics. 2022;109:953–60.

222. Rahman N. Realizing the promise of cancer predisposition genes. Nature. 2014.

223. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol. 2012;

224. Brodie A, Azaria JR, Ofran Y. How far from the SNP may the causative genes be? Nucleic Acids Res. 2016;

225. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Reviews Genetics. 2010.

226. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;

227. Nica AC, Dermitzakis ET. Expression quantitative trait loci: Present and future. Philosophical Transactions of the Royal Society B: Biological Sciences. 2013.

228. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Human Molecular Genetics. 2015.

229. Bergen SE, Petryshen TL. Genome-wide association studies of schizophrenia: Does bigger lead to better results? Current Opinion in Psychiatry. 2012.

230. Hysi PG, Valdes AM, Liu F, Furlotte NA, Evans DM, Bataille V, et al. Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. Nat Genet. 2018;

231. Lin YJ, Liao WL, Wang CH, Tsai LP, Tang CH, Chen CH, et al. Association of human height-related genetic variants with familial short stature in Han Chinese in Taiwan. Sci Rep. 2017;

232. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet. 2009;

233. Tomlinson IPM, Carvajal-Carmona LG, Dobbins SE, Tenesa A, Jones AM, Howarth K, et al. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. PLoS Genet. 2011;

234. Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes J, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. Nat Commun. 2019;

235. Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nat Genet. 2010;

236. Wang H, Schmit SL, Haiman CA, Keku TO, Kato I, Palmer JR, et al. Novel colon cancer susceptibility variants identified from a genome-wide association study in African Americans. Int J Cancer. 2017;

237. Armaghany T, Wilson JD, Chu Q, Mills G. Genetic alterations in colorectal cancer. Gastrointestinal Cancer Research. 2012;5(1):19–27.

238. Jasperson KW, Tuohy TM, Neklason DW, Burt RW. Hereditary and Familial Colon Cancer. Gastroenterology. 2010;

239. Schumacher FR, Schmit SL, Jiao S, Edlund CK, Wang H, Zhang B, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat Commun. 2015;

240. Tomlinson IPM, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet. 2008;

241. Tenesa A, Farrington SM, Prendergast JGD, Porteous ME, Walker M, Haq N, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet. 2008;

242. Le Marchand L. Genome-Wide Association Studies and Colorectal Cancer. Surgical Oncology Clinics of North America. 2009.

243. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet. 2007;

244. Jiang K, Sun Y, Wang C, Ji J, Li Y, Ye Y, et al. Genome-wide association study identifies two new susceptibility loci for colorectal cancer at 5q23.3 and 17q12 in Han Chinese. Oncotarget. 2015;

245. Wang H, Haiman CA, Burnett T, Fortini BK, Kolonel LN, Henderson BE, et al. Fine-mapping of genome-wide association study-identified risk loci for colorectal cancer in African Americans. Hum Mol Genet. 2013;

246. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet. 2013;

247. Guo X, Long J, Zeng C, Michailidou K, Ghoussaini M, Bolla MK, et al. Fine-scale mapping of the 4q24 locus identifies & pr two Independent loci associated with breast cancer risk. Cancer Epidemiology Biomarkers and Prevention. 2015;

248. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. Am J Hum Genet. 2002;

249. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nature Reviews Genetics. 2003.

250. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics. 2018.

251. Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, et al. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. PLoS Genet. 2014;

252. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nature Reviews Genetics. 2009.

253. Sillanpää MJ, Bhattacharjee M. Bayesian association-based fine mapping in small chromosomal segments. Genetics. 2005;

254. Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet. 2012;

255. Porcu E, Rüeger S, Lepik K, Agbessi M, Ahsan H, Alves I, et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nat Commun. 2019;

256. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;

257. Fernandez-Rozadilla C, Timofeeva M, Chen Z, Law PJ, Thomas M, Schmit SL, et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. Nat Genet. 2022;

258. Zhu H, Zhou X. Transcriptome-wide association studies: a view from Mendelian randomization. Quantitative Biology. 2020.

259. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;

260. Gala H, Tomlinson I. The use of Mendelian randomisation to identify causal cancer risk factors: promise and limitations. J Pathol. 2020;

261. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2023;

262. Mahajan A, Spracklen CN, Zhang W, Ng MCY, Petty LE, Kitajima H, et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. Nat Genet. 2022;

263. Polushina T, Giddaluru S, Bettella F, Espeseth T, Lundervold AJ, Djurovic S, et al. Analysis of the joint effect of SNPs to identify independent loci and allelic heterogeneity in schizophrenia GWAS data. Transl Psychiatry. 2017;

264. Steimle J, Weibel N, Olberding S, Mühlhäuser M, Hollan JD. PLink. In 2011.

265. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang H, et al. A global reference for human genetic variation. Nature. 2015;

266. Consortium EP, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C a, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2013;

267. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;

268. https://grch37.ensembl.org/info/data/biomart/index.html.

269. Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Aken B, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;

270. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;

271. Dong X, Su YR, Barfield R, Bien SA, He Q, Harrison TA, et al. A general framework for functionally informed set-based analysis: Application to a large-scale colorectal cancer study. PLoS Genet. 2020;

272. Seshan VE. clinfun: Clinical Trial Design and Data Analysis Functions. R package version 105. 2013;

273. Pomerantz MM, Werner L, Xie W, Regan MM, Lee GSM, Sun T, et al. Association of prostate cancer risk loci with disease aggressiveness and prostate cancer-specific mortality. Cancer Prevention Research. 2011;

274. Hoffmann TJ, Van Den Eeden SK, Sakoda LC, Jorgenson E, Habel LA, Graff RE, et al. A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. Cancer Discov. 2015;

275. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, et al. Discovery of common and rare genetic risk variants for colorectal cancer. Nat Genet. 2019;

276. Liang Y, Yang Z, Zhong R. Primary biliary cirrhosis and cancer risk: A systematic review and meta-analysis. Hepatology. 2012;

277. Cheng T, Thompson D, Painter J, O'Mara T, Gorman M, Martin L, et al. Meta-analysis of genome-wide association studies identifies common susceptibility polymorphisms for colorectal and endometrial cancer near SH2B3 and TSHZ. Sci Rep. 2015;5:17369.

