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

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

Ross Robert Hannah

Thesis submitted for the degree of Doctor of Philosophy

University of Edinburgh

8th September 2023
Declaration

This thesis has been composed entirely by me.

Experiments and analyses have been carried out solely by me, unless stated otherwise in specific parts.

This work has not been submitted for any other degree or professional qualification.

Signature: 

Date: 08-09-23
Acknowledgments

Firstly, I’d like to thank Duncan for being a supportive supervisor, for his mentorship in helping me to develop my academic skills and for his patience and engagement as a teacher.

Thanks to Jonine, for her enthusiasm and support in maintaining structure and optimism as I progressed through the toughest parts of the project.

Also, a big thanks to Paul, for offering (seemingly) endless advice and answers navigating through bioinformatics, for offering me important perspective when things have felt overwhelming, and all with a calm, kind and frankly therapeutic demeanour whenever this project has felt the most daunting.

I’d also like to thank the entire Sproul lab group over the past 4 years, for their council and company at every unexpected turn the project has taken.

Thanks also to Jim Wilson, and his research group, for teaching me in the ways of population genetics when it was needed it most, and for being a welcoming and friendly group.

I’d also like to thank my thesis committee for their guidance in helping to ensure this project remained on track and reached its final destination.

Most of all, a special thanks to my Dad, to Andrew and to Talia for keeping my head up, and to Helena for keeping my head straight. Thank you for your patience, empathy and for constantly grounding me and helping me to move forward when things felt impossible. I couldn’t have asked for a stronger support network, I’m lucky to have you all.

A final and emphatic thanks to the IGC PhD student group of 2019, who have given me a greater sense of community and friendship than I could have possibly hoped for these last four years. We’ll always have lunch at 12.
Table of Contents

Declaration 2
Acknowledgments 3

Table of Figures ........................................................................................................................................... 8
Table of Tables ............................................................................................................................................... 11

Abstract 12
Lay Summary 14

Chapter 1 – Introduction 16

1.1 – Epigenetics ............................................................................................................................................. 16
1.2 - Epigenetic histone modifications ............................................................................................................. 17
1.3 – Epigenetic DNA modifications ............................................................................................................... 18
1.4 – DNA methylation patterns ..................................................................................................................... 19
1.5 – DNA methylation mechanisms and enzymology ..................................................................................... 22
1.6 – Relationship between transcriptional regulation and DNA methylation in normal tissue .......... 26
1.7 – Dysregulated DNA methylation states in disease and cancer ............................................................... 28
1.8 – Hallmarks of the tumour methylome ....................................................................................................... 29
1.9 – Breast tumour DNA methylation dysregulation ...................................................................................... 34
1.10 – Relationship between genetic sequence and DNA methylation ......................................................... 36
1.11 – Quantitative trait loci and their associations with CpG methylation in normal and tumour tissues ........................................................................................................................................... 37
1.12 – Comparison of methylation changes in functional genetic elements between tumour and normal methylome ....................................................................................................................................... 41
1.13 – Mechanisms of cis-acting methylation QTLs .......................................................................................... 43
1.14 – Mechanisms of trans-acting methylation QTLs ...................................................................................... 45
1.15 – Methylation QTL studies and the challenges of analysing tissue-specific and tumour-specific methylation patterns ........................................................................................................................................... 47
1.16 – Methylation QTL studies in cancer ......................................................................................................... 48
1.17 – Breast cancer methylation mechanisms ................................................................................................. 49
# Chapter 2 – Methods

2.1 - Datasets

2.2 – R environment for pre-processing, analysis and data visualisation

2.3 - Methylation data processing and evaluation

2.4 - SNP genotype data processing and evaluation

2.5 – SNP genotype imputation

2.6 - Analysis of population structure in SNP genotype data

2.7 – Methylation QTL association analysis

2.8 – LD clumping

2.9 – Downstream analysis of methQTLs

2.10 – Annotations for enrichment analyses

2.11 - Enrichment Analyses

# Chapter 3 – Identification of methylation QTLs in breast cancer

3.1 - Introduction

3.2 – Description of CpG methylation data

3.3 - Background of pre-processing, normalisation and quality control methods for Infinium 450k CpG methylation data

3.4 - Evaluation of normalisation methods shows that ssNoob performs best at reducing technical experimental variance in breast tumour methylation samples

3.5 - Evaluation of the quality control steps for breast tumour methylation samples informs robust inclusion criteria

3.6 – Description of SNP genotype data

3.7 - Background of pre-processing, normalisation, quality control and imputation methods for Affymetrix 6.0 SNP genotype data

3.8 - Evaluation of population stratification in the cohort justified a single-ancestry European cohort of genotype samples from normal blood for analysis
3.9 - Establishing an optimised inclusion criteria for pre-processing and quality control of genotype and methylation data ................................................................................................................................................................. 87

3.10 – Background on methodology of methQTL analyses ......................................................................................................................... 92

3.11 - _RegScan_-based GWAS framework identifies methQTLs more accurately than the QTL-specific association method _Matrix eQTL_ ................................................................................................................................................. 94

3.12 – Further investigation of _RegScan_-based GWAS analysis identifies additional SNP/CpG quality control that resolves technical artefacts in methQTL analyses ........................................................................... 101

3.13 - M-value transformation addresses statistical limitations of methylation beta values but has no practical effect on methQTL identification ........................................................................................................... 104

3.14 - Analysis using methylation-adjusted covariate data identifies different methQTLs in breast tumour tissue samples ................................................................................................................................................. 106

3.15 - A stricter imputation quality score threshold for SNPs has no practical effect on the SNP set which passes quality control due to redundancy with other filtering steps ........................................................................... 111

3.16 - Discussion ......................................................................................................................................................................................... 116

Chapter 4 – _Mechanisms which drive methQTL associations in breast tumours are different from those driving methQTLs in normal blood tissue_ 123

4.1 – Introduction .................................................................................................................................................................................. 123

4.2 – There are significant SNP associations with CpG methylation patterns in breast tumours ................................................................. 125

4.3 – Clumping SNPs by linkage disequilibrium identifies independent methQTLs in breast tumours ........................................................................................................................................................................ 125

4.4 – MethQTLs in breast tumours share loci with methQTLs in normal blood tissue ........................................................................................... 128

4.5 – The ratio of _cis_- and _trans_-methQTLs is more balanced in breast tumours than in normal blood tissue ........................................................................................................................................................................ 130

4.6 – The effect size of _trans_-methQTLs is lower than _cis_-methQTLs in breast tumours ............................................................................. 131

4.7 – Discussion .................................................................................................................................................................................. 132

Chapter 5 – _The functional context of cis-methQTLs in breast tumours shows minor differences from normal tissue_ 140

5.1 – Introduction .................................................................................................................................................................................. 140

5.2 – SNP genotypes associate with local CpG methylation in breast tumour _cis_-methQTLs ................................................................................................................. 141

5.3 – _Cis_-methQTL-CpGs are depleted in promoters and CGIs in breast tumours .................................................................................... 146
5.4 – Cis-methQTL-CpGs are depleted in chromatin-based annotations for promoters and enhancers in breast tumours ................................................................. 148

5.5 – Cis-methQTL-CpGs are depleted in PMDs in breast tumours ................................................. 150

5.6 – Cis-methQTL-CpGs are enriched in CTCF sites outside of promoters in breast tumours ..... 155

5.7 – Cis-methQTL-CpGs are depleted in bivalent promoters in breast tumours .... Error! Bookmark not defined.

5.8 – A minority of breast tumour cis-methQTL-SNPs also associate with gene expression .... 156

5.9 – Discussion .................................................................................................................. 158

Chapter 6 – The functional context of trans-methQTLs in breast tumours show major similarities with normal tissue 166

6.1 – Introduction ................................................................................................................. 166

6.2 – SNP genotypes associate with both distal and interchromosomal CpG methylation in breast tumour trans-methQTLs ........................................................................................................... 167

6.3 – Trans-methQTL-CpGs are enriched in CpG islands and promoters but depleted in enhancers in breast tumours ............................................................................................................. 171

6.4 – Trans-methQTL-CpGs are depleted in chromatin-based annotations for enhancers in breast tumours .......................................................................................................................... 174

6.5 – Breast tumour trans-methQTL-CpGs are depleted in breast-specific tumour PMDs .... 176

6.6 – Trans-methQTL-CpGs share some similar mechanisms of methylation change with cis-methQTLs, through CTCF enrichment, bivalent promoter depletion and eQTL overlap .......... 177

6.7 – Discussion .................................................................................................................. 180

Chapter 7 – Discussion and conclusions 186

7.1 - Main findings of the project ......................................................................................... 186

7.2 - Implications and relevance to the broader field .......................................................... 190

7.3 - Limitations of project .................................................................................................. 192

7.4 - Future Work ............................................................................................................... 198

7.5 - Concluding Statement ............................................................................................... 201

References 186
Table of Figures

Figure 1 – Distribution of CpG methylation across the genome of normal cells .........................20
Figure 2 – Methylation states are dynamically regulated by active and passive pathways for enzymatic methylation and demethylation of CpG site .................................................................23
Figure 3 – The cancer methylome undergoes hallmark changes of global hypomethylation and focal hypermethylation .........................................................................................................31
Figure 4 – Mechanism for cis-methQTLs .................................................................................39
Figure 5 – Mechanism for trans-methQTLs .............................................................................40
Figure 6 – Proportionally enriched/depleted distribution of methQTL-CpGs in functional genomic features .........................................................................................................................44
Figure 7 – ssNoob normalisation reduces technical variation between replicates ...................73
Figure 8 – Functional and ssNoob normalisation improves bimodality of beta distribution over raw normalisation ..............................................................................................................75
Figure 9 – Undetected probe percentage has negative correlation with sample quality ......78
Figure 10 – Methylation samples with low undetected probe percentage have greater bimodality in breast tumours ......................................................................................................79
Figure 11 – TCGA ancestry annotations form discrete genetic clusters .................................86
Figure 12 – Classifying EU ancestry by genetic similarity increases genetic similarity and size of EU population cohort ........................................................................................................88
Figure 13 – Matrix eQTL identifies methQTLs of variable quality in breast tumours ..........97
Figure 14 – RegScan-based GWAS pipeline identifies non-significant methQTLs associated with cg15442702 methylation ........................................................................................................99
Figure 15 - RegScan-based GWAS pipeline identifies a significant cis-methQTL associates with cg02113055 methylation ........................................................................................................101
Figure 16 - Z-score filtering removes outlier samples and improves quantile distribution for methQTL analysis of cg15442702 ..................................................................................................102
Figure 17 - Additional variant/CpG quality filtering from Regscan-based GWAS method resolves artefact issues from Matrix eQTL analysis of cg15442702 ........................................104
Figure 18 - Transforming methylation data from beta values to M-values has a comparable effect on the methQTL analysis of cg15442702 .......................................................... 108

Figure 19 - Inclusion of methylation structure as covariate identifies unique set of methQTLs for cg15442702 ........................................................................................................ 111

Figure 20 - A strict imputation quality score threshold identifies comparable methQTLs with a nominal imputation score when analysing cg15442702 ........................................................................................................ 115

Figure 21 – Frequency of SNPs and CpGs in methQTLs identified before LD clumping had been applied .......................................................................................................................... 126

Figure 22 - Description of clumps and their associations with CpGs and secondary SNPs ...127

Figure 23 – Breast tumour methQTL-SNPs share loci with normal blood methylation QTLs ........................................................................................................................................... 129

Figure 24 – Tumour methylome shows more balanced distributional *cis:trans* ratio across the tumour genome compared with normal methylome ................................................. 131

Figure 25 – The effect size of *trans*-methQTLs is smaller than that of *cis*-methQTLs in breast tumours ........................................................................................................................................... 133

Figure 26 – The ratio of *cis:trans* increases with respect to stricter significance thresholds134

Figure 27 - rs3768921 is a significant *cis*-methQTL for methylation at cg00004073 ..........142

Figure 28 – Haploblock of genetically linked SNPs around rs3768921 colocalises with AGAP1 gene promoter ...................................................................................................................................... 144

Figure 29 – SNPs with greatest genetic linkage within rs3768921 haploblock show higher regulatory potential .......................................................................................................................... 145

Figure 30 – Breast tumour cis-methQTL-CpGs are depleted in CpG islands and promoters across Illumina reference genome annotations .......................................................................................................................... 149

Figure 31 – Breast tumour cis-methQTL-CpGs are depleted in promoters and CpG islands in *chromHMM* genome annotations across tumour and normal mammary tissue........151

Figure 32 - Breast tumour *cis*-methQTL-CpGs are depleted in common tumour, but not breast tumour, PMDs and HMDs .......................................................................................................................... 153

Figure 33 - *Cis*-methQTL-CpGs are relatively enriched in less conserved PMDS and depleted in more conserved PMDs across 30 breast tumour samples compared with non-methQTL-CpGs ........................................................................................................................................... 154

Figure 34 - Breast tumour *cis*-methQTL-SNPs share loci with normal breast expression *cis*-QTLs ........................................................................................................................................... 157
Figure 35 – Bivalent promoters may undergo stable DNA methylation through CGI hypermethylation or oncogene activation through CGI hypomethylation in tumours ........161

Figure 36 – CTCF local protection against hypermethylation is lost when CTCF binding region is disrupted by genetic variant.................................................................163

Figure 37 - rs2008600 is a significant trans-methQTL for methylation at cg00160777 ........169

Figure 38 – Haploblock of genetically linked SNPs around rs2008600 colocalises with CCSER1 intron..........................................................................................................................170

Figure 39 – SNPs with highest genetic linkage within rs2008600 haploblock have low regulatory potential.............................................................................................................172

Figure 40 – Breast tumour trans-methQTL-CpGs are enriched in CpG islands and promoters but depleted in enhancers ................................................................................................174

Figure 41 – Breast tumour trans-methQTL-CpGs are enriched in CpG islands and promoters but depleted in enhancers for chromHMM chromatin annotations across tumour and normal mammary tissue ................................................................................................176

Figure 42 – Breast tumour trans-methQTL-CpGs are depleted in PMDs but enriched in HMDs in common tumour annotations ..................................................................................178

Figure 43 – Trans-methQTL-CpGs are relatively depleted in both well and poorly conserved PMDs across 30 breast tumour samples compared with non-methQTL-CpGs ..................179

Figure 44 – Breast tumour trans-methQTL-SNPs share loci with normal breast expression cis-eQTLs ........................................................................................................................................181

Figure 45 – Cis and trans-methQTL mechanisms are largely similar across normal and tumour breast tissue methylome ................................................................................................189
Table of Tables

Table 1 - ChromHMM chromatin-derived genomic feature annotations ........................................64
Table 2 - TCGA ancestry annotations for cohort ........................................................................84
Table 3 – Methylation data quality control steps......................................................................89
Table 4 – Genotype data quality control steps.........................................................................91
Table 5 – Post imputation filtering and modelling decisions to optimise methQTL analysis of
breast tumours.............................................................................................................................117
Table 6 - Regulatory potential of SNPs with highest $R^2$ and within LD block of rs3768921 ..147
Table 7 – Regulatory potential of SNPs with the highest $R^2$ within LD block of rs2008600 ..173
Abstract

An important factor in expanding our knowledge of transcriptional regulation has been to better understand the role of epigenetic modifications in mediating transcriptional activity. One hallmark of this relationship is the observation that gene activity associates with DNA methylation patterns. Abnormal changes to the methylome are a pathological feature of many cellular disease and a hallmark of cancer. However, the mechanisms underpinning these alterations remain unclear. Current models in normal tissues have suggested that some methylation changes may occur in an allele-dependent manner whereby genetic variation associates with both local and global changes in the normal tissue methylome. Investigating the mechanisms and directionality of the interplay between DNA methylation, sequence variation and their role in gene regulation may therefore reveal important insights into tumorigenesis. Previous studies have established the effects of genetic variation on the methylome of normal tissue. However, a similar characterisation of the genetic effects on the cancer methylome has been limited due to the complexity of analysing tumour tissues which have highly heterogeneous cell content compositions.

I used a population genetics approach to examine the effects of genetic sequence variation on the methylome of breast tumours. I asked if there were loci which showed an allele-dependent association with the methylation state of CpGs (termed methQTLs) either locally in cis or distally in trans. I then asked if I could use the distinct mechanisms which underpin cis-methQTLs and trans-methQTLs to characterise the role these genetic variants may have in regulating the methylome. I developed a robust analytical framework for detecting methQTLs using breast tumour methylation data and matched normal blood germline genotype data. I compared and assessed various methods for the pre-processing, normalisation and quality control-based filtering of these datasets to establish an optimal inclusion cohort of both subjects and samples. Furthermore, I compared different per-allele association methods for identifying SNP/CpG associations and different analysis conditions (such as covariate data, statistical assumptions and confounding linkage effects) to ensure the accurate identification of independently-associated methQTLs.
I used this analytical framework to identify methQTLs across 333 subjects with matched genotype and methylation data. I identified 446,482 significant methylation QTLs, of which 13,195 were independently associating loci. There were 6,725 distinct SNPs which associated with methylation changes at 5,779 unique CpG sites. A higher proportion of breast tumour methQTLs were acting in cis (10,500) than in trans (2,695), although this ratio was more balanced relative to normal blood methQTLs. I then characterised the mechanisms underpinning cis-methQTLs and trans-methQTLs in breast tumours by analysing their enrichment in genomic elements which have functional importance. There was a depletion in tumour cis-methQTL-CpGs from CpG islands (both promoters and enhancers). Additionally, there was cis-methQTL-CpG enrichment in CTCF sites and depletion in bivalent promoters and repressed regions. This suggests that methylation change occurs in CpG island regulatory elements which are typically unmethylated and targets for hypermethylation in tumours, and therefore support a mechanism whereby instances of local hypermethylation may disrupt TF binding regions (such as CTCF sites) and polycomb-repressed bivalent promoters. In trans-methQTLs however, there was CpG enrichment in CGIs and promoters but depletion in enhancers. Meanwhile, CTCF site enrichment and bivalent promoter depletion were seen commonly across both cis- and trans-methQTL-CpGs. This suggests that methylation changes are primarily at distal promoters and supports the hypothesised mechanism that these are primarily cis-QTL mechanisms for the expression of transcription factors that subsequently programme long-range methylation effects. Overall, the concordance between the genomic profiles of methQTLs in both tumour and normal breast methylomes suggests that trans-methQTL mechanisms are more consistent compared with cis-methQTL mechanisms.

Taken together, these results show that germline genetic variation has a strong relationship with DNA methylation in breast cancer which associates with distinct methylome alterations from the normal blood methylome. Furthermore, this relationship may have a significant role in breast cancer biology by the mediation of transcriptional regulation through both local and long-range mechanisms.
Our DNA outlines specific instructions to generate and maintain a healthy cell. However, the interpretation of these instructions is dependent on how accurately our cells can read and process the information that our DNA conveys in our genes. While DNA sequence primarily defines these instructions, DNA may also be marked with chemical modifications which affect the accessibility of the sequence to be read without changing the sequence. These are called epigenetic modifications, and they contribute to the control of how active a gene may be through either direct or indirect interactions with the DNA sequence. DNA methylation is a direct epigenetic mark which appears abundantly across the genome and exclusively where the cytosine base is paired with a subsequent Guanine (CpG) in mammal DNA. Interestingly, when high DNA methylation is found within normally unmethylated gene regulatory elements (termed promoters and which coordinate gene regulation), these genes are subsequently switched off, suggesting a potential role in regulating gene activity.