278. Brandes N, Linial N, Linial M. Genetic association studies of alterations in protein function expose recessive effects on cancer predisposition. Sci Rep. 2021;

279. Kar SP, Beesley J, Al Olama AA, Michailidou K, Tyrer J, Kote-Jarai ZS, et al. Genome-wide meta-analyses of breast, ovarian, and prostate cancer association studies identify multiple new susceptibility loci shared by at least two cancer types. Cancer Discov. 2016;

280. Hu L, Li Z, Cheng J, Rao Q, Gong W, Liu M, et al. Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation. Cell. 2013;

281. Chou WC, Chou SC, Liu CY, Chen CY, Hou HA, Kuo YY, et al. TET2 mutation is an unfavorable prognostic factor in acute myeloid leukemia patients with intermediate-risk cytogenetics. Blood. 2011;

282. Knight J, Spain SL, Capon F, Hayday A, Nestle FO, Clop A, et al. Conditional analysis identifies three novel major histocompatibility complex loci associated with psoriasis. Hum Mol Genet. 2012;

283. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014;

284. Broekema R V., Bakker OB, Jonkers IH. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. Open Biol. 2020;

285. Lu Y, Kweon SS, Tanikawa C, Jia WH, Xiang YB, Cai Q, et al. Large-Scale Genome-Wide Association Study of East Asians Identifies Loci Associated With Risk for Colorectal Cancer. Gastroenterology. 2019;

286. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;

287. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res. 2014;

288. Stadhouders R, Kolovos P, Brouwer R, Zuin J, Van Den Heuvel A, Kockx C, et al. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nat Protoc. 2013;

289. Baxter JS, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V, et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. Nat Commun. 2018;

290. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: A comprehensive technique to capture the conformation of genomes. Methods. 2012;

291. Guil S, Esteller M. Cis-acting noncoding RNAs: Friends and foes. Nature Structural and Molecular Biology. 2012.

292. Jankowska AM, Szpurka H, Tiu R V., Makishima H, Afable M, Huh J, et al. Loss of heterozygosity 4q24 and TET2 mutations associated with myelodysplastic/myeloproliferative neoplasms. Blood. 2009;

293. Wu X, Zhang Y. TET-mediated active DNA demethylation: Mechanism, function and beyond. Nature Reviews Genetics. 2017.

294. Huang Y, Wang G, Liang Z, Yang Y, Cui L, Liu CY. Loss of nuclear localization of TET2 in colorectal cancer. Clinical Epigenetics. 2016.

295. Cornish AJ, Gruber AJ, Kinnersley B, Chubb D, Frangou A, Caravanga G, et al. Whole Genome Sequencing of 2,023 Colorectal Cancers Reveals Mutational Landscapes, New Driver Genes and Immune Interactions. bioRxiv. 2022;

296. Moran-Crusio K, Reavie L, Shih A, Abdel-Wahab O, Ndiaye-Lobry D, Lobry C, et al. Tet2 Loss Leads to Increased Hematopoietic Stem Cell Self-Renewal and Myeloid Transformation. Cancer Cell. 2011;

297. Cimmino L, Dolgalev I, Wang Y, Yoshimi A, Martin GH, Wang J, et al. Restoration of TET2 Function Blocks Aberrant Self-Renewal and Leukemia Progression. Cell. 2017;

298. Kim MR, Wu MJ, Zhang Y, Yang JY, Chang CJ. TET2 directs mammary luminal cell differentiation and endocrine response. Nat Commun. 2020;

299. Solary E, Bernard OA, Tefferi A, Fuks F, Vainchenker W. The Ten-Eleven Translocation-2 (TET2) gene in hematopoiesis and hematopoietic diseases. Leukemia. 2014;

300. Ko M, An J, Bandukwala HS, Chavez L, Äijö T, Pastor WA, et al. Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX. Nature. 2013;

301. Kinney SRM, Pradhan S. Ten eleven translocation enzymes and 5-hydroxymethylation in mammalian development and cancer. Adv Exp Med Biol. 2013;

302. Wu H, Zhang Y. Mechanisms and functions of Tet proteinmediated 5-methylcytosine oxidation. Genes and Development. 2011.

303. Oswald J, Engemann S, Lane N, Mayer W, Olek A, Fundele R, et al. Active demethylation of the paternal genome in the mouse zygote. Current Biology. 2000;

304. Hajkova P, Erhardt S, Lane N, Haaf T, El-Maarri O, Reik W, et al. Epigenetic reprogramming in mouse primordial germ cells. Mech Dev. 2002;

305. Lorsback RB, Moore J, Mathew S, Raimondi SC, Mukatira ST, Downing JR. TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;23) [3]. Leukemia. 2003.

306. Dawlaty MM, Breiling A, Le T, Raddatz G, Barrasa MI, Cheng AW, et al. Combined Deficiency of Tet1 and Tet2 Causes Epigenetic Abnormalities but Is Compatible with Postnatal Development. Dev Cell. 2013;

307. Koh KP, Yabuuchi A, Rao S, Huang Y, Cunniff K, Nardone J, et al. Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell. 2011;

308. Dawlaty MM, Breiling A, Le T, Barrasa MI, Raddatz G, Gao Q, et al. Loss of tet enzymes compromises proper differentiation of embryonic stem cells. Dev Cell. 2014;

309. López V, Fernández AF, Fraga MF. The role of 5-hydroxymethylcytosine in development, aging and age-related diseases. Ageing Research Reviews. 2017.

310. Nabel CS, Jia H, Ye Y, Shen L, Goldschmidt HL, Stivers JT, et al. AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. Nat Chem Biol. 2012;

311. Morgan HD, Dean W, Coker HA, Reik W, Petersen-Mahrt SK. Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: Implications for epigenetic reprogramming. Journal of Biological Chemistry. 2004;

312. Hendrich B, Hardeland U, Ng HH, Jiricny J, Bird A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature. 1999;

313. He S, Sun H, Lin L, Zhang Y, Chen J, Liang L, et al. Passive DNA demethylation preferentially up-regulates pluripotency-related genes and facilitates the generation of induced pluripotent stem cells. Journal of Biological Chemistry. 2017;

314. Yang J, Bashkenova N, Zang R, Huang X, Wang J. The roles of TET family proteins in development and stem cells. Development (Cambridge). 2020;

315. Neri F, Incarnato D, Krepelova A, Rapelli S, Pagnani A, Zecchina R, et al. Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells. Genome Biol. 2013;14(8).

316. Pastor WA, Aravind L, Rao A. TETonic shift: Biological roles of TET proteins in DNA demethylation and transcription. Nature Reviews Molecular Cell Biology. 2013.

317. Gommers-Ampt JH, Van Leeuwen F, de Beer ALJ, Vliegenthart JFG, Dizdaroglu M, Kowalak JA, et al. β-d-glucosyl-hydroxymethyluracil: A novel modified base present in the DNA of the parasitic protozoan T. brucei. Cell. 1993;

318. Yamagata K, Kobayashi A. The cysteine-rich domain of TET2 binds preferentially to mono- and dimethylated histone H3K36. J Biochem. 2017;

319. Ko M, Huang Y, Jankowska AM, Pape UJ, Tahiliani M, Bandukwala HS, et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature. 2010;

320. Garcia-Outeiral V, de la Parte C, Fidalgo M, Guallar D. The Complexity of TET2 Functions in Pluripotency and Development. Vol. 8, Frontiers in Cell and Developmental Biology. 2021.

321. Yang H, Liu Y, Bai F, Zhang JY, Ma SH, Liu J, et al. Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. Oncogene. 2013;

322. Feng Y, Li X, Cassady K, Zou Z, Zhang X. TET2 Function in Hematopoietic Malignancies, Immune Regulation, and DNA Repair. Front Oncol. 2019;

323. Inoue S, Lemonnier F, Mak TW. Roles of IDH1/2 and TET2 mutations in myeloid disorders. International Journal of Hematology. 2016.

324. Voigt P, Reinberg D. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia The Cancer Genome Atlas Research Network. N Engl J Med. 2013;

325. Duployez N, Goursaud L, Fenwarth L, Bories C, Marceau-Renaut A, Boyer T, et al. Familial myeloid malignancies with germline TET2 mutation. Leukemia. 2020;

326. Weissmann S, Alpermann T, Grossmann V, Kowarsch A, Nadarajah N, Eder C, et al. Landscape of TET2 mutations in acute myeloid leukemia. Leukemia. 2012;

327. Nibourel O, Kosmider O, Cheok M, Boissel N, Renneville A, Philippe N, et al. Incidence and prognostic value of TET2 alterations in de novo acute myeloid leukemia achieving complete remission. Blood. 2010;

328. Gaidzik VI, Paschka P, Späth D, Habdank M, Köhne CH, Germing U, et al. TET2 mutations in Acute Myeloid Leukemia (AML): Results from a comprehensive genetic and clinical analysis of the AML study group. Journal of Clinical Oncology. 2012;

329. Wang R, Gao X, Yu L. The prognostic impact of tet oncogene family member 2 mutations in patients with acute myeloid leukemia: A systematic-review and meta-analysis. BMC Cancer. 2019;

330. Figueroa ME, Abdel-Wahab O, Lu C, Ward PS, Patel J, Shih A, et al. Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. Cancer Cell. 2010;

331. Sajadian SO, Ehnert S, Vakilian H, Koutsouraki E, Damm G, Seehofer D, et al. Induction of active demethylation and 5hmC formation by 5-azacytidine is TET2 dependent and suggests new treatment strategies against hepatocellular carcinoma. Clin Epigenetics [Internet]. 2015;7(1):1–14. Available from: http://dx.doi.org/10.1186/s13148-015-0133-x

332. García MG, Carella A, Urdinguio RG, Bayón GF, Lopez V, Tejedor JR, et al. Epigenetic dysregulation of TET2 in human glioblastoma. Oncotarget. 2018;

333. Uribe-Lewis S, Stark R, Carroll T, Dunning MJ, Bachman M, Ito Y, et al. 5-hydroxymethylcytosine marks promoters in colon that resist DNA hypermethylation in cancer. Genome Biol. 2015;

334. Uribe-Lewis S, Carroll T, Menon S, Nicholson A, Manasterski PJ, Winton DJ, et al. 5-hydroxymethylcytosine and gene activity in mouse intestinal differentiation. Sci Rep. 2020;

335. Kumar D, Cinghu S, Oldfield AJ, Yang P, Jothi R. Decoding the function of bivalent chromatin in development and cancer. Genome Res. 2021;31(12).

336. Dunican DS, Mjoseng HK, Duthie L, Flyamer IM, Bickmore WA, Meehan RR. Bivalent promoter hypermethylation in cancer is linked to the H327me3/H3K4me3 ratio in embryonic stem cells. BMC Biol. 2020;18(1).

337. Balss J, Meyer J, Mueller W, Korshunov A, Hartmann C, von Deimling A. Analysis of the IDH1 codon 132 mutation in brain tumors. Acta Neuropathol. 2008;

338. Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science (1979). 2008;

339. Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, Yuan W, et al. *IDH1* and *IDH2* Mutations in Gliomas. New England Journal of Medicine. 2009;

340. Yang H, Ye D, Guan KL, Xiong Y. IDH1 and IDH2 mutations in tumorigenesis: Mechanistic insights and clinical perspectives. Clinical Cancer Research. 2012.

341. Pansuriya TC, Van Eijk R, D'Adamo P, Van Ruler MAJH, Kuijjer ML, Oosting J, et al. Somatic mosaic IDH1 and IDH2 mutations are associated with enchondroma and spindle cell hemangioma in Ollier disease and Maffucci syndrome. Nat Genet. 2011;

342. Borger DR, Tanabe KK, Fan KC, Lopez HU, Fantin VR, Straley KS, et al. Frequent Mutation of Isocitrate Dehydrogenase (IDH)1 and IDH2 in Cholangiocarcinoma Identified Through Broad-Based Tumor Genotyping. Oncologist. 2011;

343. Cairns RA, Mak TW. Oncogenic isocitrate dehydrogenase mutations: Mechanisms, models, and clinical opportunities. Cancer Discovery. 2013.

344. Wang HY, Tang K, Liang TY, Zhang WZ, Li JY, Wang W, et al. The comparison of clinical and biological characteristics between IDH1 and IDH2 mutations in gliomas. Journal of Experimental and Clinical Cancer Research. 2016;

345. Molenaar RJ, Maciejewski JP, Wilmink JW, Van Noorden CJF. Wild-type and mutated IDH1/2 enzymes and therapy responses. Oncogene. 2018.

346. Cohen AL, Holmen SL, Colman H. IDH1 and IDH2 mutations in gliomas. Curr Neurol Neurosci Rep. 2013;

347. Guengerich FP. Introduction: Metals in biology: α-Ketoglutarate/iron-dependent dioxygenases. Journal of Biological Chemistry. 2015;290(34):20700–1.

348. Pusch S, Schweizer L, Beck AC, Lehmler JM, Weissert S, Balss J, et al. D-2-Hydroxyglutarate producing neo-enzymatic activity inversely correlates with frequency of the type of isocitrate dehydrogenase 1 mutations found in glioma. Acta Neuropathol Commun. 2014;

349. Ashraf S, Noguera NI, Di Giandomenico J, Zaza S, Hasan SK, Lo-Coco F. Rapid detection of IDH2 (R140Q and R172K) mutations in acute myeloid leukemia. Ann Hematol. 2013;

350. Dang L, White DW, Gross S, Bennett BD, Bittinger MA, Driggers EM, et al. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature. 2009;

351. Guo C, Pirozzi CJ, Lopez GY, Yan H. Isocitrate dehydrogenase mutations in gliomas: Mechanisms, biomarkers and therapeutic target. Current Opinion in Neurology. 2011.