Importantly, DNA methylation is commonly dysregulated in cellular disease, whereby abnormal methylation changes may interfere with how these instructions are being interpreted by the cell. This dysregulation is especially prominent in cancer, whereby methylation is globally lost but may also be gained at the promoters of genes with significant roles in preventing tumorigenesis. Despite these observations, the exact mechanisms which underpin the abnormal tumour methylomes are not well understood. One way to find the specific sequence regions which underpin these mechanisms is to look for DNA regions where a sequence variant associates with distinct methylation changes. By this approach, models for how these mechanisms may be driven have been hypothesised which depend on finding methylation-associated variants and also on whether these associations are short-range or long-range.

In my PhD project, I studied how genetic variants associate with the breast tumour methylome to elaborate on our understanding of the underlying genetic mechanisms. I
prepared a robust set of criteria for ensuring the methylation and genetic sequence data were of a high quality and then I compared different methods for analysing these associations. Using the methods which performed best according to my criteria, I then looked for associations between genetic variants and the methylation patterns of 333 breast tumour patients. I found 13,195 genetic variants which had a direct and significant association with the tumour methylation patterns. I compared the methylation patterns of tumour and normal cells by contrasting whether or not these associations were between variants and methylation over short or long distances. I found that normal cell methylomes have a greater bias for finding short-range associations between methylation changes and sequence variants.

Finally, I investigated the “mechanism profile” of short-range and long-range associations individually through analyses of the regions in which they were found more often than would be expected by chance, and the functional importance of those regions. For short range associations, methylation changes were less likely to be found in regulatory control regions and may therefore support a mechanism where a sequence variant can compromise the binding regions for important proteins. For long-range associations, methylation changes are more likely to be found in long-range promoters and support a hypothesis that these sequence variants may control the activity of a local gene, which is subsequently responsible for changing methylation patterns elsewhere at distal parts of the DNA sequence. This reveals important insights in to how abnormal breast cancer methylation changes are mediated by genetic variation.
Chapter 1 – Introduction

1.1 – Epigenetics

The classic model of transcriptional regulation in the genome is underpinned by a complex network of sequence-specific transcription factors which regulate gene expression. Driven by sequence variation and cellular signalling, these transcription factors cooperate to activate or silence gene activity and regulate complex molecular pathways. However, as early as 1942, Waddington alluded to phenotypic changes which were not necessarily accountable to genotypic changes, and thus required explanation and distinct categorisation. Therefore, in contrast to the genetic component of gene regulation, epigenetics has been described to represent genome-wide transcriptional regulation which occurs independently from genetic variation (Inbar-Feigenberg et al., 2013).

There have been various definitions of epigenetics which primarily involve DNA modifications but which may also include important conditions (such as the heritability of DNA methylation) or indeed broader chromatin biology (such as non-coding RNAs) as key components of the term (Fincham, 1997; Volpe et al., 2002; Martienssen and Moazed, 2015). However, I will use an interpretation by Bird (2007) which focusses on the chemical and structural modifications made to DNA molecules which either propagate or signal changes to the activity of a gene. This definition attempts to unify various stricter interpretations of such marks (which, for example, may argue over the transience of heritability in their classification of epigenetic modifications) by simply considering them as responsive to the different mechanisms which may alter expression changes.

Under this definition, epigenetics has been studied for its role in transcriptional activity at both the gene and chromatin level. Chromatin has been classified in to two broad types based on its availability for transcriptional activation. Euchromatin is a term used to describe active chromatin that is open, loose and associates with high transcription factor activity, whilst heterochromatin describes chromatin which is compact and consequently
associates with transcriptional repression (Quina, Buschbeck and Di Croce, 2006). There are two primary epigenetic modifications which associate with regulatory control of the transcriptional state of the genome: histone modifications and DNA modifications (Golbabapour, Abdulla and Hajrezaei, 2011). These epigenetic modifications largely mediate transcriptional activity of the genome through distinct mechanisms, but may also co-operate to direct spatial and temporal patterns of these modifications (Allis and Jenuwein, 2016).

1.2 - Epigenetic histone modifications

Distinct profiles of epigenetic modification combinations at histone proteins in nucleosomes have been shown to subsequently associate with different transcriptional states of chromatin (Millán-Zambrano et al., 2022). Histones comprise the structural packaging of DNA molecules, which are wrapped around an octamer of the histone proteins H2A, H2B, H3 and H4 to form a stable nucleosome with DNA (McGinty and Tan, 2015). These nucleosomes are then connected by open linker DNA which is stabilised by linker histones, such as H1 (Hergeth and Schneider, 2015). While linker DNA is thought to be accessible to transcription factors and therefore active, DNA tightly wound in nucleosomes is considered inaccessible and thus inactive (Radman-Livaja and Rando, 2010). Nucleosome positioning and compaction can therefore facilitate or impede transcription regulation of gene expression (Radman-Livaja and Rando, 2010; Jiang and Zhang, 2021).

The chemical modifications found on histone proteins, most prominently on the tails of histones, have been well catalogued (Millán-Zambrano et al., 2022). The most common classes of modification include methylation, acetylation, ubiquitination and phosphorylation and these are most commonly found on lysine residues (Bannister and Kouzarides, 2011). Some of these modification classes show consistent effects on transcriptional activity (Millán-Zambrano et al., 2022). For example, active genes are dynamically marked with acetylation in H3 proteins - most commonly at H3K9 and H3K27 - by histone acetylation transferase enzymes, and which are subsequently removed by histone deacetylase enzymes.
Most of these modification classes, however, have complex contextual effects to mark active or inactive chromatin (Millán-Zambrano et al., 2022). For example, histone trimethylation at H3K4 in promoters/enhancers or in H3K36 in gene bodies may mark active gene expression, whereas histone trimethylation at H3K9 in repressed gene promoters and in constitutive heterochromatin (or H3K27 in bivalent promoters which have both active and repressive histone modifications) all associate with inactive gene activity (Lan and Shi, 2009).

1.3 – Epigenetic DNA modifications

Epigenetic regulation also includes the complex patterns of sequence-specific modifications made directly to the DNA nucleotides to mediate transcriptional regulation (Breiling and Lyko, 2015). The most common DNA modification in animals is DNA methylation, which occurs almost exclusively in the context of cytosine-guanine dinucleotide (CpG) sites, whereby a methyl group is added to the 5’-end of the cytosine base by DNA methyltransferase enzymes (Bird, 2002; Breiling and Lyko, 2015). While a CpG site most commonly shows symmetrical methylation on both the forward and reverse DNA strands, a methylated CpG may also be briefly in a hemi-methylated state during DNA replication step of the cell cycle before the new strand has been able to be methylated (Petryk et al., 2021).

There are a few alternative instances of base methylation observed in the human genome. There are also hydroxymethyl-cytosines, one intermediate from the active enzymatic oxidation-driven demethylation pathway of 5-methylcytosine back in to cytosine, but which also has a distinct genomic distribution and associates with discrete regulatory and tissue-specific developmental functions separately from 5’-methylcytosine (Shi et al., 2017; He et al., 2021). Besides cytosine, other methylated bases have been observed, such as methyl-adenine which is primarily seen in bacteria and plants but which was also reported in the human genome, although this observation in mammals has been challenged primarily as being confounded by contamination (Kumar, Chinnusamy and Mohapatra,
2018; Bochtler and Fernandes, 2021; Kong et al., 2022). For this reason, human DNA methylation is studied in the context of measuring patterns of CpG methylation.

1.4 — DNA methylation patterns

CpG sites are non-randomly distributed across the genome according to their methylation state (Kaplun et al., 2022) (Figure 1). In mammals, between 70-80% of CpG sites are fully methylated (Ehrlich et al., 1982; Suzuki and Bird, 2008). However, CpGs also make up small clusters (greater than 200 base pairs) with high CpG-density termed CpG islands (CGIs), which are canonically unmethylated (Bird, 1986; Gardiner-Garden and Frommer, 1987). This conservation of hypomethylation comes from both the protection of high CpG density regions against spontaneous deamination (relative to rest of genome), and from the CpG collaborativity which maintains hypomethylation in CGIs (Haerter et al., 2014; F. Chen et al., 2018). This hypomethylation may therefore maintain active CGIs.

Unmethylated CGIs have functional significance to transcription as they are commonly enriched in regulatory elements of the genome, such as promoters and enhancers (Rollins et al., 2006; Cain, Montibus and Oakey, 2022). Indeed, whilst over half of CGIs are normally unmethylated and overlap with transcription start sites in 60-70% of gene promoters, the other half are distal (primarily in gene bodies but also occasionally in intergenic regions, which are called orphan CGIs) and are enhancers (Deaton and Bird, 2011; Bell and Vertino, 2017). Importantly, we know that hypomethylation of these regulatory elements is important for maintaining active gene expression, as a gain of methylation in these regulatory elements has been shown to associate with transcriptional repression (Weber et al., 2007; Deaton and Bird, 2011). This is considered to most likely occur either sterically (by impeding sequence accessibility for certain transcription factors) or possibly also through the broader promotion of inactive heterochromatin (Deaton and Bird, 2011; Deaton et al., 2011). However, the directionality of this association is still not clear; for example, to what extent is this mechanism driven by DNA methylation repressing promoters
Figure 1 – Distribution of CpG methylation across the genome of normal cells

CpGs are largely methylated across the genome (70-80%), but active, CGI-rich regulatory elements (such as promoters and enhancers) are unmethylated. Gene bodies, heterochromatin and other inactive genomic regions are largely methylated. Adapted from Schübeler (2015).
as a proactive mark compared with other promoter-repressive mechanisms directing DNA methylation as a reactive mark? The mechanisms underpinning the methylation state of promoter CGIs therefore appear to be highly predictive of gene activity, but the association is still not clearly understood.

Methylation patterns outside of promoter CGIs also have distinct functional roles, primarily through the maintenance of genomic stability (Kaplun et al., 2022). Gene bodies have been shown to be fully methylated in association with highly expressed genes, and this hypermethylation has been proposed to protect genomic stability by silencing alternative ectopic regulatory elements (such as promoter CGIs) which may reside within the gene body (Maunakea et al., 2010; Neri et al., 2017). In intergenic regions, most CpGs are fully methylated and found in inactive regions of chromatin. Repressed chromatin regions may harbour retrotransposons, and so maintaining high methylation of these inactive regions may contribute to genome stability by repressing transposable elements after the embryogenic reset of germ line methylation has occurred (Zamudio et al., 2015). DNA methylation also shows broader regulatory roles beyond local sequence effects, such as chromosomal transcriptional regulation (Kaplun et al., 2022). For example, alongside histone modifications, DNA methylation has a significant role in the imprinting of maternal or paternal alleles which balances the mono-allelic or bi-allelic expression patterns that genes require in certain tissue and spatio-temporal contexts (Elhamamsy, 2017). DNA methylation patterns also has a key role in dosage compensation through the inactivation of one X-chromosome in females, whereby almost all promoter CGIs become methylated in the inactive chromosome (Gendrel and Heard, 2011; Sharp et al., 2011).

Beyond the general methylation patterns which have been observed, comparisons between the DNA methylomes of various different tissues have identified conserved tissue-specific methylation patterns within transcriptional regulatory elements that suggest a role for methylation in tissue-specific gene regulation (Zhou et al., 2017). From whole-genome bisulfite-sequencing, ~20% of CpG sites are predicted to show methylation variation between tissues (Ziller et al., 2013). Genic regions which show tissue-specific differential methylation, primarily hypomethylation in regulatory elements and less commonly hypermethylation in the gene body, associate with tissue-specific gene expression patterns
and cellular identity (Rakyan et al., 2010; Lokk et al., 2014). Indeed, Blake et al. (2020) compared four tissue types from matched individuals and showed that the number of tissue-specific differentially methylated regions (DMRs) varies between tissues (i.e. they detected more DMRs detected between liver and heart tissues than between heart and lungs) and are enriched distally to the transcriptional start site (mainly enhancers and untranslated regions). They also saw that ~60% of conserved tissue-specific DMRs (only differentially methylated in one of the four tissues) showed conserved tissue-specific hypomethylation in relevant functional genes for that tissue (for example, cardiovascular pathways for heart tissue DMRs) (Blake et al., 2020).

1.5 – DNA methylation mechanisms and enzymology

The establishment and maintenance of the DNA methylation landscape is dynamic and involves complex enzymology, the balance of which is essential for maintaining a healthy cell (Figure 2). DNA methylation is remodelled throughout the cell cycle and regulated across tissue types by DNA methyltransferase enzymes: primarily through DNMT1, DNMT3A and DNMT3B (Ren, Gao and Song, 2018). The roles of these proteins are categorised in to two groups: maintenance of methylation patterns or establishment of de novo methylation (Goll and Bestor, 2005). Maintenance methylation is defined as the recapitulation of fully methylated CpG sites from a hemi-methylated state during S-phase (Nishiyama et al., 2020; Petryk et al., 2021). DNMT1 is primarily responsible for maintenance function alongside its cofactor UHRF1, which recognises hemi-methylated DNA and recruits DNMT1 (Takebayashi et al., 2007; Xie and Qian, 2018; Petryk et al., 2021).

Alternatively, de novo methylation is established at unmethylated CpG sites across the genome by the DNMT3 proteins (Kato et al., 2007). The two forms of DNMT3 (3A and 3B) are responsible for de novo methylation and are additionally stabilised by the non-catalytic DNMT3L protein (Kato et al., 2007; Andrews et al., 2023). DNMT3A/3B are highly expressed during embryo development and less so as cells become differentiated, but they do maintain some unique functions which relate to distinct patterns of temporal and spatial
Methylation states are dynamically regulated by active and passive pathways for enzymatic methylation and demethylation of CpG site.

Methylation of CpGs is regulated by the DNMT family of methylase proteins. DNMT1 has the canonical role of targeting hemi-methylated CpGs (following DNA replication) and therefore maintains fully methylated CpGs. DNMT3 isoforms have the canonical role of targeting unmethylated CpGs and introducing de novo methylation. Meanwhile, active demethylation is regulated by the TET family of dioxygenase enzymes. Alternatively, methylation can also be lost passively by replication-dependent DNMT1 deficiency through successive cell divisions. Adapted from Lebecque et al. (2021).
expression as well as preferential sequence targeting (Chen et al., 2003; Auclair et al., 2014; Lyko, 2018). Primarily, DNMT3B is thought to methylate broadly in early embryo development, whereas DNMT3A may methylate at late embryo development and after birth (Gao et al., 2020). However, the exact roles and inter-dependence of the DNMT family of proteins on one another is still unclear as some degree of redundancy has been reported in these proteins; a maintenance role can be compensated by DNMT3 proteins in the absence of DNMT1, and in turn, DNMT1 has shown context-specific de novo methylation capability (Dodge et al., 2005; Jair et al., 2006; Feng et al., 2010; Hervouet et al., 2018; Haggerty et al., 2021).

With regard to chemical structure of the DNMT proteins, DNMT1 and DNMT3 share a regulatory N-terminus and a catalytic C-terminus, but do show some distinct differences in their domain structures. DNMT1 has a key CXXC domain which enables the targeting and binding of DNMT1 to unmethylated CpGs, which then promotes the subsequent catalytic activation of DNMT1 and therefore directs methylase activity (Pradhan et al., 2008; Kikuchi et al., 2022). Meanwhile, the two DNMT3 isoforms also maintain unique functionality despite being structurally similar (Ren, Gao and Song, 2018). DNMT3A and DNMT3B are both recruited to chromatin marks by the same set of histone-reader domains in the N-terminus: the ADD domain and PWWP domain (Jeltsch and Jurkowska, 2016). Both DNMT3A and DNMT3B use the PWWP domain to bind H3K36me3 in active gene bodies for methylation, whilst DNMT3A may preferentially bind H3K36me2 in intergenic regions and heterochromatin (Qin and Min, 2014; Weinberg et al., 2019). DNMT3A, but not DNMT3B, uses the ADD domain as an autoinhibitory mechanism through the allosteric activation of the protein upon binding unmethylated H3K4 to reduce protein activity at active promoters marked with H3K4me3 (Otani et al., 2009; Boyko et al., 2022). Additionally, the organisation of the N-terminus in DNMT3A has a regulatory role, whereby the long DNMT3A1 isoform containing an intact N-terminus has shown isoform-specific functions in post-natal development and bivalent polycomb-regulated neuro-developmental genes, and most importantly retains methylation abilities, whereas the short (DNMT3A2) isoform which lacks the N-terminus shows none of these functions (Gu et al., 2018; Weinberg et al., 2021).
DNA methylation exists in a dynamic balance where methylation can be lost and regained, particularly during development and cell differentiation (Luo, Hajkova and Ecker, 2018). Before birth, an embryo has already undergone two waves of DNA methylation erasure; firstly in the germline before sex-specific re-methylation and then subsequently during early embryogenesis to achieve totipotency before further embryonic development, after which embryonic methylation patterns have been accurately re-established (Zeng and Chen, 2019). As such, the cell must also have mechanisms for demethylation of DNA, and this loss of methylation may occur either passively or actively. Passive loss of DNA methylation occurs primarily from replication error-associated deficiency in the methylation maintenance machinery (i.e. DNMT1), which results in the gradual dilution of methylated CpGs across the genome (Wu and Zhang, 2014). This passive demethylation is involved in both waves of methylation loss (Wu and Zhang, 2014).

In contrast to passive methylation loss, there are also enzymes which actively drive rapid, replication-independent demethylation of CpGs. The process of active demethylation has been associated with the Ten-eleven translocation (TET) family of dehydrogenase proteins, which includes TET1, TET2 and TET3 (Melamed et al., 2018). While they share the same catalytic activity, TET1 and TET2 are ubiquitous in embryonic stem cells and are required for differentiation whereas TET3 is more present in oocytes and neurons (Wu and Zhang, 2017; Melamed et al., 2018). TET1 has been shown to convert the 5-methylcytosine (through oxidation intermediates) to a hydroxymethyl-cytosine, which can then be converted back into unmethylated cytosine by glycosylase enzymes (Tahiliani et al., 2009; He et al., 2011; Liu et al., 2021). Active demethylation also has a role in the first wave of germ line demethylation and a smaller role in the embryogenic demethylation of specifically the paternal genome (Zeng and Chen, 2019). This cyclical processes of methylation and demethylation maintains a dynamic balance of DNA methylation which defines development and cell identity.
1.6 – Relationship between transcriptional regulation and DNA methylation in normal tissue

DNA methylation has shown significant association with gene expression in normal healthy tissues (Moore, Le and Fan, 2013). As mentioned, one of the key observed associations has been between unmethylated CpG islands in regulatory elements and active transcriptional regulation, which has been suggested to protect transcription factor binding in these elements (Mohn et al., 2008; Moore, Le and Fan, 2013). This was first posed by the observation that when a promoter CGI for an active gene becomes hypermethylated, there is an associated reduction in gene expression which could be attributed to the binding of methyl-CpG binding proteins and the subsequent disruption of transcription factor binding (Boyes and Bird, 1991, 1992). Since then, DNA methylation has been shown to have a regulatory role both specifically at transcription factor binding sequences and also more broadly at inactive chromatin through an interaction with alternative molecular mechanisms of gene repression (Kaplun et al., 2022).

At CGI regulatory elements, DNA methylation may regulate gene expression by directly interacting with transcription factors, but the mechanism remains to be established (Yin et al., 2017). This was considered to primarily be through the steric disruption of transcription factor binding, whereby proteins could bind unmethylated sequences but not to methylated binding sites (Domcke et al., 2015; Zhu, Wang and Qian, 2016). However, some transcription factors may alternatively preferentially bind methylated CpG sites (Yin et al., 2017). Indeed, some proteins have been shown to bind distinctly separate sequences depending on the methylation context of the binding region, such as KLF4 (Hu et al., 2013). As mentioned previously, many proteins use the CXXC domain to target unmethylated CpGs or CGIs, such as DNMT1 and the TET family, amongst others (Xu et al., 2018). While not transcription factors themselves, one way methylation may also indirectly affect transcription factor activity is through the recruitment of methylation-binding domain (MBD) proteins (such as MECP1) to specifically bind methylated CpGs in inactive CpG-dense regions; these may subsequently either occupy and impede transcription factor binding
regions or recruit either chromatin remodelling or methylase proteins to alter methylation patterns (Baubec et al., 2013).

Once bound, a transcription and DNA-binding factors may reprogram local methylation either directly or with the assistance of co-factors (Zhu, Wang and Qian, 2016). For example, for the DNA-binding factor CTCF (involved in imprinting), occupancy at its binding site associates with local methylation loss (Stadler et al., 2011), while it has been suggested that CTCF can bind regardless of methylation state before mediating any local demethylation to maintain active DNA elements (Feldmann et al., 2013). In another example, TET1 is recruited (by the PPY1 nuclear receptor) as a co-factor to demethylate the local DNA sequence (Fujiki et al., 2013). Thus, DNA methylation can be directly regulated by transcription and DNA-binding factors.

While DNA methylation has an independent role as an epigenetic modification, it is worth highlighting how it has also been shown to cooperate alongside other molecular mechanisms to direct chromatin repression and promote inactive chromatin, often through the recruitment of methylation proteins (Laisné et al., 2018). While I have previously discussed histone protein marks as an independent epigenetic mark, these modifications cooperate with DNA methylation by recruiting and directing DNMT enzymes and their methylase activity through characteristic histone mark profiles which correspond with active or repressed gene/chromatin activity and at different genomic features (Rose and Klose, 2014). For example, while both DNMT3A and DNMT3B bind to H3K36me3 in gene bodies, DNMT3A preferentially binds H3K36me2 in inactive gene bodies and intergenic regions (Weinberg et al., 2019). Similarly, whilst DNMT3B is directed to gene bodies through H3K36me3 marks, it has been recently suggested that the DNMT3B has a role conserving the homeostasis of methylation at H3K9me3-marked heterochromatin against hypomethylation, facilitated by the N-terminus of the protein (Taglini et al., 2022).

DNA methylation also has an important relationship with the polycomb-mediated system of transcriptional repression, another major chromatin repressive mechanism (Morey and Helin, 2010). Polycomb mediates the repression of developmental genes which require to be poised for rapid activation from a repressed state in different tissue contexts.
(Boyer et al., 2006). Such genes have bivalent promoters that contain both active (H3K4me3) and repressive histone marks (H3K27me3) so they may be readily switched “on or off” (Bernstein et al., 2006). There are two polycomb-repressive complexes (PRC) which enact this repression. H3K27me3 marks are first established by the PRC2 complex at bivalent promoters, alongside H3K4me3, and PRC1 in turn is often recruited to H3K27me3 and drives chromatin compaction (Morey and Helin, 2010). As such, polycomb can be considered a relatively transient and reversible mechanism of repression in comparison to a more robust repressive mark in DNA methylation (Bird, 2002).

In fact, polycomb-mediated repression and DNA methylation commonly show a mutually exclusive relationship at bivalent promoters, whereby the presence of H3K27me3 marks or DNA methylation may antagonise one another (Mager et al., 2003; Grossniklaus and Paro, 2014; van Mierlo et al., 2019). Indeed, polycomb may promote hypomethylation at developmental genes through TET proteins (Li et al., 2018). Similarly, PRC2 has been suggested to preferentially bind unmethylated CGIs over heavily methylated genomic regions as a loss of DNA methylation results in H3K27me3 being distributed away from the promoter, along with PRC2 activity (Reddington et al., 2013; McLaughlin et al., 2019; Uckelmann and Davidovich, 2021). DNA methylation can therefore interact with polycomb-mediated repression to regulate the activity of chromatin in different contexts.

1.7 – Dysregulated DNA methylation states in disease and cancer

DNA methylation dysregulation is a defining characteristic of poor cellular health and has been studied as a hallmark, and potentially useful biomarker, for numerous cellular diseases and overall cell health (Jin and Liu, 2018). For example, DNA methylation dysregulation is a hallmark of ageing which correlates with cell age and proliferation (López-Otín et al., 2013). During ageing, the functional capacity of the cell becomes steadily poorer as a consequence of the declining fidelity with which molecular damage is able to be repaired (López-Otín et al., 2013). Correlated DNA methylation for most tissues has been shown to be globally hypomethylated but focally hypermethylated at certain gene
promoters which are suggested to associate with regulating age-associated processes, such as cell proliferation (Fuke et al., 2004; Unnikrishnan et al., 2018).

From these observations, epigenetic clocks have been established to leverage consistently conserved DNA methylation at sufficiently reliable CpG sites. These may predict a molecular age that reflects the chronological age while also accounting for the effects of environmental stresses which have been shown to exacerbate epigenetic/methylation changes and functional decline in tissues, such as obesity or air pollution (Horvath and Raj, 2018). Furthermore, these clocks have been shown to be reliable both between and within different tissue and animal species (Horvath and Raj, 2018; Ake Lu et al., 2021). This suggests that cell age and proliferation have an important and evolutionary conserved relationship with DNA methylation and that age-related changes in DNA methylation may therefore have some role in propagating genomic instability through transcriptional dysregulation.

Abnormal methylation patterns are also seen in age-related diseases which show high proliferation and differentiation, suggesting a role for DNA methylation changes in enabling cellular disease. For example, hypertension and other cardiovascular diseases show abnormally low levels of DNA methylation (Zhang and Zeng, 2016; Kazmi et al., 2020). Methylation dysregulation may therefore have a functional role in the pathogenesis of cardiovascular disease through the control of gene expression, and individual CpG sites have been reported as biomarkers for cardiovascular disease in population studies (Fernández-Sanlés et al., 2021). Alzheimer’s disease, amongst other neurodegenerative diseases, has also been linked to DNA methylation dysregulation and accelerated epigenetic aging through transcriptional dysregulation (Irmer and Jin, 2012; Huo et al., 2019).

1.8 – Hallmarks of the tumour methylome

Perhaps the most significant example of age-related aberrant proliferation in cellular disease however is cancer. Ageing and cancer have shared hallmarks in genomic instability
(due to somatic mutations) and in epigenetic dysregulation, whereby mutations to epigenetic modifiers drive the gene expression changes that underpin increased cell proliferation, aging and tumorigenesis (Shen and Laird, 2013; Darwiche, 2020). However, while the methylome patterns of cancer share some hallmarks with ageing and replicative senescence, tumours have shown distinct methylation changes which arise independently from cellular senescence, such as promoter hypermethylation in tumours (Xie et al., 2018; Chen, Ganz and Sehl, 2022). Regardless of whether these methylation changes occur dependently or independently of ageing and cellular senescence, focal hypermethylation and genome-wide hypomethylated domains have been established as two key hallmarks of the cancer methylome (Lakshminarasimhan and Liang, 2016) (Figure 3). However, it is still not fully understood how these two hallmarks manifest through the progression of cancer and if they have a direct role in transcriptional dysregulation.

The first hallmark feature of the tumour methylome is focal promoter hypermethylation. Focal hypermethylation of breast cancer-associated promoters in healthy breast tissue has shown that methylation is maintained and prevents senescence, showing an active role for promoter hypermethylation in cancer progression (Saunderson et al., 2017). A relatively rare instance of promoter hypermethylation (but with perhaps the most direct causal role in tumorigenesis) is of tumour-suppressor gene promoters (Esteller, 2002; Ng and Yu, 2015; Kim et al., 2019). Tumour suppressor genes are typically cell-cycle regulators or DNA repair proteins and may consequently enable cancer progression if they become silenced (Sun and Yang, 2010). One example is BRCA1, a tumour suppressor gene involved in both DNA repair and cell cycle control, which shows promoter hypermethylation in both breast and ovarian cancer (Chiang et al., 2006; Zhang and Long, 2015; Vos, van Diest and Moelans, 2018). In another example, hypermethylation and silencing of the MLH1 mismatch repair gene has a negative effect on the stability of microsatellite repeats (Li et al., 2013).

In contrast, the majority of focal promoter hypermethylation in tumours occurs at developmental genes which are often found alongside facultative heterochromatin and under polycomb-mediated repression (Reddington, Sproul and Meehan, 2014). Indeed, over 75% of hypermethylated CGI promoters are polycomb-repressed bivalent promoters which
Figure 3 – The cancer methylome undergoes hallmark changes of global hypomethylation and focal hypermethylation

Methylation in CpG islands (enriched in promoters and enhancers) are altered to enable oncogenic activity (hypomethylated regulatory elements) and reduce tumour-suppressor functionality (hypermethylated regulatory elements) in the cancer cell. Heterochromatin and gene bodies are hypomethylated. Adapted from Skvortsova, Stirzaker and Taberlay (2019).
are associated with various different polycomb proteins which comprise the PRC1 and PRC2 protein complexes (Easwaran et al., 2012). PRC2-regulated bivalent CpG island promoters are canonically hypermethylated and silenced in tumours, but there has also been evidence for upregulation (in CGIs within distal enhancers and in promoter CGIs which lose H3K27me3), whereby these upregulated genes show high tumour-specificity (Zheng et al., 2021). One possible mechanism has also been the role of EZH2, one of the three PRC2 complex proteins, which may bind DNMTs and methylate CpGs at H3K27 methylated chromatin (Viré et al., 2006). Mutations in DNMT proteins also show a role in the hypermethylation of polycomb-repressed bivalent promoters. It has been shown in mice studies that missense mutations in DNMT3A which compromise binding to H3K36me2/3 results in abnormal hypermethylation at polycomb-associated unmethylated bivalent promoters in developmental regions (Heyn et al., 2019). These promoters concurrently lose their bivalent histone marks which consequently drives abnormal growth disorders, such as microcephalic dwarfism (Heyn et al., 2019).

The second hallmark feature of the tumour methylome is those domains which show a loss of methylation across the genome. This hypomethylation was one of the earliest methylation changes observed in cancers (Gama-Sosa et al., 1983), but hypomethylation has since been shown to associate with tumour formation. For example, DNMT1 hypomorphs in mice (with a 90% loss of expression levels) resulted in genome-wide loss of methylation across all tissues and consequently developed T-cell lymphomas through chromosomal instability (Gaudet et al., 2003). Hypomethylation has been shown to occur primarily at intergenic and heterochromatic regions and to accelerate with tumour progression (Batra et al., 2021). These regions contain methylated repetitive regions and transposable elements that normally preserve the stability of the genome, but which are unmethylated in tumours (Batra et al., 2021). For example, LINE1 (long interspersed nuclear element 1) is a mobile retrotransposon which, when hypomethylated, is a predictor of tumour progression (Wolff et al., 2010; Karami et al., 2015). However, intragenic hypomethylation may also drive tumorigenesis if normally high levels of gene body methylation are lost, allowing alternative promoters or enhancers become derepressed (Kulis et al., 2012). Indeed, primary liver tumours show increased transcription of hypomethylated, CpG-poor alternative promoters,
alongside downregulated transcription of reference promoters, for genes which specifically drive hepatocarcinogenesis (Nepal and Andersen, 2023).

In cancer, hypomethylated regions in heterochromatin are called partially methylated domains (PMDs), which are gene-poor inactive regions that can span hundreds of kilobases and which colocalise with repressive histone marks (such as H3K9me3 and H3K27me3) (Lister et al., 2009; Hawkins et al., 2010). PMDs cover up to 40% of the genome and show reduced methylation levels below 70% of what is observed in normal tissue (Lister et al., 2009; Schroeder and LaSalle, 2013). PMD hypomethylation shows cell type-specific variation, with greater heterogeneity in differentiating and proliferating cells (Salhab et al., 2018). Occurring in late-replicating domains of the genome, PMD formation has been suggested to be a consequence of poor maintenance of DNA methylation during DNA replication, which leads to the gradual, cell division-dependent loss of CpG methylation (Endicott et al., 2022). PMDs have associate with specific sequence contexts (specifically with the “WCGW” base sequence) and with greater mitotic age (Zhou et al., 2018a).

Finally, one of the most striking features of cancer, which has been most challenging to tumour characterisation, has been the combination of cell-specific tumour methylation patterns and the heterogeneity of intra-tumour cell composition (Zhu et al., 2021). This heterogeneity is a combination of somatic genetic mutations, epigenetic changes and the tumour microenvironment (such as cell-cell interactions between invading, stromal and epithelial cells) (Lawson et al., 2018). As a consequence, bulk sequencing has had limited ability to disentangle cell-specific tumour methylation changes as cell subpopulations may show varied methylation patterns and collectively dictate tumour characteristics (such as metastasis or drug resistance) (Beyes, Bediaga and Zippo, 2021). This may also lead to the misclassification of molecular subtype from single-sample analyses.
1.9 – Breast tumour DNA methylation dysregulation

In the 2020 GLOBOCAN estimates of cancer prevalence and mortality, female breast cancer has become the most prevalent cancer for new diagnoses globally (representing 11.7% of total new global cancer diagnoses) and remaining the highest cancer mortality rate in females (16% of total female cancer deaths) (Sung et al., 2021). As these numbers are predicted to continue rising (to over 3 million global new cases and 1 million global deaths by 2040), it is crucial to understand how this disease functions at a molecular/epigenetic level, how it reacts to therapeutic approaches, and how epigenetic dysregulation may be leveraged to improve treatment options (Vietri et al., 2021; Arnold et al., 2022). For example, some promising studies using epigenetic treatments (such as DNMT inhibitors to lower DNA methylation) have shown a synergistic effect with anti-cancer therapeutics (specifically where the tumour has shown high drug resistance, such as triple negative breast cancer), but further work is needed to address concerns over reported cytotoxic side effects (Su et al., 2018; Buocikova et al., 2020). Furthermore, the improved analysis of epigenetic biomarkers from circulating tumour DNA in plasma could provide a reliable, non-invasive alternative for early diagnosis and screening that would eliminate the need for the invasive biopsies required for current biomarkers, and which have already shown promise in breast cancers (Zhao et al., 2014; Salta et al., 2018; Stewart and Tsui, 2018). A better understanding of the epigenetic changes in breast tumour is therefore a key requirement for improved therapeutic treatments.

Breast tumours show some breast-specific methylation patterns which relate to their specific pathology, and these can be attributed to the combination of gene and hormone receptor expression (although it is difficult to assign these changes to either one of these individually). Breast tumours are commonly classified in to five expression-based molecular subtypes which are defined by the transcriptional signature of a 50-gene classifier, called PAM50 (Parker et al., 2009). This primarily characterises the state of three hormone receptors in breast cancer patients: oestrogen receptor (ER+), progesterone receptor (PR+) and Human epidermal growth factor receptor 2 (HER2+) (Malhotra et al., 2010). These five molecular subtypes are, in decreasing order of their prognosis and histological grade:
normal-like, luminal A, luminal B, HER2-positive and basal-like triple negative (Sørlie et al., 2003; Malhotra et al., 2010). Luminal A and luminal B are both ER+ and PR+, whilst most are negative for HER2 receptor. There are three points of distinction for Luminal B compared to A: lower PR expression, higher proliferative markers (such as the Ki-67 antigen marker and mutations in the p53 gene) and less differentiation (Soliman and Yussif, 2016). Alternatively, HER2+ and basal-like are both ER- and PR-, whilst basal-like is also negative for the HER2+ receptor. Beyond these receptors, subtypes may be further defined by antigen markers, differentially expressed genes and non-coding RNAs which also acts as expression biomarkers for proliferation and subtype, and so are often included in these subtype profiles (Yersal and Barutca, 2014).

DNA methylation patterns in breast cancer has been used to identify epigenotypes, which describe discrete DNA methylation profiles for molecular subtypes as a means for improved subtype classification (Bediaga et al., 2010). There were three DNA methylation subtypes initially reported, which roughly translate to basal like (low methylation), luminal B (hypermethylated promoters) and luminal A, with HER2+ and normal-like subtypes unable to be confidently categorised from methylation data alone (Rønneberg et al., 2011; Stefansson et al., 2015). Moreover, machine learning approaches have been used to classify cancer subtypes using expression profiles and methylation profiles in to these predictors (Adorján et al., 2002; L. Chen et al., 2018). Holm et al. expanded on previous work where they used unsupervised clustering to collectively define seven methylation-based epitypes, which included normal-like, basal-like, HER+ and then four luminal epitypes (Holm et al., 2010, 2016). These epitype-based classifications however are still limited in their ability for accurate classification and require improved modelling to improve their reliability. For example, there is still difficulty in categorising luminal A and B subtypes from methylation data alone, whereby higher proliferation rates remain the strongest indicator of luminal B tumours. Indeed, a more recent study managed to identify luminal, basal and HER2+ subtypes from DNA methylation patterns using a supervised classifier of probes from the Infinium 450k array, but similarly could not differentiate between luminal A and luminal B (Chen et al., 2019). This suggests that we still do not fully understand how breast tumour methylation relates to molecular subtypes and cell heterogeneity.
Breast tumour studies have also related the general hallmark features of the tumour methylome to breast tumour-specific biology. Breast tumours which are positive for ESR1 (Oestrogen sensitive receptor 1) show focal hypermethylation at oestrogen-regulated enhancers, which control expression of key regulators for ESR1 and reduce ESR1 binding (Stone et al., 2015). Additionally, breast tumour PMDs contain genes which remain repressed due to chromatin repression and through repressive histone modifications (H3K27me3 and H3K9me3) which are mutually exclusive with DNA methylation (Hon et al., 2012). In fact, Brinkman et al. reported a relationship between these features in breast cancer by showing that PMDs are hypervariable across tumours and drive other aberrant methylome features (such as focal hypermethylation inside PMD CGIs) by promoting a strictly intermediate methylation level regardless of typically hypomethylated genetic elements (Brinkman et al., 2019).