352. Rakheja D, Medeiros LJ, Bevan S, Chen W. The Emerging Role of D-2-Hydroxyglutarate as an Oncometabolite in Hematolymphoid and Central Nervous System Neoplasms. Front Oncol. 2013;

353. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012;

354. Bardella C, Al-Dalahmah O, Krell D, Brazauskas P, Al-Qahtani K, Tomkova M, et al. Expression of Idh1R132Hin the Murine Subventricular Zone Stem Cell Niche Recapitulates Features of Early Gliomagenesis. Cancer Cell. 2016;

355. Lim DA, Cha S, Mayo MC, Chen MH, Keles E, VandenBerg S, et al. Relationship of glioblastoma multiforme to neural stem cell regions predicts invasive and multifocal tumor phenotype. Neuro Oncol. 2007;

356. Chan SM, Majeti R. Role of DNMT3A, TET2, and IDH1/2 mutations in pre-leukemic stem cells in acute myeloid leukemia. Int J Hematol. 2013;

357. Kim H, Kim M, Im SK, Fang S. Mouse Cre-LoxP system: general principles to determine tissue-specific roles of target genes. Vol. 34, Laboratory Animal Research. 2018.

358. Rutlin M, Rastelli D, Kuo WT, Estep JA, Louis A, Riccomagno MM, et al. The Villin1 Gene Promoter Drives Cre Recombinase Expression in Extraintestinal Tissues. CMGH. 2020;10(4).

359. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;

360. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics. 2017;

361. Guinney J, Dienstmann R, Wang X, De Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;

362. Hinoue T, Weisenberger DJ, Lange CPE, Shen H, Byun HM, Van Den Berg D, et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res. 2012;

363. https://www.cancer.ox.ac.uk/research/networks/scort.

364. https://www.dana-farber.org/.

365. Kassambara A. ggpubr: “ggplot2” based publication ready plots. R package version 0.2. https://CRANR-project.org/package=ggpubr. 2020;

366. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;

367. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. J Stat Softw. 2017;

368. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;

369. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;

370. https://cran.r-project.org/web/packages/PairedData/PairedData.pdf.

371. Court F, Arnaud P. An annotated list of bivalent chromatin regions in human ES cells: A new tool for cancer epigenetic research. Oncotarget. 2017;

372. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;

373. Brenet F, Moh M, Funk P, Feierstein E, Viale AJ, Socci ND, et al. DNA methylation of the first exon is tightly linked to transcriptional silencing. PLoS One. 2011;

374. Fernandes MS, Carneiro F, Oliveira C, Seruca R. Colorectal cancer and RASSF family-A special emphasis on RASSF1A. International Journal of Cancer. 2013.

375. Guo Y, Shu L, Zhang C, Su ZY, Kong ANT. Curcumin inhibits anchorage-independent growth of HT29 human colon cancer cells by targeting epigenetic restoration of the tumor suppressor gene DLEC1. Biochem Pharmacol. 2015;

376. Ying J, Poon FF, Yu J, Geng H, Wong AHY, Qiu GH, et al. DLEC1 is a functional 3p22.3 tumour suppressor silenced by promoter CpG methylation in colon and gastric cancers. Br J Cancer. 2009;

377. Kolla V, Zhuang T, Higashi M, Naraparaju K, Brodeur GM. Role of CHD5 in human cancers: 10 years later. Cancer Research. 2014.

378. Weber S, Koschade SE, Hoffmann CM, Dubash TD, Giessler KM, Dieter SM, et al. The notch target gene HEYL modulates metastasis forming capacity of colorectal cancer patient-derived spheroid cells in vivo. BMC Cancer. 2019;

379. Yao X, Hu W, Zhang J, Huang C, Zhao H, Yao X. Application of cAMP-dependent catalytic subunit β (PRKACB) low expression in predicting worse overall survival: A potential therapeutic target for colorectal carcinoma. J Cancer. 2020;

380. Ma TL, Zhu P, Chen JX, Hu YH, Xie J. SIX3 function in cancer: progression and comprehensive analysis. Cancer Gene Ther. 2022;29:1542–9.

381. Crist RC, Roth JJ, Waldman SA, Buchberg AM. A conserved tissue-specific homeodomain-less isoform of MEIS1 is downregulated in colorectal cancer. PLoS One. 2011;

382. Liu X, Wen J, Li C, Wang H, Wang J, Zou H. High-Yield Methylation Markers for Stool-Based Detection of Colorectal Cancer. Dig Dis Sci. 2020;65:1710–9.

383. Wang Y, Dan L, Li Q, Li L, Zhong L, Shao B, et al. ZMYND10, an epigenetically regulated tumor suppressor, exerts tumor-suppressive functions via miR145-5p/NEDD9 axis in breast cancer. Clin Epigenetics. 2019;

384. Bergheim J, Semaan A, Gevensleben H, Groening S, Knoblich A, Dietrich J, et al. Potential of quantitative SEPT9 and SHOX2 methylation in plasmatic circulating cell-free DNA as auxiliary staging parameter in colorectal cancer: A prospective observational cohort study. Br J Cancer. 2018;

385. Chung HH, Lee C Te, Hu JM, Chou YC, Lin YW, Shih YL. NKX6.1 represses tumorigenesis, metastasis, and chemoresistance in colorectal cancer. Int J Mol Sci. 2020;

386. OuYang C, Xie Y, Fu Q, Xu G. SYNPO2 suppresses hypoxia-induced proliferation and migration of colorectal cancer cells by regulating YAP-KLF5 axis. Tissue Cell. 2021;

387. Harvey KF, Zhang X, Thomas DM. The Hippo pathway and human cancer. Nature Reviews Cancer. 2013.

388. Barrett CW, Ning W, Chen X, Smith JJ, Washington MK, Hill KE, et al. Tumor suppressor function of the plasma glutathione peroxidase Gpx3 in colitis-associated carcinoma. Cancer Res. 2013;

389. Noda H, Miyaji Y, Nakanishi A, Konishi F, Miki Y. Frequent reduced expression of alpha-1B-adrenergic receptor caused by aberrant promoter methylation in gastric cancers. Br J Cancer. 2007;

390. Amunjela JN, Tucker SJ. POPDC proteins as potential novel therapeutic targets in cancer. Drug Discovery Today. 2016.