1.10 – Relationship between genetic sequence and DNA methylation

Having already discussed models for the interaction between transcriptional regulation and DNA methylation, one key mechanism is how genetic sequence variation may associate with different CpG methylation states in an allele-dependent manner. Indeed, genetic factors may explain around 10-20% of methylation variation, especially at CpG sites which show variable and intermediate methylation levels (Gertz et al., 2011; Hannon, Knox, et al., 2018). This suggests there are loci where there is some allele-dependent genetic regulatory control over methylation changes, and these changes can arise from either germline sequence variation or somatic mutations. While somatic mutations progressively accumulate with age (and in replication-associated diseases, such as cancer), germline mutations can predispose the genome to further somatic mutation by disrupting cellular maintenance processes, such as DNA repair pathways (Ramroop, Gerber and Toland, 2019; Qing et al., 2020; Vali-Pour et al., 2022).

This relationship between DNA methylation and sequence variation manifests as allele-specific methylation patterns across the genome which are faithfully conserved across
individuals and tissue-specific (Tycko, 2010; Do et al., 2016). This is in contrast to the patterns of random monoallelic expression observed in a cell subpopulation, whereby an allele which is cell type-specific (as opposed to inter-individual SNP variation arising from cis-acting polymorphisms) associates with gene expression and/or CpG methylation patterns (da Rocha and Gendrel, 2019). Allele-specific methylation associates with chromosomal imprinting and dosage compensation (through X-inactivation), but has also been observed in non-imprinted autosomal regions (Kerkel et al., 2008; Wang, Lou and Wang, 2019). In fact, allele-specific methylation is observed for as high as 10% of common SNPs across the genome (Hellman and Chess, 2010). One model for the mechanism underpinning sequence-dependent allele-specific methylation is that proteins which have sequence polymorphisms in their binding sequences may consequently bind with allelic sensitivity and preference (Shoemaker et al., 2010; Do et al., 2017). Therefore, conservation of allele-specific methylation at repressed chromatin across a population suggests there are indeed important genetic associations with both methylation and expression.

1.11 – Quantitative trait loci and their associations with CpG methylation in normal and tumour tissues

While studies have examined the molecular mechanisms underpinning allele-specific methylation in individuals, association studies have been used to explore this relationship in larger populations. The mapping of genetic loci which associate with a quantitative molecular phenotypes, termed quantitative trait loci (QTL), has become an important approach to analyse how genetics may drive complex traits and to interrogate the genomic function of a locus (Mackay, 2001). For example, a significant advancement in our understanding of the regulatory roles of sequence variants has been through the use of large transcriptomic datasets to identify SNPs which associate with gene expression (eQTLs) in an allele-dependent manner (Morley et al., 2004; Schadt et al., 2008). These eQTLs have been used to develop our understanding of population-specific and tissue-specific patterns of expression, as well as disease-associated differential gene expression (Montgomery and
Dermitzakis, 2011; Nica and Dermitzakis, 2013). QTLs are therefore tools which can relate genetic variance to variance in different molecular phenotypes.

Moreover, by exploiting DNA methylation data resources, studies can observe SNPs which have an allele-dependent association with methylation, at one or more CpG sites, and can therefore describe methylation QTLs (methQTLs) (Heijmans et al., 2007; Bell et al., 2011; Villicaña and Bell, 2021). MethQTLs have become valuable tools to identify SNPs which may have a functional role in the genetic regulation of DNA methylation patterns at an increasingly large number of loci (Hop et al., 2020). However, like all QTLs, methQTLs are limited in how much can be inferred beyond an association; discerning the causal direction of the association and the mechanisms responsible for the association requires more comprehensive analysis (Pierce et al., 2018; Huan et al., 2019; Neumeyer, Hemani and Zeggini, 2020). For this reason, methQTLs require additional functional characterisation using the genomic context of the SNP or corresponding CpG, and their colocalisation with QTLs for other molecular phenotype data, such as expression QTLs, proteomic QTLs, histone modifications, DNase I accessibility and transcription factor binding (Banovich et al., 2014).

One of the most defining features of the mechanism driving a methQTL is the distance between the SNP and the corresponding CpG site (Villicaña and Bell, 2021). A methQTL which associates with local CpG site methylation levels is termed a cis-methQTL (Figure 4). Alternatively, long-range SNP-CpG associations are termed trans-methQTLs (Figure 5). Whilst the threshold for a local cis-association has been debated (typically a window between 500Mb and 2Mb in cis-methQTL studies), Fauman and Hyde, (2022) showed that a cis-QTL distance of 1Mb is optimal in several trait-associated loci (including expression, proteins and metabolites) by modelling with the assumption that cis associations should be distance-dependent whilst trans associations should have a normal distribution of distances (McRae et al., 2018; Morrow et al., 2018). While both fundamentally alter how transcription factors interact with DNA, the underlying mechanisms behind cis-methQTLs and trans-methQTLs are thought to be different, and thus are beneficial to study and understand as separate models (Villicaña and Bell, 2021).
Quantitative trait loci (QTLs) represent allele-dependent associations between genotype and molecular phenotypes, such as gene expression (A) or CpG methylation (B). C) Allele-dependent associations between SNPs and local CpG sites (less than 500kb distance) are termed cis-methQTLs. The proposed cis-methQTL mechanism is that transcription factors which remodel local methylation (with the assistance of co-factors) may only bind unmethylated binding regions, but a SNP in this binding region may disrupt binding either through methylation sensitivity or sterically if it attracts methyl CpG binding proteins which occupy the binding region. D) If a QTL associates with a local CpG that is also in the regulatory element for a gene, then it may also associate with gene expression and thus be a cis-acting expression-methylation QTL, a common occurrence for cis-methQTLs.
**trans-methQTL mechanism:**

![Diagram of trans-methQTL mechanism](image)

**Figure 5 – Mechanism for trans-methQTLs**

Allele-dependent associations between SNPs and distal CpG sites (more than 500kb distance or interchromosomal) are termed *trans*-methQTLs. The most prominent hypothesised *trans*-methQTL mechanism is that they are primarily cis-eQTLs in relatively inactive regions of the genome for genes which transcribe transcription factors capable of modifying methylation and enabling the long-range remodelling of methylation at distal CpG sites. These transcription factor could be methylase enzymes (such as DNMTs) or may be epigenetic regulators of genome wide methylation changes (such as CDCA7).
Thus far, there have been more studies successfully detecting \textit{cis} associations between genetic variation and CpG methylation than detecting \textit{trans} associations, for two key reasons. Firstly, the microarray technology used for measuring CpG methylation has an inherent genic bias in the genomic distribution of CpGs which are measured, as they were originally designed to focus on genes, which favours the detection of \textit{cis}-methQTL-CpGs which are enriched in gene promoters (Pidsley \textit{et al.}, 2016). Secondly, the effect size of distal associations is considered to be inherently weaker compared to the direct local effects of \textit{cis} associations, and thus, the greater sample sizes required to detect \textit{trans}-methQTLs have only recently become readily available in a small number of tissues (Hannon, Gorrie-Stone, \textit{et al.}, 2018). However, future studies will take advantage of the increasingly improved non-genic genomic coverage and larger sample sizes for methylation data cohorts to address these limitations.

\textbf{1.12 – Comparison of methylation changes in functional genetic elements between tumour and normal methylome}

A key part of characterising possible methQTL mechanisms is to first understand how the methQTL-associated SNPs and CpGs are distributed across the genome respectively, and which genetic features these methylation changes are occurring in. Specifically, enrichment or depletion of methylation changes in certain functional genetic elements, such as CpG islands, promoters, enhancers or transcription factor binding sites, may all suggest possible underlying methQTL mechanisms. To do this, these elements must be considered individually by the functional relevance of their methylation states in normal and tumour tissues.

Studies which have observed differential methylation between tumour and normal tissues at CGIs in regulatory elements (such as promoters and enhancers) and in the surrounding regions (which are called shores and shelves) have investigated this dysregulation as a key step in tumour progression (Ili \textit{et al.}, 2020; Muse \textit{et al.}, 2020). Promoters and enhancers share similar chromatin features to one another, but promoters
are proximal to the transcription start site while enhancers are distal (located in gene bodies or intergenic regions) (Core et al., 2014). Indeed the distinction of enhancers from promoters is primarily from lower transcript stability and the presence of H3K4me1 histone mark (Core et al., 2014).

For example, whilst promoters are typically hypomethylated in normal tissue, the tumour methylome shows instances whereby promoter CGIs have become hypermethylated, and therefore silenced. This occurs primarily at bivalent promoters for developmental genes (found in regions of low methylation called DNA methylation valleys in facultative heterochromatin) and more rarely in tumour-suppressor genes (Sproul and Meehan, 2013; Nishiyama and Nakanishi, 2021). Furthermore, tumour methylation in enhancers has been shown to be highly variable across multiple tumour types relative to others genomic features, such as promoters or overall CpG islands (Bell et al., 2016). Indeed, the majority of enhancers are differentially hypomethylated compared to normal tissue and likely have a role in chromatin activation during cancer progression (Bell et al., 2016).

Another important feature inferring these mechanisms is where methylation changes are in the binding sequences of transcription factors, especially if those proteins have a potential role in broader methylation regulation. A good example is the CTCF transcription factor, which binds to regularly interspaced binding regions to mediate 3D chromatin structure (separating in to domains of euchromatin and heterochromatin) (Dehingia et al., 2022). CTCF also acts as an insulator to regulate long-range transcriptional interactions between genes and distal enhancers (Dehingia et al., 2022). Occupancy of the CTCF protein has been shown to maintain local hypomethylation at binding sequences and to actively protect against hypermethylation (Damaschke et al., 2020; Segueni and Noordermeer, 2022). This is thought to be through a interference with DNMT methylation proteins through either steric hinderance or through PARP-1 signalling, which moderates the insulator functionality of CTCF and recruits CTCF to the DNA damage response (Guastafierro et al., 2008; Han et al., 2017). In tumours, there is frequent CTCF copy-number loss and consequently hypermethylation at many of these binding sites through the genome, a few of which were shown to reside in genes which were consistently downregulated across both breast and prostate cancer (Damaschke et al., 2020). However,
this effect shows tumour-specific variance in where methylation is most affected; there is only a 2% overlap in CpGs affected by CTCF-loss hypermethylation between prostate and breast cancer (Damaschke et al., 2020).

### 1.13 – Mechanisms of *cis*-acting methylation QTLs

To date, *cis*-methQTL and *trans*-methQTL profiles have been well studied in normal blood (Min et al., 2021), but less so in other tissues due to the limited data resources for this level of description in non-normal tissues. Building on the observations of other studies, which have defined the genomic architecture of *cis* - and *trans*-methQTLs, a recent meta-analysis by a DNA methylation consortium aimed to consolidate the results from various methQTL analyses which were carried out in normal blood (Min et al., 2021). This allowed for a sufficiently powered characterisation of the “functional context” of both *cis*-methQTLs and *trans*-methQTLs in normal tissue (Figure 6), a term which I will use to represent aspects of the genomic positioning and their enrichments in functional regions, of the SNPs and CpGs respectively in a methQTL association. The functional context of a methQTL may therefore be used to imply potential mechanisms driving a SNP-CpG association.

In normal tissue, *cis*-methQTLs-SNPs have been shown to be enriched in active regions of the genome, such as promoters and enhancers (Shi et al., 2014; Min et al., 2021). Meanwhile, *cis*-methQTL associated CpGs were enriched in non-genic regions and enhancers but depleted in CGIs and most promoters (Villicaña and Bell, 2021). This *cis*-methQTL-CpG depletion in CGIs is thought to be a consequence of high evolutionary conservation of CGI hypomethylation, which would reduce DNA methylation variation in these CGIs and also reduce allele-dependent associations (Long et al., 2013; Al Adhami et al., 2022). Indeed, *cis*-methQTLs could hypothetically affect gene activity if the *cis*-methQTL-CpG resides in an enhancer/promoter and thus regulates gene expression.

Indeed, the strongest mechanistic hypothesis has been that *cis*-methQTL-SNPs disrupt the binding regions of transcription factors which reprogramme methylation at CpG sites in local regulatory elements (Figure 4C-D) (Banovich et al., 2014; Wu et al., 2018). As such, *cis*-methQTLs commonly impact multiple CpGs within a 3kb window.
Figure 6 – Proportionally enriched/depleted distribution of methQTL-CpGs in functional genomic features

MethQTLs have characteristic genomic profiles in normal blood for the proportional distribution of QTL SNPs, and their corresponding CpG sites, which are dependent on the methQTL acting in cis (A) or in trans (B). A) For cis-methQTLs, SNP enrichment in active regulatory regions and corresponding CpG enrichment in local enhancers fits the cis-methQTL mechanism of disrupting binding of transcription factors which modify local CpG methylation. B) For trans-methQTLs, SNP enrichment in inactive regions (such as intergenic and heterochromatin) and corresponding CpG enrichment in distal gene promoters (or enhancers) fits the trans-methQTL mechanism of being a cis-methQTL for a transcription factor with long-range methylation remodelling activity.
These CpGs may often be in the promoter/enhancers of a gene and thus mediate expression, as studies have shown that cis-eQTLs are often also cis-methQTLs (Pierce et al., 2018; Zhao et al., 2019). As the power of methQTL analyses has improved, studies have also shown that the effect size and detection rate of cis-methQTLs is larger than that of trans-methQTLs, an observation that would also align with the idea of a relatively direct cis-methQTL mechanism (Hannon, Gorrie-Stone, et al., 2018).

This change in methylation may be an active mechanism, where these changes are also within binding sites for methylation-sensitive binding factors, such as DNMT3A or the TET proteins which may drive these local changes (Wang et al., 2019). Alternatively, methylation has been suggested to change passively if transcription factors with protective roles against methylation changes become compromised, such as I have described for CTCF binding (Stadler et al., 2011; Damaschke et al., 2020). Through the mapping of methQTLs and ASM, this relationship between sequence-specific transcription factor binding and the protection of local hypermethylation has been suggested as a key cis-acting mechanism for methylation changes (Do et al., 2017). An alternative mechanism may be that the CpG density of cis-methQTL-CpGs may affect methylation changes, as age-associated methylation loss has been shown to be greater in low-CpG density regions due to the collaborativity of CpGs to maintain methylation levels in CpG dense regions (Higham et al., 2022).

### 1.14 – Mechanisms of trans-acting methylation QTLs

In contrast, trans-methQTL-SNPs in normal blood have been shown to be enriched in repressed, heterochromatic regions (Figure 5) (Min et al., 2021). Meanwhile, trans-methQTL-CpGs are enriched in CGIs and promoters, but depleted in enhancers, gene bodies and inactive heterochromatic regions (Villicaña and Bell, 2021). Additionally, more than 25% of trans-methQTLs-SNPs have been proposed to be located in a telomeric (17%) or within a 1Mb subtelomeric (7%) regions in whole blood (McRae et al., 2018). Trans-methQTLs have a less established canonical mechanism associated with its genomic profile since these loci drive methylation changes distally and often indirectly.
The strongest and simplest model for a trans-methQTL mechanism is whereby the QTL is primarily a cis-eQTL for a transcription factor which then acts to reprogram methylation at a distal CpG site (Pai, Pritchard and Gilad, 2015; Bonder et al., 2017; Huan et al., 2019). Indeed, this mediation of transcriptional regulation by a cis-eQTL may also occur by a cis-methQTL if the methylation state of a cis-CpG also affects expression of a protein, as has been shown to be the case for a GTPase enzyme that can regulate DNA methylation (Shi et al., 2014). This may also occur with the transcription of proteins which associate with broad genome-wide epigenetic regulatory effects, such as CDCA7 which correlates with methylation changes across the genome (Hop et al., 2020). An individual trans-methQTL may also have multiple trans-CpGs if the transcription factor is a broader epigenetic regulator. For example, variants in the intron of SENP7, which remove SUMO modifications from proteins, have been shown to be cis-eQTLs for a transcript that reduces methylation at various trans-CpG sites (Lemire et al., 2015).

More complex models might add that the SNP/CpG association may occur from more indirect trans-methQTL effects. This has been suggested from the observation that there is lower trans heritability explained by trans-methQTLs relative to cis-methQTLs (Gaunt et al., 2016). Mechanistically, this could arise if methylation changes were dependent on a polygenic complex of proteins (such as co-factors alongside methylation enzymes) and the heritability attributed to each of these multiple genes would therefore be individually lower. This suggestion may be supported by the fact that cis-methQTLs have typically had a higher detection rate than trans-methQTLs until recently where larger cohort sizes have had sufficient power to identify and study trans-methQTLs (Hannon, Gorrie-Stone, et al., 2018). Alternatively, one explanation for unexplained mechanisms for trans associations could be that long-range, 3D genome organisation may position distal sequences in proximity (for example, through looping), although there has been little evidence for this thus far; in this instance, local mechanisms of methylation change, typically associated with cis-methQTLs, could also be the driving mechanisms behind trans-methQTLs (Do et al., 2017). One final suggestion has been that a coding mutation in a gene could become a trans-methQTL if that protein has a disrupted methylation binding domain and consequently recruits methylation proteins to aberrant regions of the genome. This loss of selectivity for DNA binding has been shown for the MeCP2 gene, and also in other
transcription factors without methylation binding domains (Zhu, Wang and Qian, 2016; Franklin, 2019). Trans associations therefore have a variety of mechanisms which drive these long range methQTLs, but our understanding of this will continue to steadily improve with the availability of larger datasets and sufficient power to detect and describe these mechanisms more accurately.

1.15 – Methylation QTL studies and the challenges of analysing tissue-specific and tumour-specific methylation patterns

While DNA methylation patterns do show tissue-specific patterns, there have been differing reports of the proportion of methQTLs which are either shared or unique to a tissue. For example, Oliva et al. (2022) looked across 9 tissue types and found that more than 5% of cis-acting methQTLs were tissue-specific while another study of methQTLs across 4 tissue types (B-cells, T-cells, biopsies from ileum and rectum) found tissue-specific methQTL proportions ranged between 6.5% and 10.1% (Scherer et al., 2021). Meanwhile, one study by Gutierrez-Arcelus et al. (2015) assessed the number of methylated sites from methQTLs across of three cell types (lymphoblastoid, T-cells and fibroblasts) and found that most (between 34% and 52%) of the methylated CpG sites they identified were shared between tissues.

While methQTL studies are abundant, there remains some key challenges which have limited the practicality of analysing different tissues, and particularly in cancer where heterogeneity of tumour cell composition makes methylation patterns complex to study (Beyes, Bediaga and Zippo, 2021). For example, an effective analysis of cell-type specific methQTLs would require sufficiently large cohort sizes which is especially limiting in methylation studies; studies profiling methylation with single-cell resolution have been expensive, have had poor coverage and suffered from relatively small sample sizes, and thus new approaches are being developed to address these limitations (Clark et al., 2016; Rahmani et al., 2019). Recently, the Expanding GTex (eGTEx) project has addressed the lack of non-blood methylation data (amongst data for other molecular features, such as histone
modifications and DNA accessibility) by extending the GtEX expression data project with methylation data to provide researchers with a cohort of subjects with matched expression and methylation data (Stranger et al., 2017; Oliva et al., 2022). Overall, there is still some practical challenges to overcome in this effort to better understand tissue-specific methylation patterns.

### 1.16 – Methylation QTL studies in cancer

Perhaps the most significant non-blood tissue that require greater understanding and research are tumour tissues, although studies of these have thus far been sparse. For example, tumour methylation has been complicated to study due to tissue-specific methylation patterns and the innate heterogeneity of tumour cell content (Beyes, Bediaga and Zippo, 2021). Another reason is the distinction of understanding the different implications of methQTLs from germline genotype variation and somatic (tumour-associated) genomic variation in methQTL detection. An accumulation of common somatic mutations primarily drives tumour progression, but there is also the significant role of germline variants with high functional-relevance which may predispose the genome to higher cancer risk and susceptibility (Ramroop, Gerber and Toland, 2019; Qing et al., 2020). Indeed, while some of these germline variants are identified in GWAS, many have low effect sizes but may affect molecular processes which are key in the process of somatic mutation (e.g. DNA repair pathway) (Vali-Pour et al., 2022).

There are various examples of recent tumour studies which have developed our understanding of the tumour methylome using methQTLs. One study compared cell type-specific melanoma methQTLs and tissue-of-origin melanocyte methQTLs, which colocalised with eQTLs and GWAS hits, and showed that almost half of the melanocyte methQTLs were conserved in melanoma (Zhang et al., 2021). Another study also describes tumour-specific methQTLs (whereby the methylation effect is exclusive to the tumour) in prostate cancer which colocalised with eQTLs and subsequently found novel biomarkers for disease aggression (Houlahan et al., 2019). Using over 3,000 primary tumour samples from 13
cancers, one study identified and catalogued pan-cancer risk-associated methQTLs which contained 21% of all cancer variants tested and which harboured novel cancer-associated genes (Heyn et al., 2014). Indeed, this idea has been further adapted in to a methQTL database for 23 different human cancers from The Cancer Genome Atlas data, some of which are associated with GWAS loci or survival time (Gong et al., 2019). However, while most of these studies have focussed on identifying methQTLs as clinical tools (for risk assessment, prognosis or survival analysis), none have thus far described the molecular mechanisms that underpin methQTLs in tumours tissues or how these mechanisms change between cis-acting and trans-acting associations.

1.17 – Breast cancer methylation mechanisms

There are studies focussing on the mechanisms of breast cancer epigenetics. One such study performed an expression-methylation-QTL analysis in breast tumours and found that DNA methylation at enhancers and at binding regions for three transcription factors responsible for oestrogen-receptor signalling (ERα, FOXA1 and GATA3) were sufficient to delineate molecular subtypes (Fleischer et al., 2017). Methylation changes have been shown to occur relatively early in breast cancer, in ductal carcinoma rather than invasive tumour cells (Vietri et al., 2021). Another study focussed on identifying differentially methylated genic CpGs with prognostic potential and identified 368 CpGs, the majority of which (80%) were hypermethylated, downregulated promoters and the remainder were hypomethylated, upregulated gene bodies (de Almeida et al., 2019). Additionally, one study merged genotype, transcriptome and methylome data to normalise effects from the tumour microenvironment to ask more accurately how these three layers interact together in breast cancer progression (Batra et al., 2021). They found some consistent trends, including some CpGs which underwent replication-dependent hypomethylation, but also found CGI instability, mainly in hypermethylated, CpG-dense enhancers (and some promoters) and less so in hypomethylated, CpG-sparse enhancers, where the degree of methylation change also positively correlated with more aggressive tumour grade (Batra et al., 2021). However, the
underlying mechanisms of breast cancer methylation dysregulation still require further characterisation.

Beyond the CpGs themselves, there are also strong functional links between methylation changes and the transcription factor genes these CpGs reside in (Vietri et al., 2021). Tumour suppressor genes undergo early hypermethylation in breast cancer (Brooks, Cairns and Zeleniuch-Jacquotte, 2009). For example, the cell-cell adhesion pathway transcription factor E-cadherin associates with aggressive breast cancer when silenced (Sharh et al., 2014). Meanwhile, hypermethylation and silencing of IL15RA alters regulation of other cell-adhesion genes involved in cell-adhesion, thus driving breast cancer progression (H. Yang et al., 2019). This functional link is not limited to hypermethylation however, as hypomethylation of the circadian regulatory gene CRY2 in ER- breast tumours also drive tumour progression (Mao et al., 2015). Specifically, triple-negative breast cancers show the greatest epigenetic alterations, with BRCA1 silencing and TET1 overexpression amongst many other hypomethylated genes (Zhu et al., 2015; Good et al., 2018). Thus, there is a need to take population approaches which have been used to profile methylation mechanism in normal tissue, and use this to verify the methylation changes we might expect to see in breast tumours given widespread methylation dysregulation.

1.18 – Project Aims

Population approaches have been used to characterise allele-dependent genetic associations with methylation patterns in normal tissues. However, using these approaches to characterise allele-dependent methylation changes in tumour tissues is challenging due to their complexity, such as how these associations are confounded by the heterogeneous cell composition of tumour tissue. The complication of tumour analysis necessitates a bespoke, effective and practical analytical framework. Thus far, many different pipelines have been described and applied, but few include the necessary initial evaluation of the steps and methods within these pipelines (such as pre-processing steps and quality control),
to ensure an analytical framework is optimised for identifying methylation QTLs in a specific dataset.

Thus far, models have been described which characterise individual methQTLs (primarily on their SNP-CpG distance, as cis and trans acting mechanisms have been hypothesised). However, whilst many studies have focussed on a single methQTL to infer a causal gene, few studies have considered the broader genomic distribution of methQTL SNPs and their corresponding CpGs. As such, a novel approach may be to describe the functional context for an individual methQTL, which considers the genomic positioning of SNPs and their corresponding CpGs as well as any functional elements they reside in. Such analysis could provide novel insights and identify regions related to molecular interactions that may occur between SNPs and CpG methylation (for example, transcription factor binding disruption in cis-methQTLs). Thus, I will use the functional context of methQTLs identified in the cancer methylome to compare against the normal methylome and reveal important differences which expand our understanding of how genetic sequence variation may interact with CpG methylation in tumours.

I will analyse methQTLs in breast cancer, which currently has one of the highest prevalence and has the poorest mortality rate in the context of female cancer cases (Sung et al., 2021), but which also has well-defined transcription factor biology (Zacksenhaus et al., 2017) and sufficiently large resources of genotype and methylation data available for analysis. While expression QTL (eQTL) analyses are common in breast tumours (Quiroz-Zárate et al., 2017), resources and tools for methQTL studies are rarer. I will use a computational approach which allows for the study of methQTLs across a cohort of breast tumour patients. In order to understand the context of any methQTL findings, I will compare breast tumour methQTLs against methQTLs identified in normal tissue to see if they are potentially tumour-specific. While normal breast tissue data would be the ideal comparator, such data is too sparse in literature for sufficiently powered analysis, and so I will compare against normal blood methQTL findings instead. Given the widespread methylation dysregulation observed in breast tumours, I hypothesize that germline genetic sequence polymorphisms may in part drive tumour-specific methylation changes (independently from tumour-associated somatic mutations). I hypothesise that this will manifest through
differences between the normal and tumour breast methylomes, such as how the methQTLs are distributed and their distinct functional context.

First, I will develop a robust per-allele association analysis for identifying methQTLs in breast tumours. As there are myriad methods available for both data pre-processing and for the methQTL analyses, I will evaluate these analytic methods in order to optimise the per-allele association analysis to the data I will use. Then, I will use this optimised methQTL association analysis to answer three key analytical questions about breast cancer methylation:

1. Are there methQTLs in breast tumour tissue, and how do these compare to methQTLs identified in normal blood tissue?

2. What is the functional context of cis-acting methQTLs in breast tumours, and how does this differ from that of cis-methQTLs in normal blood?

3. What is the functional context of trans-acting methQTLs in breast tumours, and how does this differ from that of trans-methQTLs in normal blood?

By characterising the functional context of cis and trans acting methQTLs in the abnormal tumour methylome, I aim to expand the current understanding of how genetic variation may drive methylation dysregulation in breast tumours.
Chapter 2 – Methods

2.1 - Datasets

TCGA breast tumour data:

I accessed breast cancer methylation data from The Cancer Genome Project (TCGA) breast cancer project (TCGA-BRCA) using the Genomic Data Commons (GDC) data portal from the National Cancer Institute (Collins and Barker, 2007; Ciriello et al., 2015; Thennavan et al., 2021). TCGA projects have a strict process for histological quality control at a centralised site. For example, tumour samples must have a tumour content greater than 60%, a necrosis percentage lower than 20%, and a matched normal sample (either blood or breast).

I downloaded methylation sample data for invasive ductal carcinoma which had been measured on Illumina’s Infinium 450k microarrays. This contained 704 female cases which had a total of 783 samples which were from primary breast tumour sites, matched normal blood, matched tumour-adjacent normal breast and metastatic samples.

For access to TCGA genotype data, I had to comply with standards for the ethical and protected use of this data, as set by the NIH and their policy for best practice. I signed and agreed to an NIH data protection plan, alongside the NIH data use certification agreement for the TCGA project. In short, these detailed best practices to ensure the confidential data was stored and protected on a secure university data server, and not copied or moved from here. It also ensures the individuals will remain anonymous and that any data breach will be dealt with promptly and appropriately. To this end, the TCGA genotype data was stored in an individual directory with individual security and permissions requirements.
I downloaded genotype sample data for matched normal blood samples and/or matched tumour-adjacent normal breast tissue which was measured using the Affymetrix 6 SNP microarray. This contained 1,109 female cases with 2,265 samples which were from primary tumour, matched normal blood and normal tumour-adjacent breast tissue. Matched normal blood represents the closest approximation available of a normal tissue for comparison. However, somatic variation between different tissues means there remains a minor limitation in how accurately comparisons can be made between these tissues (Acha-Sagredo, Ganguli and Ciccarelli, 2022).

**GTEx expression data:**

The Genotype Tissue Expression (GTEx) project is a repository for expression and genotype data (primarily from arrays, whole genome sequencing, whole exome sequencing and RNA-sequencing) from various tissues. GTEx contains a repository of reported eQTLs (ranging from eQTLs in genes/SNPs of interest to eQTLs in single-cell data) that may be used to study tissue-specific gene expression and regulation (Carithers et al., 2015). GTEx (Version 8) has a standardised analysis method which includes recording RNA-seq data, pre-processing, expression quantification and eQTL mapping using FastQTL (GTEx Portal). The breast mammary tissue eQTLs were identified from pre-processed bulk tissue RNA-seq data for 396 samples with matched expression and genotype data. This contained 1,441,105 eQTLs which had been determined to be significant (FDR < 0.05).

**2.2 – R environment for pre-processing, analysis and data visualisation**

All pre-processing and analyses were performed using base packages in R (version 4.1.0), unless specified otherwise. All required R packages were loaded using Bioconductor (version 1.30.19).
All plots (including enrichment analyses, histograms, bar charts, Manhattan plots, QQ-plots, violin plots, PCA scatterplots) were made using the R package ggplot2 (version 3.4.1, Wickham, 2009), unless specified otherwise.

2.3 - Methylation data processing and evaluation

Normalisation:

TCGA methylation data and annotations were loaded into the methylation pre-processing R package, minfi (version 1.40) (Aryee et al., 2014). A sample information sheet was generated from metadata files taken from the repository to access hybridisation sample features (eg. array IDs, sample type, race, sex, TCGA barcode). After loading the data in to minfi, I used the ssNoob function for normalising the methylated and unmethylated probe signals. This method generated tables with the normalised methylation values (as beta values) and with the detection rates (p-values for detected probes) for each probe and sample.

Quality Control:

I performed quality control filtering using the detection p-value table to calculate detection rates per probe and removed probes with detection rates greater than 1%. I also removed probes which were in sex chromosomes or which were non-CpGs. Probes were removed if they were in a masking probe annotation from Zhou et al. (2017), which highlights probes with either feint signals from poor hybridisation or spurious signals from cross-reactivity. Sample-level filtering was also used for quality control. Samples were included which were annotated as female and from primary breast tumour tissue (from annotation file), and which were European by principal component clustering (see section 2.5). In addition, samples were filtered out if they had a median methylated:unmethylated probe signal ratio greater than 10.5 (which is a minfi quality control command) or if they had a detection rate greater than 1%. Finally, the beta values were transformed in to M-
values, which are considered more statistically robust and improves how well extreme methylation values (near 0 and 1) can be analysed. I performed this transformation in R using the formula $M = \log_2(B/1-B)$.

**Bimodality assessment:**

For assessing the bimodality of the CpG beta value distributions, I used the R package `diptest` (version 0.76-0) (Ringach, 2021). I performed a “dip test” on the beta value distributions, which tests the null hypothesis that a distribution is unimodal (Hartigan and Hartigan, 1985). This produced a dip statistic which can then be assessed and contrasted against other bimodal distributions to directly compare bimodality.

2.4 - SNP genotype data processing and evaluation

**Normalisation:**

TCGA genotype data was loaded in to the R package `CRLMM` (Corrected Robust Linear Modelling with Maximum likelihood distance; version 1.40) to be processed (Carvalho et al., 2007). The data was genotyped and normalised using `CRLMM`, which was designed to address the shortcomings of between-lab studies resulting from batch effects when processing and calling single-nucleotide polymorphisms in large SNP microarray data (Carvalho et al., 2007). CRLMM uses quantile normalisation when pre-processing array data (Carvalho et al., 2007). Following this, a SNP call table is generated, along with a probability score for each SNP call, which can be analysed alongside a SNP annotation file (which is also accessed from the GDC portal).

**Quality Control:**

SNP probes were removed if they were non-autosomal and if they failed two quality scores from `CRLMM`; probes were excluded if their SNP QC score was less than 0.25 and if
the average SNP probability score for any probe was less than 0.95. Probes were also filtered out by the genomic profile of the SNP; probes were excluded if their genotype missingness rate across all samples exceeded 1%, their Hardy-Weinberg equilibrium significance p-value was below 1x10^{-6} or if their minor allele frequency was lower than 5%. Additionally, in preparation for imputation, SNP probes were also removed if they had any allele mismatches. Meanwhile, sample filtering steps only included samples which were female and from primary blood tissue (from annotation file), and which were European following principal component clustering (Section 2.5). Samples were also excluded if they were duplicates, their genotype missingness rate across all probes was higher than 3% or their signal to noise ratio was less than 4.

**PLINK file conversion/formatting:**

The SNP call table and SNP annotation files are required to be in the format of PED and MAP files to be compatible with *Plink* (version 1.90b1g). Unable to find a suitable tool for generating PED and MAP files from a SNP call file, I used R to load the SNP call file, SNP annotation file and the sample sheet in to R and then manually constructed PED and MAP files to be compatible with Plink. I then used *Plink* to convert PED and MAP files to be in the *Plink* binary file formats (BED, BIM and FAM). Finally, the VCF file format is a prerequisite for imputation. Thus, I used *Plink* to recode the B files in to VCF files. Finally, I used *vcftools* (version 0.1.13, Danecek et al., 2011) to then obtain a genotype matrix (coded as 0,1 and 2) text file from the B-files, that was more readily accessed and analysed compared with the VCF file format.

**2.5 – SNP genotype imputation**

I used the Michigan Imputation server (MIS) to impute the processed genotype data, using their Minimac4 genome imputation tool (Das et al., 2016). Using the MIS interface, I selected the haplotype reference consortium reference panel (HRC version r1.1 in human genome build hg19), which included ~64,000 haplotypes from samples which were
predominantly of European ancestry. I also selected the European reference population, the
hg19 genome array build and I indicated that this genotype data was unphased. I also used
the MIS interface for an initial round of quality-control correlation-based filtering (using R²)
to filter out imputed genotypes if they had a haplotype R² correlation less than 0.3. I
uploaded the VCF files to the server to be imputed.

I also did further post-imputation filtering using some Linux-based tools for
additional quality control by filtering the VCF files using bcftools (version 1.16, Danecek et
al., 2021). I applied a minimum R² threshold of 0.4 and removed any SNPs which had a rare
minor allele (less than 0.05 or greater than 0.95). I also removed imputed SNPs which were
monomorphic or which had a missingness rate across samples greater than 3%. I then
converted VCF files to a BGEN format (using qctool version 2.0.8) and obtained a genotype
matrix from the VCF files (using vcftools version 0.1.13), which were both required for the
methQTL association analysis.

2.6 - Analysis of population structure in SNP genotype data

Principal component analysis:

To address population structure, I calculated principal components from the
unimputed genotype data. I used Plink (version 2.0, Purcell et al., 2007) to calculate the top
20 principal components for the genotype data. I then used R to plot the first two
components against one another and coloured principal component clusters by the ancestry
annotation assigned from the sample metadata which was also acquired from the GDC data
portal.

Ancestry outlier analysis:

I adapted a custom R tool developed around a basic principal component analysis
which additionally highlighted outlier samples which genetically deviate from the majority
ancestry in the cohort (written by David Clark & Paul Timmers). This was originally designed to remove samples which introduced any genetic diversity in contrast with the principal ancestry of a cohort (for example, identifying ethnic outliers when studying island cohorts). These outliers are identified where they exceed a set distance calculation from the ancestry cluster. I used this tool to identify samples which did not genetically cluster with the principal European ancestry in the cohort. This tool also plots genotype clusters for reference ancestries to compare against the subject data.

I downloaded reference genotype data from *HapMap* phase 3 for the two other major populations which were present in the cohort and metadata: African American ancestry and Asian ancestry (Altshuler *et al.*, 2010); I used ASW (African ancestry in southwest USA) and CHB (Han Chinese in Beijing, China) to broadly represent African American and Asian ancestries respectively. I used these reference population data to anchor a plot of the first two principal components in comparison with the primarily European ancestry of the TCGA genotype data. The tool works by first merging *Plink* files (BED, BIM and FAM files) for both the cohort and reference genotype data according to chromosome. The code then selects for only the reference populations of interest you have specified from the reference files. The code then calculates principal components from these merged files and plots them by population, identifying outliers using a tolerance threshold relative to the centroid of the principal ancestry cluster.

### 2.7 – Methylation QTL association analysis

**Covariate data:**

I included covariate data for age but not for sex as the samples were filtered to be all female. I also included twenty principal components (the maximum number of components from the analysis) which were calculated for unimputed genotype data to account for population-driven substructure in the data (Section 2.5).
MethQTL identification with GWAS-based association analysis:

I carried out a methQTL association analysis in Linux. I used a per-allele association analysis that was designed as a method for genome-wide association studies (GWAS) between genotype and molecular phenotypes of interest. This association analysis was developed around the linear regression tool RegScan (Haller et al., 2015) and by also adjusting for covariate data. The tool uses two steps of data preparation; the first step applies quality checks on the phenotype data after merging them together and the second step fits the covariates to generate residuals which will be used in the regression. The phenotypic residuals are then regressed against the genotypes using RegScan, one chromosome at a time. The chromosome regression results are then merged in to one collection of summary statistics.