391. Shibata D, Mori Y, Cai K, Zhang L, Yin J, Elahi A, et al. RAB32 hypermethylation and microsatellite instability in gastric and endometrial adenocarcinomas. Int J Cancer. 2006;

392. Fan G, Sun L, Shan P, Zhang X, Huan J, Zhang X, et al. Loss of KLF14 triggers centrosome amplification and tumorigenesis. Nat Commun. 2015;

393. Ying K, Wang C, Liu S, Kuang Y, Tao Q, Hu X. Diverse Ras-related GTPase DIRAS2, downregulated by PSMD2 in a proteasome-mediated way, inhibits colorectal cancer proliferation by blocking NF-κB signaling. Int J Biol Sci. 2022;

394. Zhang XW, Wang XF, Ni SJ, Qin W, Zhao LQ, Hua RX, et al. UBTD1 induces cellular senescence through an UBTD1-Mdm2/p53 positive feedback loop. Journal of Pathology. 2015;

395. Erfani M, Zamini M, Tamaddon G, Hosseini S, Mokarram P. Expression and methylation status of BTG2, PPP1CA, and PEG3 genes in colon adenocarcinoma cell lines: promising treatment targets. Gastroenterol Hepatol Bed Bench. 2022;15(4):395–405.

396. Lock FE, Underhill-Day N, Dunwell T, Matallanas D, Cooper W, Hesson L, et al. The RASSF8 candidate tumor suppressor inhibits cell growth and regulates the Wnt and NF-kappaB signaling pathways. Oncogene. 2010;29(30):4307–16.

397. Shah M, Cardenas R, Wang B, Persson J, Mongan NP, Grabowska A, et al. HOXC8 regulates self-renewal, differentiation and transformation of breast cancer stem cells. Mol Cancer. 2017;

398. Li X, Wang J, Zhang C, Lin C, Zhang J, Zhang W, et al. Circular RNA circITGA7 inhibits colorectal cancer growth and metastasis by modulating the Ras pathway and upregulating transcription of its host gene ITGA7. Journal of Pathology. 2018;

399. Yang WT, Chen M, Xu R, Zheng PS. PRDM4 inhibits cell proliferation and tumorigenesis by inactivating the PI3K/AKT signaling pathway through targeting of PTEN in cervical carcinoma. Oncogene. 2021;

400. Chen Z, Wang Z, Chen J, Li Z. Fibulin-5 is down-regulated in colorectal cancer and correlated with clinicopathologic characteristics. Clin Lab. 2018;

401. Tufan T, Yang J, Tummala KS, Cingoz H, Kuscu C, Adair SJ, et al. ISL2 is an epigenetically silenced tumor suppressor and regulator of metabolism in pancreatic cancer. bioRxiv. 2020;

402. Tobelaim WS, Beaurivage C, Champagne A, Pomerleau V, Simoneau A, Chababi W, et al. Tumour-promoting role of SOCS1 in colorectal cancer cells. Sci Rep. 2015;

403. Wang G, Wang F, Meng Z, Wang N, Zhou C, Zhang J, et al. Uncovering potential genes in colorectal cancer based on integrated and DNA methylation analysis in the gene expression omnibus database. BMC Cancer. 2022;

404. Xu X, Chang X, Xu Y, Deng P, Wang J, Zhang C, et al. SAMD14 promoter methylation is strongly associated with gene expression and poor prognosis in gastric cancer. Int J Clin Oncol. 2020;

405. Model F, Osborn N, Ahlquist D, Gruetzmann R, Molnar B, Sipos F, et al. Identification and validation of colorectal neoplasia-specific methylation markers for accurate classification of disease. Molecular Cancer Research. 2007;

406. Dunn BK. Hypomethylation: One side of a larger picture. In: Annals of the New York Academy of Sciences. 2003.

407. Verma N, Pan H, Doré LC, Shukla A, Li Q V., Pelham-Webb B, et al. TET proteins safeguard bivalent promoters from de novo methylation in human embryonic stem cells. Nat Genet. 2018;

408. Moser AR, Pitot HC, Dove WF. A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse. Science (1979). 1990;

409. Leclerc D, Deng L, Trasler J, Rozen R. ApcMin/+ mouse model of colon cancer: Gene expression profiling in tumors. J Cell Biochem. 2004;

410. Kamdar S, Isserlin R, Van Der Kwast T, Zlotta AR, Bader GD, Fleshner NE, et al. Exploring targets of TET2-mediated methylation reprogramming as potential discriminators of prostate cancer progression. Clin Epigenetics. 2019;

411. Haffner MC, Chaux A, Meeker AK, Esopi DM, Gerber J, Pellakuru LG, et al. Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. Oncotarget. 2011;

412. Wilson ER, Helton NM, Heath SE, Fulton RS, Payton JE, Welch JS, et al. Focal disruption of DNA methylation dynamics at enhancers in IDH-mutant AML cells. Leukemia. 2022;

413. Pelosof L, Yerram S, Armstrong T, Chu N, Danilova L, Yanagisawa B, et al. GPX3 promoter methylation predicts platinum sensitivity in colorectal cancer. Epigenetics. 2017;

414. Naumov VA, Generozov E V., Zaharjevskaya NB, Matushkina DS, Larin AK, Chernyshov S V., et al. Genome-scale analysis of DNA methylation in colorectal cancer using Infinium HumanMethylation450 BeadChips. Epigenetics. 2013;

415. Arechederra M, Daian F, Yim A, Bazai SK, Richelme S, Dono R, et al. Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. Nat Commun. 2018;

416. Bernhart SH, Kretzmer H, Holdt LM, Jühling F, Ammerpohl O, Bergmann AK, et al. Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. Sci Rep. 2016;

417. Lee HY, Tang DW, Liu CY, Cho EC. A novel HDAC1/2 inhibitor suppresses colorectal cancer through apoptosis induction and cell cycle regulation. Chem Biol Interact. 2022;

418. Han S, Liu Y, Cai SJ, Qian M, Ding J, Larion M, et al. IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. British Journal of Cancer. 2020.

419. Fujii T, Khawaja MR, DiNardo CD, Atkins JT, Janku F. Targeting Isocitrate Dehydrogenase (IDH) in cancer. Discov Med. 2016;

420. Yen K, Wang F, Schalm S, Hansen E, Straley K, Kernytsky A, et al. Mutation Selective IDH Inhibitors Mediate Histone and DNA Methylation Changes. Blood. 2012;

421. Liu X, Gong Y. Isocitrate dehydrogenase inhibitors in acute myeloid leukemia. Biomarker Research. 2019.

422. Ushijima T, Suzuki H. The Origin of CIMP, At Last. Cancer Cell. 2019.

423. Koh G, Degasperi A, Zou X, Momen S, Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. Nature Reviews Cancer. 2021.