While this GWAS pipeline was established to find associations with a single phenotype, the significantly large number of methylation phenotypes (398,230 CpGs) required that I use the task array functionality of the Linux cluster job schedulers to analyse each of the CpGs as individual phenotypes simultaneously. I also included additional RegScan options, including a maximum p-value limit of 5x10^{-8} (by converting to a t-statistic and setting a corresponding t-limit of 5.579959). Samples were also filtered by Z-scores, which is the number of standard deviations that the methylation values diverge from the mean. I applied two rounds of Z-sample filtering based on separate thresholds; the first round was applied to raw methylation values and had a maximum Z-score cut off of 5 while the second round applied to the residual methylation values and had a maximum Z-score cut-off of 3. After successfully running the association analysis, I then used a combination of Linux and R to merge the methQTL output files in to a single output file of methQTLs and then filter for only associations which passed the multiple-testing Bonferroni adjusted maximum p-value threshold of 5x10^{-14}.

Assessment of association signal quality:

I visually assessed different aspects of the association signals identified from the methQTL analysis. I generated Manhattan plots by plotting the chromosomal position
against the significance p-value of each association signal. I included some reference P-value significance markers for scale of significance; this included nominal significance (less than 5x10^{-5}) and genome-wide significance (less than 5x10^{-8}). I also generated Quantile-Quantile (QQ) plots to show if the p-value statistics from these associations followed a uniform distribution. In a standard GWAS study, a QQ-plot of these statistics should only show a small tail of p-values departing from the uniform distribution, representing a peak of significant SNP hits. A QQ-plot showing a large divergence from the uniform distribution could suggest genomic inflation has occurred. As such, I additionally calculated a genomic inflation factor to assess how the observed p-value significance compared with the expected significance. This is defined as the ratio of the median observed chi-squared test statistic of methQTL p-values against the median expected test statistics of a chi-squared normal distribution. This value is around 1 if there is no overinflation or underinflation of significance (Devlin and Roeder, 1999).

2.8 – LD clumping

In order to identify a lead SNP which will represent a haploblock of genetically linked SNPs, I used LD clumping. I used Plink (version 1.90b1g) to perform LD clumping of the imputed SNP genotype data. I also used an association file which contained summary statistic annotations for each methQTL (such as P-value and beta for each association), as well as co-ordinates for the SNP and CpG respectively. To decide on the optimal LD clumping conditions, I referred to the clumping conditions described in the polygenic risk tool, LDPred2 (Privé, Arbel and Vilhjálmsdóttir, 2020). Here, Privé et al. (2019) describe a relationship between the window size and the R^2 threshold, whereby the window size is equivalent to a base window size (set as 500) divided by the R^2 threshold. After some initial testing on variations in the window size and the R^2 threshold, I used this formula to calculate the optimal clump conditions for identifying the maximum number of significant methQTLs within an acceptable window size range proposed in Privé et al. (which ranged between 100 and 5000). Ultimately, clumps spanned a window size of 2MB from the lead SNP and contained secondary SNPs which had both a R^2 correlation with the lead SNP greater than
0.25 and SNP associations with p-value significance of less than 5\times 10^{-14}. This generated tables with lead SNPs for each clump, and which indexed all SNPS which had been assigned that clump for any given phenotype.

2.9 – Downstream analysis of methQTLs

Effect size:

To assess the effect size of methQTLs, I analysed the minor allele beta values from each of the methQTLs to approximate the effect size that each methQTL-SNP had on the methylation of a corresponding CpG. I considered both the directional beta value distribution and the absolute beta value distribution in histograms. To compare between cis- and trans-methQTLs I then also considered a violin plot to show the summary statistics of these two distributions and compare the mean beta values. For statistical significance, I used a Wilcoxon-rank test to determine that the distributions are significantly different from one another.

QTL overlap (clumping index):

I looked for overlapping SNPs between two QTL datasets. Since different datasets, which have been clumped separately, may have different lead SNPs representing a common QTL, I assigned each SNP on the array a clump index number relative to the clumping I had performed for the imputed TCGA genotype data. I then looked for common clump IDs between these datasets, rather than for common SNPs between two QTL datasets.

LD Proxy:

Linkage disequilibrium between candidate SNPs was interrogated using the informatics web tool LDLink, which offers a suite of approaches for analysing genetic linkage (Machiela and Chanock, 2015a). I used the tool LDProxy, which allows for the visualisation of the linkage block which a SNP of interest resides in, and offers additional functional
annotation on these SNPs. I entered a *rs.id* for a SNP of interest, along with selecting the appropriate genome build (hg38) and a European-based reference population (British in England and Scotland, or GBR). I also selected a window size of 500kb and a LD measure of \( R^2 \). Finally, I selected which functional annotation resource to use, and chose *RegulomeDB* as a source. This generates a \( R^2 \) plot for the designated SNP and any other SNPs in the surrounding haploblock, a functional annotation (including coding status and potential to have a regulatory role) and an additional table format for these SNPs which can be analysed.

2.10 – Annotations for enrichment analyses

**Illumina Infinium 450k (UCSC/ENCODE) Annotations:**

Infinium arrays use annotations of the genomic distributions for their CpG probes in various functional elements. These are primarily taken from two sources: The ENCODE project and the University/College of Santa Cruz (UCSC) genome browser. The ENCODE project (*encyclopedia of DNA elements*) can integrate genomic and transcriptomic annotations from experimental data, and also includes epigenetic and chromatin-related data-derived annotations as well (ENCODE Project Consortium, 2012). Meanwhile, the UCSC genome browser was built around the human genome project, and consequently is a tool for browsing annotations overlaid with the updated versions of the human genome (Rosenbloom *et al.*, 2013). CGIs had been predicted by the UCSC, which classed them as 500 base windows with GC content greater than 50% and a minimum observed expected CpG ratio of 0.6. Promoters and enhancers were predicted from ENCODE, which makes these predictions based primarily from DNase-seq, bisulfite sequencing, ChIP-seq and gene expression data to indicate regulatory elements (Pazin, 2015). I downloaded Infinium 450k annotations from the Illumina product files for the array (https://emea.support.illumina.com/downloads/infinium_humanmethylation450_product_files.html).

**ChromHMM genomic annotations:**
I used a *chromHMM* annotations from Taberlay *et al.* (2014) for genomic features from two cell lines: HMEC (normal breast tissue) and MCF7 (breast tumour tissue). These were derived from ChIP-seq data of histone marks or proteins which enabled genomic features to be identified. The histone marks used to determine genomic annotations were H3K4me3, H3K27ac, H3K4me1, H3K27me3. Furthermore, CTCF and RNA polymerase II was also included. The genomic features, identified using chromatin profiles of the combinations of these marks, included; bivalent promoters, promoters, enhancers, CTCF sites, repressed (resembling facultative heterochromatin), transcribed regions and heterochromatin (resembling constitutive heterochromatin). I downloaded chromHMM annotations from the ENCODE website (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57498) as BED files for both HMEC and MCF7 cell lines and converted them from hg19 to hg38 using *LiftOver*. The exact chromatin profile for each annotation is summarised in an adapted figure from Taberlay *et al.* (2014) (Table 1).

![ChromHMM genomic profiles](image)

**Table 1 - ChromHMM chromatin-derived genomic feature annotations**

Chromatin profile of ChIP-seq data for histone marks and proteins which form a chromatin profile for genomic features which can be used to classify genomic annotations from chromatin data. Shown here for two cell lines: MCF7 (breast tumour) and HMEC (normal mammary epithelial).
Common tumour PMD/HMD annotations:

Zhou et al. (2018) used low methylation levels and low CpG density to classify a shared set of partially methylated domains (PMDs). These PMDs were common across at least 21/30 tumour samples (N = 417; genomic coverage = 13%) and were collected from eight cancer types (including breast) which each also had one tumour-adjacent normal tissue sample. Additionally, a shared set of highly methylated domains (HMDs) were subsequently classified as having no PMDs in any of the tumour samples (N = 830; genomic coverage = 32%). These PMD annotations were identified from whole-genome bisulfite-sequencing (WGBS) data using an HMM-based two-state classification tool – called MethPipe (Song et al., 2013) – which was applied to a window size of 10kb in each sample. I downloaded these from https://zwdzwd.s3.amazonaws.com/pmd/PMD_coordinates_hg38.bed.gz

Breast tumour PMD annotations:

Brinkman et al. (2019) described PMDs across 30 primary breast tumour samples which had been collected using whole-genome bisulfite-sequencing and also processed using MethPipe, as described above (Song et al., 2013). PMDs were detected using the R package MethylSeekR, which was designed to identify tracks of hypomethylated regions in the aim of identifying regulatory elements from base-level high resolution bisulfite sequencing data (Burger et al., 2013). I downloaded these PMD annotations from https://zenodo.org/record/1467025#.YvO-yOzMJqt

2.11 - Enrichment Analyses

LiftOver of annotation files from hg19 to hg38:
Before the enrichment analysis, I had to ensure all annotations and datasets were on the same genome build. This is because I use the *bedtools* package (version 2.29.2; Quinlan and Hall, 2010) to find the intersect between methQTL-CpGs which overlap with these annotations (using the BED files for the annotation and my methQTL-CpGs respectively) and thus I need to have genomic coordinates from a matched genome build. The genomic annotations (both Illumina and *chromHMM*) and PMD annotations (both Zhou common tumour and Brinkman breast tumour) were all in hg19 whilst the methQTLs were in hg38. I used the *LiftOver* command from the UCSC Linux package (version 326) to convert the annotations so that all files were in hg38. Using the *LiftOver* command in *Linux*, I selected an annotation file for conversion along with a specific, relevant chain file containing the genome build conversion (eg. 19 to 38), which can be downloaded from the UCSC.

**Genomic feature enrichment (Odds ratio Score):**

I assessed enrichment and depletion of methQTL-CpGs in genomic features. I used Odds ratio (OR) scores to represent if CpGs were observed in a genomic feature proportionally more or less likely than would be expected. I calculated OR scores from the baseline number of methQTL-CpGs which were present and which were absent from a genomic feature of interest, and then repeated this with CpGs from the array which were not in a methQTL to establish the proportional occurrence of these CpGs in a genomic feature. An OR score then relates these as a ratio. I used *bedtools* (version 2.29.2) to find the intersect between methQTL-CpGs and these genomic features. I calculated OR scores using a Fisher’s test, from which I also obtained a p-value to represent the significance of this OR score. I also recorded 95% confidence intervals, and incorporated these as error bars in the OR score diagrams. An OR score of 1 would indicate there is no proportional difference between methQTL-CpGs and non-methQTL-CpGs in a genomic feature. The three different annotations used for OR scores are described in section 2.9.

**CTCF overlap enrichment analysis:**

To look for methQTL-CpG overlap with CTCF binding sites, I first used Odds ratio scores for enrichment in *chromHMM* CTCF state for CTCF sites. I also downloaded two other
genomic annotations to further describe the methQTLs relative to CTCF sites. The first was
the mammary epithelial tissue ENCODE-derived CTCF region (Pazin, 2015). I downloaded
this from ENCODE under the accession ID: ENCF338TGS
(https://www.encodeproject.org/files/ENCFF338TGS/). The second was the CTCF site
sequence motif from the JASPAR CORE open-access database of experimentally defined
transcription factor binding sites in eukaryotes (Castro-Mondragon et al., 2022). I
downloaded this from Jaspar website (https://jaspar.genereg.net/matrix/MA0139.1/). I
performed this analysis in Linux, using the bedtools package (version 2.29.2) to find the
intersect between the methQTL-SNPs and methQTL-CpGs.
Chapter 3 – Identification of methylation QTLs in breast cancer

3.1 - Introduction

An effective analysis of methQTLs requires an evaluation of the various methods available in two key steps; for pre-processing and filtering the datasets through quality control before the analysis and subsequently for identifying significant SNP/CpG associations. Deciding between pre-processing options requires a comparison of the efficiency with which various normalisation methods can reduce technical variance from batch effects, but also requires exclusion criteria for samples and subjects which generates the highest quality dataset for analysis (Wilhelm-Benartzi et al., 2013; Aryee et al., 2014; Zhao et al., 2017). Following this, there are different approaches to identify associations, ranging from general association tools to methods developed specifically for QTL identification. Furthermore, the model parameters selected, and the substructure accounted for through covariate data in this analysis (such as population structure), may have a strong effect on which methQTLs are identified, and more specifically, which biological effects are being observed (Sul, Martin and Eskin, 2018). In order to make such comparisons, there must be measurable comparators to reliably define poor analysis quality for these methodological decisions, such as genomic inflation. Finally, important caveats of association studies must also be addressed, such as the confounding effects of high linkage disequilibrium on the identification of independently associating variants (Neumeyer, Hemani and Zeggini, 2020).

In this chapter, I aim to establish an optimised methQTL analysis framework that will effectively identify methQTLs which represent genetically programmed tumour methylome changes in a breast cancer cohort. I will do this by considering the merits and caveats of the methods available for each of these steps. I will achieve this by defining metrics for
evaluating the quality of these methQTLs and subsequently comparing these metrics across methods. In a first step, I will analyse pre-processing methods to ensure the data has been sufficiently normalised and that the cohort consists of samples and subjects which have met stringent quality control thresholds. In the next step, I will then make a series of comparisons between the modelling choices which can affect the methQTLs identified from the analysis, including statistical properties and assumptions of the data, addressing population sub-structure (amongst other covariates), and finally ensuring methQTLs are representing independently-associating SNP effects.

3.2 – Description of CpG methylation data

The Cancer Genome Atlas (TCGA) is a genomics resource which aims to characterise various cancers through the study of primary tumour and matched tumour-adjacent normal samples, collecting a range of molecular data (including methylation) using microarrays (Collins and Barker, 2007; Ciriello et al., 2015; Thennavan et al., 2021). I first asked how much DNA methylation data was available for breast tumours from the TCGA. For clarity, I will refer to each individual of the cohort as a subject, and each microarray hybridisation as a sample. The methylation data included 783 samples from 704 different subjects, and each sample had CpG methylation measurements for 485,512 CpG probe sites from the Infinium 450k microarray. Of these samples, 710 were primary solid tumour samples and 73 were matched tumour-adjacent normal solid tissue samples. Furthermore, there were five samples which had a biological replicate, which are defined as multiple DNA portions taken from a single tissue sample. Additionally, there were also two samples which had a technical replicate, which are defined as multiple extractions of a single DNA portion. Overall, for the purposes of this analysis, I will focus on the analysis of primary breast tumour methylation data to represent the tumour methylome.

The most appropriate analysis would be between normal genotype samples and tumour methylation samples which were both from breast tissue. However, there are two challenges with this; the low sample number for matched tumour-adjacent breast tissue
and the caveat that tumour-adjacent normal tissues are not necessarily an accurate representation of normal tissue. Normal-adjacent cell populations may not be entirely normal, and there are other interactions between these neighbouring cell populations; tumour-adjacent tissues have been shown to be affected by pro-inflammatory signalling signals from proximal tumours (Aran et al., 2017). Therefore, to maintain sufficient power for association analysis and control over the cell content of the normal samples, I will focus on samples from normal blood tissue instead of normal breast eQTLs.

3.3 - Background of pre-processing, normalisation and quality control methods for Infinium 450k CpG methylation data

CpG methylation can be measured from a sample in various ways. The two most common approaches include either sequencing whole genome (or whole exome) bisulfite-treated DNA or using CpG probe microarrays (Huang, Huang and Wang, 2010). Thus far, microarrays have proven to be the most affordable and popular option in practice (Huang, Huang and Wang, 2010). Whilst originally only covering 27,000 CpG sites, Infinium microarrays have evolved to cover almost 1 million CpG sites genome-wide (Morris and Beck, 2015; Pidsley et al., 2016). Many methylation studies have used the Infinium 450k array, which was designed to have specifically high gene/CGI coverage and measures over 480,000 CpG probes by genotyping bisulfite-converted DNA and highlighting methylated CpG sites throughout the genome (Bibikova et al., 2011). Briefly, this array uses both single-probe assays (where the methylation state is determined by the colour channel measured during single base extension) and multi-probe assays (where there are distinct probes for unmethylated and methylated sequence) to measure methylated and unmethylated signals at CpG sites with high depth of coverage (Dedeurwaerder et al., 2014). The measured beta values represent the percentage of methylation for a population per CpG site; They are either methylated (1), unmethylated (0), or an intermediate value if only some of the population have methylation at this CpG site (which can occur during different phases of the cell cycle and DNA replication) (Zhang et al., 2017). As such, the average CpG methylation distribution across the genome is therefore bimodal.
There are four key steps when pre-processing Infinium 450k methylation data: normalisation, sample filtering, probe filtering and batch correction (Wilhelm-Benartzi et al., 2013; Dedeurwaerder et al., 2014). An optimal normalisation approach will reduce technical variation in the beta distributions without removing biological variation, such as from different cell types. Building on previous methods, functional normalisation incorporates control probe signals to assess unwanted variation and improve technical replicate correlation over established approaches to preserve biological signals (Aryee et al., 2014; Fortin et al., 2014). However, there is still between-sample variation and modest batch effects in this approach. A more recent study described a single-sample adaption of NOOB normalisation called ssNoob; this addresses the issues of batch effects and background signal correction for better normalisation of cross-generation methylation data and single samples (Triche et al., 2013; Fortin, Triche and Hansen, 2017). When comparing ssNoob and functional normalisation, the authors reported that while both performed highly in reducing variation, ssNoob performed better in joint normalisation of data from mixed arrays. These are just some of the choice methods for broadly normalising array methylation data, but there are numerous approaches which may have different advantages, in various specific contexts, which have been compared and systematically evaluated elsewhere (Welsh et al., 2023).

The main challenge with sample filtering is the inconsistency with which it is applied across pipelines (Morris and Beck, 2015). Sample quality is often assessed by control probes which represent technical variance, by measuring bisulfite conversion and hybridisation quality (Bibikova et al., 2011). In Infinium arrays, a clustering approach to evaluate raw methylation and unmethylated probe signals can identify poor samples as outliers/sub-clusters with relatively lower signals (Bibikova et al., 2009). These two approaches have been adapted by methylation processing pipelines, such as minfi, for basic quality control and processing (Morris and Beck, 2015). However, neither of these approaches are quantitative and so sample filtering remains modest in the Infinium quality control workflow.
When filtering CpG probes, SNPs which disrupt efficient hybridisation of the probe binding region can compromise probe hybridisation efficiency (Price et al., 2013). Additionally, non-specific binding of cross-reactive probes can generate spurious signals (Chen et al., 2013). Removal of probes with weak signals which are undetected in the background also accounts for poor quality probes (Heiss and Just, 2019). These filtering criteria are well described in a study from Zhou, Laird and Shen, (2017) who assessed probe quality functionally for their hybridisation quality, rather than from probe SNP positions.