424. Paolillo C, Londin E, Fortina P. Next generation sequencing in cancer: opportunities and challenges for precision cancer medicine. Scand J Clin Lab Invest. 2016;

425. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;

426. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., et al. Signatures of mutational processes in human cancer. Nature. 2013;

427. van den Heuvel GRM, Kroeze LI, Ligtenberg MJL, Grünberg K, Jansen EAM, von Rhein D, et al. Mutational signature analysis in non-small cell lung cancer patients with a high tumor mutational burden. Respir Res. 2021;

428. Zhuravleva E, O'Rourke CJ, Andersen JB. Mutational signatures and processes in hepatobiliary cancers. Nature Reviews Gastroenterology and Hepatology. 2022.

429. Litchfield K, Reading JL, Puttick C, Thakkar K, Abbosh C, Bentham R, et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. Cell. 2021;

430. Chopra N, Tovey H, Pearson A, Cutts R, Toms C, Proszek P, et al. Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative breast cancer. Nat Commun. 2020;

431. Totsuka Y, Watanabe M, Lin Y. New horizons of DNA adductome for exploring environmental causes of cancer. Cancer Science. 2021.

432. Stower H. Chemotherapy signatures. Nature Medicine. 2020.

433. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019;

434. Ehrlich M, Norris KF, Wang RY, Kuo KC, Gehrke CW. DNA cytosine methylation and heat-induced deamination. Biosci Rep. 1986;

435. Ehrlich M, Zhang XY, Inamdar NM. Spontaneous deamination of cytosine and 5-methylcytosine residues in DNA and replacement of 5-methylcytosine residues with cytosine residues. Mutation Research/Reviews in Genetic Toxicology. 1990;

436. Wiebauer K, Jiricny J. In vitro correction of G o T mispairs to G o C pairs in nuclear extracts from human cells. Nature. 1989;

437. Shen J cheng, Rideout WM, Jones PA. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. 1994;

438. Bhagwat AS, Hao W, Townes JP, Lee H, Tang H, Foster PL. Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in Escherichia coli. Proc Natl Acad Sci U S A. 2016;

439. Crosnier C, Stamataki D, Lewis J. Organizing cell renewal in the intestine: Stem cells, signals and combinatorial control. Nature Reviews Genetics. 2006.

440. Sumigray KD, Terwilliger M, Lechler T. Morphogenesis and Compartmentalization of the Intestinal Crypt. Dev Cell. 2018;

441. Vasovcak P, Krepelova A, Menigatti M, Puchmajerova A, Skapa P, Augustinakova A, et al. Unique mutational profile associated with a loss of TDG expression in the rectal cancer of a patient with a constitutional PMS2 deficiency. DNA Repair (Amst). 2012;

442. Bellacosa A. Role of MED1 (MBD4) gene in DNA repair and human cancer. Journal of Cellular Physiology. 2001.

443. Pidugu LS, Bright H, Lin WJ, Majumdar C, Van Ostrand RP, David SS, et al. Structural Insights into the Mechanism of Base Excision by MBD4. J Mol Biol. 2021;

444. Hashimoto H, Zhang X, Cheng X. Excision of thymine and 5-hydroxymethyluracil by the MBD4 DNA glycosylase domain: Structural basis and implications for active DNA demethylation. Nucleic Acids Res. 2012;

445. Srivastava DK, Vande Berg BJ, Prasad R, Molina JT, Beard WA, Tomkinson AE, et al. Mammalian abasic site base excision repair: Identification of the reaction sequence and rate-determining steps. Journal of Biological Chemistry. 1998;

446. Sanders MA, Chew E, Flensburg C, Zeilemaker A, Miller SE, Al Hinai AS, et al. MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML. Blood. 2018;

447. Tanakaya K, Kumamoto K, Tada Y, Eguchi H, Ishibashi K, Idani H, et al. A germline MBD4 mutation was identified in a patient with colorectal oligopolyposis and early-onset cancer: A case report. Oncol Rep. 2019;

448. Rodrigues M, Mobuchon L, Houy A, Fiévet A, Gardrat S, Barnhill RL, et al. Outlier response to anti-PD1 in uveal melanoma reveals germline MBD4 mutations in hypermutated tumors. Nat Commun. 2018;

449. Riccio A, Aaltonen LA, Godwin AK, Loukola A, Percesepe A, Salovaara R, et al. The DNA repair gene MBD4 (MED1) is mutated in human carcinomas with microsatellite instability [3]. Nature Genetics. 1999.

450. Wong E, Yang K, Kuraguchi M, Werling U, Avdievich E, Fan K, et al. Mbd4 inactivation increases C→T transition mutations and promotes gastrointestinal tumor formation. Proc Natl Acad Sci U S A. 2002;

451. Degasperi A, Zou X, Dias Amarante T, Martinez-Martinez A, Ching Chiek Koh G, Dias JM, et al. Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. Science (1979). 2022;376(6591).

452. Fang H, Barbour JA, Poulos RC, Katainen R, Aaltonen LA, Wong JWH. Mutational processes of distinct POLE exonuclease domain mutants drive an enrichment of a specific TP53 mutation in colorectal cancer. PLoS Genet. 2020;

453. Poulos RC, Olivier J, Wong JWH. The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. Nucleic Acids Res. 2017;

454. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015;

455. Beletskii A, Bhagwat AS. Transcription-induced mutations: Increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. Proc Natl Acad Sci U S A. 1996;

456. Mugal CF, Von Grünberg HH, Peifer M. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. Mol Biol Evol. 2009;

457. Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC. Microsatellite instability detection by next generation sequencing. Clin Chem. 2014;

458. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. Annals of Oncology. 2015;

459. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

460. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, Mcewen R, et al. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;

461. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;

462. Islam SMA, Diaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. Cell Genomics. 2022;2(11):100179.

463. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. Nature Biotechnology. 2010.

464. Marchal C, Sasaki T, Vera D, Wilson K, Sima J, Rivera-Mulia JC, et al. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. Nat Protoc. 2018;

465. https://www.gencodegenes.org/human/.

466. Vöhringer H, Hoeck A Van, Cuppen E, Gerstung M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. Nat Commun. 2021;

467. Damiola F, Pertesi M, Oliver J, Le Calvez-Kelm F, Voegele C, Young EL, et al. Rare key functional domain missense substitutions in MRE11A, RAD50, and NBN contribute to breast cancer susceptibility: Results from a Breast Cancer Family Registry case-control mutation-screening study. Breast Cancer Research. 2014;

468. Duffy MJ, Synnott NC, Crown J. Mutant p53 in breast cancer: potential as a therapeutic target and biomarker. Breast Cancer Research and Treatment. 2018.