3.4 - Evaluation of normalisation methods shows that ssNoob performs best at reducing technical experimental variance in breast tumour methylation samples

Given the large number of different approaches to normalising Infinium 450k methylation data, it was necessary to evaluate the performance of a selection of available methods for normalising the TCGA methylation data. I first chose two approaches to compare against a raw normalisation approach (where there is no normalisation applied), both been shown to perform well at reducing technical variance when evaluated against other approaches; functional and ssNoob normalisation (Fortin, Triche and Hansen, 2017). While there are other methods which have been evaluated in more detail elsewhere (such as SWAN and Quantile normalisation methods discussed in the Fortin paper), I only considered functional and ssNoob normalisation as they are two recent and popular approaches with relatively high performance reported (Fortin, Triche and Hansen, 2017).

I compared the beta value correlation (measured with spearman’s correlation coefficient Rho, ρ) within the five biological and two technical replicate sets. I chose Spearman’s over Pearson’s correlation metric as Spearman’s makes less statistical assumptions about the linearity of the beta profiles. Replicate correlations were all high following each of the normalisation approaches (Figure 7A), while technical replicates showed higher correlation than biological replicates. However, the median correlation in the most correlated replicate was marginally higher following ssNoob (ρ = 0.976) compared to
Figure 7 – ssNoob normalisation reduces technical variation between replicates:

Jitter plot showing the correlation between the beta values measured in replicate samples using different normalisation methods for A) all probes after quality control (N = 408,820) and B) only mid-range probes with an average beta value across samples between 0.2 and 0.8 (Raw N = 179,317; Functional N = 163,938, ssNoob N = 164,382). Correlation is measured using Spearman’s Rho.

Replicate samples are of two types: There are five biological replicates (black) and two technical replicates (red). Raw = no additional normalisation applied; Functional = Functional normalisation; ssNoob = ssNoob normalisation. The blue bar denotes median correlation for each normalisation method marked.
functional ($\rho = 0.972$) and raw ($\rho = 0.972$) normalisation. In addition, the replicates showed the smallest standard deviation following ssNoob normalisation (0.009) compared to functional (0.018) and raw (0.013) normalisation. Furthermore, I considered that there would be a large proportional bias towards probes with extreme beta values (near 0 and 1), and so wanted to additionally check how these normalisation approaches would be affected when only considering probes which had an average methylation beta value between 0.2 and 0.8 across all samples - which I will refer to as mid-range probes. Replicate correlations were marginally smaller for mid-range probes when compared with the correlation scores for all probes, but the three normalisation methods showed similarly comparable replicate correlations (Figure 7B).

Finally, as an alternative measurement of normalisation performance, I asked if the effects of different normalisation methods would be reflected in the beta distributions themselves. A standard beta distribution has two modes which represent unmethylated and methylated probes and with only a minority of probes which show intermediate methylation levels (Zhang et al., 2017). I therefore wanted to compare the bimodality of the beta distributions in replicates following the application of each normalisation method. To measure and compare bimodality, I used an R package which uses the “dip test”, which assesses if there are multiple modes in a distribution and represents this multimodality as a coefficient (Hartigan and Hartigan, 1985). I first measured the bimodality of the different normalisation approaches for the replicates showing the highest correlation. As expected, the raw normalisation resulted in the lowest dip test (dip statistic = 0.067). However, Functional and ssNoob normalisation both had a dip statistic of 0.087. Furthermore, this was reflected in a density plot where the replicate showed a comparable distribution following both functional and ssNoob normalisation, contrasted with the weaker bimodality of raw normalisation (Figure 8A). Interestingly, a Wilcoxon ranked test (WRT) showed that their distributions were still significantly different (WRT P-value = 2.9x10^{-5}), suggesting that there are differences between the distributions. However, the differences between these two distributions do not significantly affect their overall modality.
**Figure 8** – Functional and ssNoob normalisation improves bimodality of beta distribution over raw normalisation

**A)** The beta density distributions between the replicates with the highest correlation after raw (Orange), functional (Blue) and ssNoob (Red) normalisation. 

**B)** The beta density distributions between the replicates with the lowest correlation after raw (Orange), functional (Blue) and ssNoob (Red) normalisation. Bandwidth and probe number were constant at 0.023 and 408,809 across all density plots (Probe N after quality control = 408,820).
To ensure my findings were not specific to one sample, I also tested if this finding was consistent in the replicates which showed the lowest correlation in beta distributions. This similarly showed relatively lower dip statistic for raw normalisation (0.071) while both ssNoob and functional normalisation had a dip statistic of 0.098 besides having significantly different distributions (WRT P-value = $1.2 \times 10^{-6}$; Figure 8B). There is therefore no significant effect on bimodality between ssNoob and Functional normalisation. However, taken together, these results still show that ssNoob normalisation performs better at reducing technical variance from the breast tumour methylation dataset than functional or raw normalisation methods.

3.5 - Evaluation of the quality control steps for breast tumour methylation samples informs robust inclusion criteria

Following normalisation, stringent quality control steps are used to filter out low-quality samples/probes which may confound biological effects with technical artifacts (Wang, Wu and Wang, 2018). As such, metrics which reliably assess the quality of samples are crucial to an effective analysis. For example, the proportion of undetected probes in each sample is an effective quality metric as this identifies low-quality samples which have a detrimental effect on probe hybridisation and consequently on signal strength (Heiss and Just, 2019). I wanted to first choose one of these metrics and validate it in the breast tumour sample data to evaluate the accuracy with which it reflects sample quality. I therefore asked how undetected probe percentage correlates with sample quality in the TCGA methylation data. To do this, I considered a simple comparator which would represent poor sample quality based on an assumption that the beta value distribution of the majority of samples would be a similar profile, and that a small number of outliers would have a relatively dissimilar beta distribution to the majority of samples. Thus, this comparator would identify outlier samples which have diverged from the average sample beta profile. I used the mean beta distribution across all samples as a standard to generate a sample correlation score for each sample based on their similarity to this mean beta profile. I could therefore establish if samples showing poorer correlation with the mean beta profile (i.e.
poor quality sample outliers) also showed relatively larger percentages of undetected probes.

I plotted a scatterplot of the undetected probe percentage and the sample correlation score and calculated a linear regression between the two. The plot showed no evidence suggesting a strong linear relationship, but instead showed a large clustering where the undetected probe percentage was low and the sample correlation was high (Figure 9). However, while the effect size (beta) was modest, there was a significant negative linear association between undetected probe percentage and the sample correlation score (Effect size beta = -0.176; P-value = 6.1x10^{-7}). While significant, this model could benefit from further statistical analysis, such as for heteroskedasticity in the correlation values. Regardless, this shows that high undetected probe percentage has a modest association with outlier samples, but justifies its inclusion as one of many metrics which collectively ensure a high standard of quality control.

As a final assessment to ensure the accuracy of undetected probe percentage as a quality control metric in the breast tumour data, I also analysed the beta profiles of samples which were above and below an undetected probe percentage threshold for sample quality. A 1% minimum threshold has been used as a common conservative threshold in literature, and the majority of samples (90.92%) in the tumour data was below this threshold (Espinal et al., 2017; Heiss and Just, 2019) (Figure 10). There was also an observable difference between the mean beta distribution plots of samples with greater/less than 1% undetected probes (Figure 10B); high quality samples showed a significantly different beta distribution from low quality samples (Wilcoxon Ranked Test; P = 0.012). Moreover, the high-quality sample group also showed a modestly more bimodal beta distribution (Dip statistic = 0.077) compared to the low-quality sample group (Dip statistic = 0.074). Therefore, the undetected probe percentage associates with sample quality in breast tumour methylation data.
Figure 9 – Undetected probe percentage has negative correlation with sample quality

Scatterplot of undetected probe percentage is plotted against a correlation score for how similar a sample’s beta distribution is from the average beta profile of the entire sample cohort (Sample N = 793; Probe N = 485,512). Red line indicates linear regression between the sample correlation score and the undetected probe percentage ($P = 6.1 \times 10^{-7}$; Effect size beta = -0.176).
Figure 10 – Methylation samples with low undetected probe percentage have greater bimodality in breast tumours

A) Histogram of the undetected probe percentage across tumour breast tissue. B) Density plots for the average beta distribution for CpG probes. Samples which have an undetected probe percentage (UPP) less/more than 1% are in red/black respectively. Number of probes (485,512) and bandwidth (0.023) are consistent across both densities.
3.6 – Description of SNP genotype data

I next focussed on establishing the inclusion criteria for the genotype data from normal tissue samples. Normal tissue represents germline polymorphisms which are independent of somatic tumour mutations, and thus I will analyse normal tissue genotype data for associations with tumour methylation patterns. I therefore asked how much normal tissue genotype data was available from TCGA. The genotype data from TCGA breast cancer project contained matched samples from both normal and tumour tissue. Overall, this dataset had 2,265 array hybridisations from 1,098 subjects measured using the Affymetrix 6 variant microarray (906,600 variant probes). Of these samples, 1,109 were from breast tumour whilst 138 were from tumour-adjacent normal breast tissue. Additionally, 1,018 samples were from normal blood tissue. As mentioned previously, I chose to focus on samples from normal blood tissue rather than matched normal breast tissue, in order to maintain statistical power for association analyses and to ensure the cell content of the normal samples is consistent from tumour-associated effects.

3.7 - Background of pre-processing, normalisation, quality control and imputation methods for Affymetrix 6.0 SNP genotype data

In comparison to methylation data processing, methods and challenges for pre-processing and genotype calling for large genetic datasets have been relatively well established primarily due to the breadth of GWAS studies that are being performed (Affymetrix Inc, 2006; Carvalho et al., 2007; Korn et al., 2008). Genotype can be assayed from either whole genome/exome sequencing or by microarrays, the latter of which has proven to be the more popular and affordable option (Verlouw et al., 2021). Affymetrix/ThermoFischer have released arrays which have grown from 100,000 probes to 900,000 individual SNPs in the popular Affymetrix6 genotype array, while the most recent Affymetrix microarray may have as high as 2.6 million markers with Axiom microarrays (McCall and Almudevar, 2012; Verlouw et al., 2021). Illumina also have a suite of genotyping microarrays, such as the Omni5 which has over 4 million variants and more than 80%
genomic coverage (Verlouw et al., 2021). However, genome coverage is not the sole indicator of the quality of microarrays, as the subsequent quality of imputation has become an equally important feature and one which not necessarily correlates with genomic coverage (Verlouw et al., 2021). Regardless of the microarray, there are some shared caveats that must be addressed when analysing genotyping array data, namely through population bias and genotyping errors.

Pre-processing of genotype Affymetrix 6.0 array data involves quantifying raw intensity signals from probes which may then be used to call genotypes (Affymetrix Inc, 2006; Carvalho, Louis and Irizarry, 2010). As these intensities are mapped during genotype calling, posterior probabilities for each of the three genotypes (AA, AB and BB) are then used to generate a confidence measure for the genotype call, which must be specific to the SNP due to the variation in probe intensities (Irizarry et al., 2003; Carvalho, Louis and Irizarry, 2010). Genotype calling is thus a common source of error, as it must address common challenges; low sample size and batch effects can induce uncertainty when calling genotypes. Batch effects are defined as the variance observed when measuring probe signals and copy number in samples from a cohort where subjects have been measured under different conditions, or “batches” (Leek et al., 2010). These can therefore create spurious genotype-phenotype associations, due to different calling methods used or population-specific genotypic differences, and so must be accounted for in the genotype calling approach (Miclaus et al., 2010; Seo et al., 2019).

There are commonly used methods which have great genotyping accuracy through the use of robust linear models, such as BRLMM and CRLMM, which call SNPs after being trained on known genotypes from haplotype map (HapMap) calls to set confidence thresholds (Carvalho, Louis and Irizarry, 2010). BRLMM is a Bayesian robust linear model from Affymetrix which normalises probe intensities for each allele at a SNP, but concerns over its accuracy and susceptibility to batch effects has been a limitation (Affymetrix Inc, 2006; Miclaus et al., 2010). An evolution of this, the corrected robust linear model CRLMM, is reported to have better quality control metrics for SNPs and consequently better batch correction by using known genotypes from HapMap (Carvalho et al., 2007; Lin et al., 2008). One recent study improved batch correction by first assessing probe signal heterogeneity.
(using multivariate-ANOVA) and then clustering samples by probe signal similarity and calling SNPs accordingly (Seo et al., 2019). This addressed an issue whereby calling genotypes with large sample sizes is only accurate when there are no significant differences between the probe signals of multiple batches. With regards to subject-level filtering, quality control steps are largely similar to other datasets, such as methylation filtering. However, samples are primarily filtered by their SNP calling quality confidence score, amongst other less routinely used examples which have been covered in quality control reviews (Zhao et al., 2017).

Beyond batch effects, the presence of population structure within these large genotyped cohorts has a significant effect on association studies. While the aim of population studies is to identify genetic variation which associates with a trait, there is commonly confounding variation where subjects share a degree of ancestry, and thus share some degree of related genetic variation (Sul, Martin and Eskin, 2018). Thus, it is essential for association models to account for population structure to ensure the associated genetic variation identified is excluding common ancestral variation. This can be addressed by including ancestry as a covariate in current methods which primarily use linear mixed models for an association analysis, but there may still be incomplete reporting of ancestry or unknown relatedness (Yu et al., 2006; Zhou and Stephens, 2012; Sul, Martin and Eskin, 2018). One method for addressing population stratification is to analyse principal components of genotype data and to observe how the cohort consequently forms ancestry-associated clusters (Price et al., 2006). Principal components can then be used as a covariate to ensure that the model accounts for this genetic structure and thus limits spurious associations.

The final step of this process is the imputation of microarray-based genotype data. Imputation involves using genotyped SNPs from a microarray, and their linkage disequilibrium with surrounding SNPs, to infer the genotypes of those surrounding SNPs and improve genomic coverage (Li et al., 2009). This addresses the issue of probe bias from the limited genomic coverage of microarrays and consequently allows for genome-wide genotyped data. However, imputation also necessary to make comparisons across different studies (or different genotyping platforms) which may use different sets of SNP probes or
quality control methods, and consequently have a small set of overlapping SNPs (Marchini and Howie, 2010).

Imputation importantly requires a reference genome from a genetically similar reference population to accurately infer this linkage and achieve high confidence genotyping for sequences which lie between microarray probes, as imputation with genetically dissimilar reference populations can lead to severe inaccuracies (Malhotra et al., 2014). Imputation quality is assessed using correlation scores ($R^2$) for each variant to represent how well the SNP fits the reference haplotype. For the genotype imputation tool Minimac2, $R^2$ can be defined as the squared correlation between the observed (imputed) genotype dosage variance and the expected genotype dosage variance (assuming allele frequencies are in hardy Weinberg equilibrium) (Fuchsberger, Abecasis and Hinds, 2015; Minimac Info File - Genome Analysis Wiki). Reference panels for European populations are widely available through projects such as the Haplotype reference consortium (HRC), which is a large reference panel specifically focussing on low-frequency variants as candidate loci for disease (McCarthy et al., 2016). Indeed, the 1000 genomes project have sampled individuals from populations across Africa, Europe, Americas, south Asia and east Asia (1000 Genomes Project Consortium et al., 2015). However, while these reference resources are abundant for European ancestries, these resources are often less extensive in less developed populations, such as in African populations or smaller, native communities (Malhotra et al., 2014; Vergara et al., 2018; Schurz et al., 2019). While tools are available to perform imputation locally, the large data storage needed for a range of reference genome sequences, and the high computational demand of imputation, has led to the development of imputation servers which host large-scale post-GWAS imputation/analysis, such as the Michigan imputation server which uses Minimac2 (Fuchsberger, Abecasis and Hinds, 2015; Das et al., 2016).
3.8 - Evaluation of population stratification in the cohort justified a single-ancestry European cohort of genotype samples from normal blood for analysis

Population structure within a cohort can cause bias in genetic association studies. Where two ancestries show phenotypic variance, all genetic factors delineating these populations will appear to be significantly associated with the phenotype regardless of relevance or causality for that trait (Sul, Martin and Eskin, 2018). As such, any degree of known/unknown shared ancestry between samples can create spurious associations of false positive SNP signals. I therefore assessed the population structure in the TCGA cohort before I began my analyses. The TCGA cohort was collected from subjects living in the United States of America. The cohort metadata included self-reported ancestry annotation for each subject. I examined these annotations for 1,018 subjects who had genotyped whole blood samples, and asked whether population structure might exist in the cohort (Table 2). This analysis revealed that the majority (67.09%) of subjects were annotated as being of European ancestry. However, a portion of subjects were annotated as African-American or Asian ancestry (17.19% and 5.99% respectively). In addition, 9.53% of subjects had no reported ancestry. These annotations suggest that population structure is likely to exist in the TCGA cohort.

<table>
<thead>
<tr>
<th>TCGA ancestry annotation</th>
<th>Number of subjects</th>
<th>Percentage of subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>European</td>
<td>683</td>
<td>67.09%</td>
</tr>
<tr>
<td>African American</td>
<td>175</td>
<td>17.19%</td>
</tr>
<tr>
<td>Asian</td>
<td>61</td>
<td>5.99%</td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>2</td>
<td>0.20%</td>
</tr>
<tr>
<td>Not reported</td>
<td>97</td>
<td>9.53%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,018</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 - TCGA ancestry annotations for cohort

TCGA ancestry for the 1,018 subjects with normal blood SNP array genotype data.
Following these observations, I decided to focus on analysing subjects from the largest ancestry within the cohort, which was European ancestry. This focus will minimise population structure whilst providing the greatest statistical power for an association study. I proceeded to assess how accurately the European ancestry was described by the TCGA annotations; these annotations could contain reporting errors that could potentially reduce the quality of these classifications. To evaluate their accuracy, I used Plink (v1.90b1g, Purcell et al., 2007) to generate principal components (PCs) from the cohort genotype data to capture genetic similarity between individuals. By comparing the subject clustering against the TCGA annotations, I can examine their correlation. This assessment revealed three distinct groupings whereby subject ancestry appeared to cluster accordingly with their TCGA annotations (Figure 11). Whilst most individuals fell within these groupings, a number of subjects were ambiguous in their clustering, raising questions over their classification accuracy by annotation. Additionally, many non-reported subjects were located within one of these three clusters - most commonly in the European grouping. The self-reported TCGA annotation accuracy therefore does not entirely capture genetic similarity.

I therefore asked if I could more accurately classify ancestry to reduce population structure. To classify the ancestry, I used unsupervised PC clustering whereby subjects are categorised based on genetic similarity. I adapted a custom R tool (written by David Clark & Paul Timmers for internal lab group analyses, and thus unpublished) which had been developed for identifying outlier ancestries (from rare population backgrounds) and generating a single ancestry cohort from island populations. This is achieved by comparing the cohort genotype data against reference genotype data for common backgrounds to filter out population outliers. While this analysis therefore removes outliers from a single majority population (without knowledge of annotations), it may also be adapted for the purposes of classifying ancestry where the desired population is the majority in a cohort. I therefore used this tool to identify outliers from genotype clustering to assign European ancestry by genetic similarity. As such, any subjects which clustered in the European ancestry grouping was classified as such (regardless of previous TCGA annotations). Additionally, any subject which was beyond the cluster centroid of the European ancestry group was designated as an outlier. To ensure the non-European clusters were distinct, I
Figure 11 – TCGA ancestry annotations form discrete genetic clusters

Scatter plot of the first two principal components (PCs) from subject genotype data from 1,018 subjects (annotated according to TCGA ancestry). Brackets show percentage of variation represented by each respective principal component. Legend describes four TCGA ancestry annotation categories; European (purple; 683 subjects), African American (green; 175 subjects), Asian (red; 61 subjects) and Other (blue; 99 subjects) – which contains American Indian/Alaskan native and those with an unreported ancestry.
also compared the European cluster in TCGA data against alternative background reference principal components from the 1000 Genome Project; African American ancestry was represented by the American South West (ASW) reference and Asian ancestry was represented by Han Chinese in Beijing (CHB). This approach derived a more uniform European cohort which was genetically distinct and with reduced population structure.