469. Zhu J, Zhao C, Zhuang T, Jonsson P, Sinha I, Williams C, et al. RING finger protein 31 promotes p53 degradation in breast cancer cells. Oncogene. 2016;

470. Yang J, Ren Y, Wang L, Li B, Chen Y, Zhao W, et al. PTEN mutation spectrum in breast cancers and breast hyperplasia. J Cancer Res Clin Oncol. 2010;

471. Heitzer E, Sunitsch S, Gilg MM, Lohberger B, Rinner B, Kashofer K, et al. Expanded molecular profiling of myxofibrosarcoma reveals potentially actionable targets. Modern Pathology. 2017;

472. Ognjanovic S, Olivier M, Bergemann TL, Hainaut P. Sarcomas in TP53 germline mutation carriers: A review of the IARC TP53 database. Cancer. 2012;

473. Singh N, Singh R, Bowen RC, Abdel-Rahman MH, Singh AD. Uveal Melanoma in BAP1 Tumor Predisposition Syndrome: Estimation of Risk. Am J Ophthalmol. 2021;

474. Van Raamsdonk CD, Griewank KG, Crosby MB, Garrido MC, Vemula S, Wiesner T, et al. Mutations in GNA11 in Uveal Melanoma . New England Journal of Medicine. 2010;

475. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;

476. Clock-Like Mutational Signatures Are Present in Human Cancers. Cancer Discov. 2016;

477. Bralten LBC, French PJ. Genetic alterations in Glioma. Cancers. 2011.

478. Jamaspishvili T, Berman DM, Ross AE, Scher HI, De Marzo AM, Squire JA, et al. Clinical implications of PTEN loss in prostate cancer. Nature Reviews Urology. 2018.

479. Derrien AC, Rodrigues M, Eeckhoutte A, Dayot S, Houy A, Mobuchon L, et al. Germline mbd4 mutations and predisposition to uveal melanoma. J Natl Cancer Inst. 2021;

480. Saint-Ghislain M, Derrien AC, Geoffois L, Gastaud L, Lesimple T, Negrier S, et al. MBD4 deficiency is predictive of response to immune checkpoint inhibitors in metastatic uveal melanoma patients. Eur J Cancer. 2022;173:105–12.

481. Bellacosa A, Cicchillitti L, Schepis F, Riccio A, Yeung AT, Matsumoto Y, et al. MED1, a novel human methyl-CpG-binding endonuclease, interacts with DNA mismatch repair protein MLH1. Proc Natl Acad Sci U S A. 1999;

482. Grigera F, Bellacosa A, Kenter AL. Complex relationship between mismatch repair proteins and mbd4 during immunoglobulin class switch recombination. PLoS One. 2013;

483. Lohman TM, Bjornson KP. Mechanisms of helicase-catalyzed DNA unwinding. Annual Review of Biochemistry. 1996.

484. Clevers H. XThe intestinal crypt, a prototype stem cell compartment. Cell. 2013.

485. Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. 2018;

486. Kunkel TA, Bebenek K. DNA replication fidelity. Annual Review of Biochemistry. 2000.

487. Kunkel TA, Erie DA. Eukaryotic Mismatch Repair in Relation to DNA Replication. Annu Rev Genet. 2015;

488. Strauss BS, Sagher D, Acharya S. Role of proofreading and mismatch repair in maintaining the stability of nucleotide repeats in DNA. Nucleic Acids Res. 1997;

489. Prindle MJ, Loeb LA. DNA polymerase delta in dna replication and genome maintenance. Environmental and Molecular Mutagenesis. 2012.

490. Hübscher U, Nasheuer HP, Syväoja JE. Eukaryotic DNA polymerases, a growing family. Trends in Biochemical Sciences. 2000.

491. Jain R, Aggarwal AK, Rechkoblit O. Eukaryotic DNA polymerases. Current Opinion in Structural Biology. 2018.

492. Henninger EE, Pursell ZF. DNA polymerase ε and its roles in genome stability. IUBMB Life. 2014.

493. Chatterjee N, Walker GC. Mechanisms of DNA damage, repair, and mutagenesis. Environmental and Molecular Mutagenesis. 2017.

494. Hoitsma NM, Whitaker AM, Schaich MA, Smith MR, Fairlamb MS, Freudenthal BD. Structure and function relationships in mammalian DNA polymerases. Cellular and Molecular Life Sciences. 2020.

495. Khare V, Eckert KA. The proofreading 3′ → 5′ exonuclease activity of DNA polymerases: A kinetic barrier to translesion DNA synthesis. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis. 2002.

496. McConechy MK, Talhouk A, Leung S, Chiu D, Yang W, Senz J, et al. Endometrial carcinomas with POLE exonuclease domain mutations have a favorable prognosis. Clinical Cancer Research. 2016;

497. Castellucci E, He T, Goldstein DY, Halmos B, Chuy J. DNA Polymerase ε Deficiency Leading to an Ultramutator Phenotype: A Novel Clinically Relevant Entity. Oncologist. 2017;

498. Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R, et al. Comprehensive Analysis of Hypermutation in Human Cancer. Cell. 2017;

499. Soriano I, Vazquez E, Leon N De, Bertrand S, Heitzer E, Toumazou S, et al. Expression of the cancer-associated DNA polymerase ϵ P286R in fission yeast leads to translesion synthesis polymerase dependent hypermutation and defective DNA replication. PLoS Genet. 2021;

500. Rosa RCA, Yurchenko AA, Chahud F, Ribeiro-Silva A, Brunaldi MO, Silva WA, et al. First description of ultramutated endometrial cancer caused by germline loss-of-function and somatic exonuclease domain mutations in pole gene. Genet Mol Biol. 2020;

501. Church DN, Stelloo E, Nout RA, Valtcheva N, Depreeuw J, Haar N Ter, et al. Prognostic significance of POLE proofreading mutations in endometrial cancer. J Natl Cancer Inst. 2015;

502. Heitzer E, Tomlinson I. Replicative DNA polymerase mutations in cancer. Current Opinion in Genetics and Development. 2014.

503. Li Y, Bian Y, Wang K, Wan XP. POLE mutations improve the prognosis of endometrial cancer via regulating cellular metabolism through AMF/AMFR signal transduction. BMC Med Genet. 2019;

504. Domingo E, Freeman-Mills L, Rayner E, Glaire M, Briggs S, Vermeulen L, et al. Somatic POLE proofreading domain mutation, immune response, and prognosis in colorectal cancer: a retrospective, pooled biomarker study. Lancet Gastroenterol Hepatol. 2016;

505. Kolodner RD, Marsischky GT. Eukaryotic DNA mismatch repair. Curr Opin Genet Dev. 1999;

506. Sachadyn P. Conservation and diversity of MutS proteins. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis. 2010;

507. Schofield MJ, Hsieh P. Dna Mismatch Repair: Molecular Mechanisms and Biological Function. Annual Review of Microbiology. 2003.

508. Jiricny J. The multifaceted mismatch-repair system. Nature Reviews Molecular Cell Biology. 2006.

509. Mendillo ML, Hargreaves V V., Jamison JW, Mo AO, Li S, Putnam CD, et al. A conserved MutS homolog connector domain interface interacts with MutL homologs. Proc Natl Acad Sci U S A. 2009;

510. Guarné A. The functions of MutL in mismatch repair: The power of multitasking. In: Progress in Molecular Biology and Translational Science. 2012.

511. Zhang Y, Yuan F, Presnell SR, Tian K, Gao Y, Tomkinson AE, et al. Reconstitution of 5′-directed human mismatch repair in a purified system. Cell. 2005;

512. Lee BI, Nguyen LH, Barsky D, Fernandes M, Wilson DM. Molecular interactions of human Exo1 with DNA. Nucleic Acids Res. 2002;

513. Dzantiev L, Constantin N, Genschel J, Iyer RR, Burgers PM, Modrich P. A defined human system that supports bidirectional mismatch-provoked excision. Mol Cell. 2004;

514. Fernandez-Leiro R, Bhairosing-Kok D, Kunetsky V, Laffeber C, Winterwerp HH, Groothuizen F, et al. The selection process of licensing a DNA mismatch for repair. Nat Struct Mol Biol. 2021;

515. Putnam CD. Strand discrimination in DNA mismatch repair. DNA Repair. 2021.

516. Sameer AS, Nissar S, Fatima K. Mismatch repair pathway: Molecules, functions, and role in colorectal carcinogenesis. European Journal of Cancer Prevention. 2014.

517. Nojadeh JN, Sharif SB, Sakhinia E. Microsatellite instability in colorectal cancer. EXCLI Journal. 2018.

518. Fang H, Zhu X, Yang H, Oh J, Barbour JA, Wong JWH. Deficiency of replication-independent DNA mismatch repair drives a 5-methylcytosine deamination mutational signature in cancer. Sci Adv. 2021;

519. Sanders MA, Vohringer H, Forster VJ, Moore L, Campbell BB, Hooks Y, et al. Life without mismatch repair. bioRxiv. 2021;

520. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell. 2016;

521. León-Castillo A, Britton H, McConechy MK, McAlpine JN, Nout R, Kommoss S, et al. Interpretation of somatic POLE mutations in endometrial carcinoma. Journal of Pathology. 2020;

522. Hamzaoui N, Alarcon F, Leulliot N, Guimbaud R, Buecher B, Colas C, et al. Genetic, structural, and functional characterization of POLE polymerase proofreading variants allows cancer risk prediction. Genetics in Medicine. 2020;

523. Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. Am J Hum Genet. 2012;

524. Tomkova M, McClellan M, Kriaucionis S, Schuster-Böckler B. DNA Replication and associated repair pathways are involved in the mutagenesis of methylated cytosine. DNA Repair (Amst). 2018;

525. Jansen AML, Tops CMJ, Ruano D, van Eijk R, Wijnen JT, ten Broeke S, et al. The complexity of screening PMS2 in DNA isolated from formalin-fixed paraffin-embedded material. European Journal of Human Genetics. 2020;

526. Andrianova MA, Bazykin GA, Nikolaev SI, Seplyarskiy VB. Human mismatch repair system balances mutation rates between strands by removing more mismatches from the lagging strand. Genome Res. 2017;

527. Pandey M, Syed S, Donmez I, Patel G, Ha T, Patel SS. Coordinating DNA replication by means of priming loop and differential synthesis rate. Nature. 2009;

528. Langley AR, Gräf S, Smith JC, Krude T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). Nucleic Acids Res. 2016;

529. Jaksik R, Wheeler DA, Kimmel M. Detection and characterization of constitutive replication origins defined by DNA polymerase epsilon. BMC Biol. 2023;21:41.

530. Guilbaud G, Murat P, Wilkes HS, Lerner LK, Sale JE, Krude T. Determination of human DNA replication origin position and efficiency reveals principles of initiation zone organisation. Nucleic Acids Res. 2022;50(13):7436–50.

531. Giner-Calabuig M, De Leon S, Wang J, Fehlmann TD, Ukaegbu C, Gibson J, et al. Mutational signature profiling classifies subtypes of clinically different mismatch-repair-deficient tumours with a differential immunogenic response potential. Br J Cancer. 2022;

532. Ortega J, Lee GS, Gu L, Yang W, Li GM. Mispair-bound human MutS–MutL complex triggers DNA incisions and activates mismatch repair. Cell Res. 2021;

533. Buitrago D, Labrador M, Arcon JP, Lema R, Flores O, Esteve-Codina A, et al. Impact of DNA methylation on 3D genome structure. Nat Commun. 2021;12:3243.

534. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;

535. https://bcac.ccge.medschl.cam.ac.uk/.

536. http://practical.icr.ac.uk/blog/.

537. Heyn P, Logan C V., Fluteau A, Challis RC, Auchynnikava T, Martin CA, et al. Gain-of-function DNMT3A mutations cause microcephalic dwarfism and hypermethylation of Polycomb-regulated regions. Nat Genet. 2019;

538. Anteneh H, Fang J, Song J. Structural basis for impairment of DNA methylation by the DNMT3A R882H mutation. Nat Commun. 2020;

539. Nosho K, Shima K, Irahara N, Kure S, Baba Y, Kirkner GJ, et al. DNMT3B expression might contribute to CpG island methylator phenotype in colorectal cancer. Clinical Cancer Research. 2009;

540. Mo HY, An CH, Choi EJ, Yoo NJ, Lee SH. Somatic mutation and loss of expression of a candidate tumor suppressor gene TET3 in gastric and colorectal cancers. Pathol Res Pract. 2020;

541. Kamori T, Oki E, Shimada Y, Hu Q, Hisamatsu Y, Ando K, et al. The effects of ARID1A mutations on colorectal cancer and associations with PD-L1 expression by stromal cells. Cancer Rep. 2022;

542. Erfani M, Hosseini SV, Mokhtari M, Zamani M, Tahmasebi K, Alizadeh Naini M, et al. Altered ARID1A expression in colorectal cancer. BMC Cancer. 2020;

543. https://www.jax.org/.

544. Lallemand Y, Luria V, Haffner-Krausz R, Lonai P. Maternally expressed PGK-Cre transgene as a tool for early and uniform activation of the Cre site-specific recombinase. Transgenic Res. 1998;

545. Liu Q, Zhu X, Lindström M, Shi Y, Zheng J, Hao X, et al. Yeast mismatch repair components are required for stable inheritance of gene silencing. PLoS Genet. 2020;