Out of 1,018 subjects, 774 were classified as belonging to European ancestry by clustering, an increase from the 683 subjects who were annotated as such in the TCGA (Figure 12). Out of these 774 subjects, 651 had been previously annotated as European, highlighting those 32 subjects of the 683 annotated European subjects who were misreported as such. There were 37 subjects which were annotated as having African American or Asian ancestry (31 and 6 respectively) which were classed as European by clustering. Additionally, out of the 96 subjects with no reported ancestry annotation, 86 were classified as European ancestry. This classification of European ancestry by clustering has therefore generated a single-population European cohort bolstered by samples which were inaccurately self-reported from TCGA annotations.

3.9 - Establishing an optimised inclusion criteria for pre-processing and quality control of genotype and methylation data

After applying ssNoob normalisation to the methylation data, using the R package minfi, I established inclusion criteria for quality control filtering of methylation samples (Table 3). Methylation samples were first removed if they were male. Samples were classified as being of high quality if they had less than 1% undetected probe percentage and if the methylated and unmethylated probe signals (measured as the log2 of the median probe signal) were both high and clustered (both greater than 10.5) a method of quality control adapted from the minfi R package (Aryee et al., 2014). Samples were also specifically from European ancestry and from primary tumours. Additionally, filtering of Infinium 450k microarray probes also had distinct inclusion criteria. A significant number of probes (64,144) were first removed for having confounding effects, which includes probes which
Figure 12 – Classifying EU ancestry by genetic similarity increases genetic similarity and size of EU population cohort

Scatter plot of principal components for entire cohort are plotted against those for two reference populations; American south west (ASW, Red) representing African American ancestry and Han Chinese in Beijing (CHB, Green) representing Asian ancestry. Subjects representing the EU subset of cohort are plotted in black dots (N = 774). Any samples classified as an outlier from this dataset (i.e. Non-EU) are plotted in black crosses (N = 244). Outliers are identified as being outside the EU PC cluster and/or within close proximity to reference clusters.
### Sample Quality Control Steps:

<table>
<thead>
<tr>
<th>Filtering Step</th>
<th>Inclusion criteria</th>
<th>No of samples removed</th>
<th>Samples remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Female-only</td>
<td>10</td>
<td>783</td>
</tr>
<tr>
<td>Undetected probe percentage</td>
<td>&lt; 1% undetected probes</td>
<td>65</td>
<td>718</td>
</tr>
<tr>
<td>Median methylation/unmethylated probe signal</td>
<td>&gt; 10.5</td>
<td>21</td>
<td>697</td>
</tr>
<tr>
<td>Ancestry</td>
<td>EU by clustering</td>
<td>135</td>
<td>562</td>
</tr>
<tr>
<td>Sample-type</td>
<td>Primary tumour</td>
<td>71</td>
<td>482</td>
</tr>
</tbody>
</table>

### Probe Quality Control Steps:

<table>
<thead>
<tr>
<th>Filtering Step</th>
<th>Inclusion criteria</th>
<th>No of samples removed</th>
<th>Samples remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masking probes</td>
<td>Remove with Zhou masking CpG annotations</td>
<td>64,144</td>
<td>421,368</td>
</tr>
<tr>
<td>Sex chromosomes</td>
<td>Autosomal chromosomes</td>
<td>9,944</td>
<td>411,424</td>
</tr>
<tr>
<td>Non-CpG probes</td>
<td>Remove probes from non-CpGs</td>
<td>3,091</td>
<td>408,333</td>
</tr>
<tr>
<td>Probe call missingness</td>
<td>&lt; 1%</td>
<td>13,530</td>
<td>398,230</td>
</tr>
</tbody>
</table>

Table 3 – Methylation data quality control steps

Stages of quality control for both CpG probes and samples separately. Final probe/sample number is in red.
either have SNPs in/near the probe 3’ ends or alternatively probes with non-unique mapping, and which have been well described by Zhou, Laird and Shen (2017). Probes were also removed which were not in CpGs at all or which were in sex chromosomes; sex chromosome methylation mechanisms are more complex (due to imprinting of maternal/paternal alleles and chromosomal dosage control by X-chromosome inactivation) and thus are harder to analyse alongside autosomes (Gendrel and Heard, 2011; Sharp et al., 2011; Elhamamsy, 2017). Finally, probes were removed if the percentage of probe calls which were missing was higher than the conservative minimum threshold of 1%. Overall, after applying this optimised process of normalisation and quality control, there were 482 subject samples and 398,230 CpG probes remaining to be used in the methQTL analysis.

After the genotype data had been normalised using the R package CRLMM, I established similar inclusion criteria for genotype samples (Table 4) (Carvalho et al., 2007; Carvalho, Louis and Irizarry, 2010). Samples were included if they were female and from European ancestry. Samples were removed which were duplicates or which were not from normal blood tissues. Additionally, samples were removed if the signal-to-noise ratio was less than 4, a threshold which has been evaluated by de Andrade et al. (2011). Finally, samples were also excluded where the percentage of missing genotype calls was greater than 3%. This is greater than the 1% threshold applied to methylation data, but after checking those samples which were between 1-3%, I was comfortable this still represented good quality samples and thus I retained them to maximise sample size. Furthermore, in addition to sample-based filtering, SNP probes were also subject to quality control steps. SNP probes were removed if they were in sex chromosomes, as these are not the focus of this study. SNPs were included if they had a minor allele frequency greater than 5%, less than 1% of missing genotype calls and were significantly in hardy Weinberg equilibrium (P < 1x10^-6). I also filtered probes according to two CRLMM quality metrics; the average SNP probability score and a SNP QC score which assesses SNP quality based on posterior probabilities for a given SNP that any associated batch effects are an outlier (Scharpf et al., 2011). Probes were removed if their average probability score was less than 0.95 and their SNP QC score was less than 0.25, which is the recommended threshold from the CRLMM package for an accurate genotype call (Carvalho, Louis and Irizarry, 2010; de Andrade et al., 2011). I also removed SNPs which had any allele mismatches between samples and the
### Sample Quality Control Steps:

<table>
<thead>
<tr>
<th>Filtering Step</th>
<th>Inclusion criteria</th>
<th>No of samples removed</th>
<th>No of samples remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Female-only</td>
<td>24</td>
<td>2,231</td>
</tr>
<tr>
<td>Duplicate</td>
<td>Remove duplicates</td>
<td>4</td>
<td>2,227</td>
</tr>
<tr>
<td>Signal:Noise ratio</td>
<td>SNR &gt; 4</td>
<td>52</td>
<td>2,175</td>
</tr>
<tr>
<td>Ancestry</td>
<td>EU by clustering</td>
<td>234</td>
<td>1,787</td>
</tr>
<tr>
<td>Sample-type</td>
<td>Normal blood</td>
<td>1,045</td>
<td>742</td>
</tr>
<tr>
<td>Genotype missingness</td>
<td>&lt; 3%</td>
<td>85</td>
<td><strong>657</strong></td>
</tr>
</tbody>
</table>

### Probe Quality Control Steps:

<table>
<thead>
<tr>
<th>Filtering Step</th>
<th>Inclusion criteria</th>
<th>No of samples removed</th>
<th>No of samples remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average SNP probability score</td>
<td>&gt; 0.95</td>
<td>0</td>
<td>906,600</td>
</tr>
<tr>
<td>SNP QC score</td>
<td>&gt; 0.25</td>
<td>0</td>
<td>906,600</td>
</tr>
<tr>
<td>Non-autosomal filtering</td>
<td>Remove sex chromosomes</td>
<td>38,573</td>
<td>868,027</td>
</tr>
<tr>
<td>Genotype missingness</td>
<td>&lt; 1%</td>
<td>82,684</td>
<td>785,343</td>
</tr>
<tr>
<td>HWE significance</td>
<td>P &lt; 1x10^-6</td>
<td>5,777</td>
<td>779,566</td>
</tr>
<tr>
<td>MAF threshold</td>
<td>&gt; 5%</td>
<td>169,622</td>
<td>609,944</td>
</tr>
<tr>
<td>MIS pre-filtering</td>
<td>No allele-mismatches</td>
<td>1,559</td>
<td><strong>608,261</strong></td>
</tr>
</tbody>
</table>

Table 4 – Genotype data quality control steps

Stages of quality control for both CpG probes and samples separately. MIS; Michigan imputation server. MAF; Minor allele frequency. HWE; Hardy-Weinberg equilibrium. Final sample/variant numbers are in red.
reference panel, which was also a prerequisite step for imputation. Overall, after applying this optimised process of normalisation and quality control, there were 657 subjects and 608,261 SNP probes which were the final genotype dataset for analysis.

Finally, I imputed the genotype dataset to infer local SNP alleles based on reference haplotypes. This is an important step in order to make comparisons of significant SNP hits between different GWAS studies, as these studies may remove different SNPs depending on their own quality control process. I used the Michigan imputation server (MIS; used the Minimac4 1.7.1 pipeline) to generate a dataset of fully imputed SNPs for each of the subjects using these probes (Das et al., 2016). I used the haplotype reference consortium (HRC; r1.1 2016) panel as this is specifically for the European population (McCarthy et al., 2016). The fully imputed dataset included 15,005,778 imputed SNPs after basic post-imputation filtering ($R^2 > 0.4$). Overall, of the 657 subjects with pre-processed genotype data the 482 with pre-processed methylation data, there was 333 matched subjects who had matched data from both.

3.10 – Background on methodology of methQTL analyses

While methQTL analyses are fundamentally association studies, and so can use general regression algorithms, studies have often chosen to use QTL-specific approaches which are optimised for large-scale associations with large numbers of simultaneous phenotypes. Software for methQTL studies has typically been unspecific to methQTLs, as regression-based methods developed for eQTL analyses have been sufficiently capable of identifying methQTLs. For example, Matrix eQTL is a computationally efficient QTL-optimised process which uses large matrix operations to rapidly test for transcript/SNP combination associations in large transcriptomic and genomic datasets, and has been successfully adapted in methQTL studies (Shabalin, 2012; Huan et al., 2019; Min et al., 2021).
One important limitation to adapting eQTL regression approaches for methQTL identification is that the bimodal methylation beta distribution violates regression model assumptions of normality and heteroskedasticity. One data transformation which addresses this issue is the M-value which improves normality and the variance error around the extreme values, but these benefits are still limited in small sample sizes (Irizarry et al., 2008; Du et al., 2010). Another consideration which is typically not included in such models is that multiple SNPs or CpGs in a single locus may have correlated methylation associations due to their genetic linkage and proximity (Saito and Suyama, 2015).

As such, recent methQTL-specific methods have prioritised addressing these limitations by designing models which account for this bimodal beta distribution and correlations. One novel tool, called Methylation-Aware Genotype Association in R (MAGAR) first clusters CpGs by their shared methylation associations and then performs a methQTL analysis before analysing for cell type-specific methylation effects, which therefore begins to consider methQTLs for CpG haplotypes rather than individual CpG sites (Scherer et al., 2021). Alternatively, Lyu et al. (2021) use genetic haplotypes of clustered variants and incorporate the beta methylation distribution to improve detection of methQTLs (especially in small sample sizes) in a methylation random field method.

There is one major caveat with the interpretation of methQTL analyses which also applies to general association studies: addressing the linkage disequilibrium between neighbouring SNPs and identifying causal variants. Linkage disequilibrium is defined as the inverse correlation between the high frequency with which local SNPs are inherited together (as a haploblock) and the distance between these SNPs (Slatkin, 2008; Sved and Hill, 2018). As such, while a QTL for a phenotype may harbour multiple genetically linked SNPs, which each show a significant association with that trait, these may not all necessarily have a direct causative effect on that phenotype. In this case, one (or multiple) SNPs may show a direct independent effect and association with a trait (termed a lead SNP) whilst the rest of the variants in the haploblock all associate indirectly through genetic linkage with these lead SNPs (termed secondary SNPs). Importantly, lead SNPs are therefore determined as by the relatively strongest significance of their associations, but this does not necessarily determine that these SNPs are functionally causal.
It is therefore important to discern the lead SNPs from the indirectly-associated secondary SNPs when addressing linkage disequilibrium (Neumeyer, Hemani and Zeggini, 2020). There are various methods which aim to identify independent SNP associations depending on the study aims. One approach to identify independently associating SNPs is a conditional analysis, which considers that if a secondary SNP has a linear relationship with both the CpG and independently of the residuals of the lead SNP/CPG linear regression, then this would infer that the secondary SNP has independent effects on the CpG beyond the lead SNP (Yang et al., 2012). In the instance that there is no need to discern a causative SNP from a haploblock, and the interest is simply for a single SNP association (typically the most significant) to represent a QTL, one option is to only consider one SNP/CPG association which represents a “clump” of associated SNPs (Privé et al., 2019). The process to further infer a causal SNP from a list of candidate independent associations may involve functionally annotating these SNPs and the genes they reside in to tie genetics to trait- or disease-associated biological pathways (Cano-Gamez and Trynka, 2020). This can be achieved using functional SNP enrichment analyses, gene ontological analyses or colocalisation of QTLs, but these methods may still fall short of clearly identifying a single causative SNP if the SNP has a less pronounced functional effect.

3.11 - *RegScan*-based GWAS framework identifies methQTLs more accurately than the QTL-specific association method *Matrix eQTL*

It was important to establish whether the improved efficiency of QTL-specific tools balances with high accuracy of methQTL identification by comparing against general association tools. In order to compare between two methQTL analyses, with different permutations regarding modelling or method options, I require reliable metrics for comparing the resulting association signals. There are two approaches to make such comparisons between the results of methQTL analyses. Firstly, I can visually assess the association signals using a Manhattan plot of the significance values for each association against their genomic location. This ensures the plots appear well-calibrated, whereby most
signals are below a significance threshold but a small number of significant peaks are present.

The second approach is to use tests of significance for genetic variation, such as the heritability or the genomic inflation factor (GIF). The GIF represents the ratio of how the median of the observed test statistic deviates from the median of the expected statistic distribution (Devlin and Roeder, 1999). Whilst high inflation (GIF > 1) can indicate confounding effects (such as strong genetic linkage or from population structure), low inflation (GIF < 1) can suggest either technical artefacts (stemming from how genotype or traits have been measured) or statistical artefacts (van den Berg et al., 2019). Additionally, the uniform distribution of SNP significance and genomic inflation can be visually assessed by a Quantile-Quantile (QQ) plot, which uses quantiles to evaluate the distribution of sample p-values. QQ-plots compare observed sample quantiles against expected theoretical quantiles whereby a linear relationship along a diagonal reference line would infer that the sample distribution is normal. For p-values from an association study, this plot should follow the reference line with only a small inflation of the observed p-values at the lowest quantiles representing a single peak of highly significant associations. QQ-plots can therefore highlight distributional differences (such as genomic inflation) if the observed quantiles are relatively higher than theoretical quantiles (and thus “above” the reference line) for the majority of the distribution. For these reasons, whilst there is no standard scoring of “quality” for association signals, I will define high methQTL quality for the purposes of this study as a description of how typical the significance of these associations is within these plots and metrics.

I first asked if a QTL-specific approach would accurately identify methQTLs in breast tumours. I chose Matrix eQTL, an approach designed for the efficient identification of expression QTLs but which has also been adapted effectively for large-scale methQTL studies (Shabalin, 2012; Huan et al., 2019; Min et al., 2021). I first tested Matrix eQTL using a subset of the Affymetrix 6 array genotype data as a smaller and faster preliminary analysis of all CpGs simultaneously. Across the 333 subjects with matched genotype and methylation data, Matrix eQTL identified 47,888 significant methQTLs (Bonferroni corrected P < 5x10^{-12}) from 608,261 SNPs and 398,230 CpGs. Following a successful analysis of the array genotype
data, I then repeated this Matrix eQTL analysis using the imputed genotype dataset (15 million SNPs). This identified ~470 million significant methQTL associations - which also reproduced the associations identified in the array genotype analysis. This suggests that specific methods such as Matrix eQTL can detect methylation QTLs in breast tumours.

After successfully identifying methQTL associations by Matrix eQTL, I next evaluated the quality of the methQTLs identified by assessing the association signals. For this, I analysed methQTLs for individual CpGs in parallel (as opposed to a single analysis of multiple phenotypes) using the imputed genotype data and compared the results. For these CpGs, I visually assessed Manhattan plots for any evidence of technical artefacts reflecting issues with how the methylation and genotype data had been pre-processed and analysed. While the GIF is a useful comparator, it is only accurate where all associations have been stored regardless of significance so that the p-value distribution is balanced. Indeed, the vast majority of associations are p-values which are not significant and which I did not store to reduce the computational storage burden. Therefore, I did not use the GIF as a metric of quality in this particular analysis.

The methQTLs identified from the imputed genotype analyses of individual CpGs were of mixed quality. For example, for one CpG which had associated with cis-methQTL-SNPs in the array genotype analysis (cg02113055), the Manhattan plot appeared to be of high quality, showing a single peak of significance ($P = 5 \times 10^{-161}$) (Figure 13A). However, for one CpG which had no significant associations in the array genotype analysis (cg15442702), the Manhattan plot appeared to be of low quality, showing severe technical artefacts. Instead of a single peak, significant association signals were uniform across the genome and produced a banding effect around a specific p-value ($P < 5 \times 10^{-20}$) (Figure 13B). This suggested that whilst Matrix eQTL could identify methQTLs, these associations were of inconsistent and unreliable quality. Either insufficient quality control had possibly led to technical artefacts and artificial banding of significance, or there were differences in how the array and imputed genotype datasets had been pre-processed.
Figure 13 – Matrix eQTL identifies methQTLs of variable quality in breast tumours

Manhattan plots showing genome-wide methQTL signals and their significance for cg02113055 (A) and cg15442702 (B). Blue and red horizontal lines denote nominal significance thresholds of $5 \times 10^{-5}$ and $5 \times 10^{-8}$ respectively for scale of significance. CpG position marked by vertical blue line.
I next sought to understand how widely this affected CpGs, and if there were any particular profiles for the CpGs which were showing this artefactual banding in their association signals. To see how different CpGs were affected, I selected six CpGs which were of different methQTL characteristics in the array genotype analyses to compare against CpGs which had shown artefactual banding in the imputed genotype analysis. I considered two CpGs which showed cis associations, two which showed trans associations and two CpGs which showed no significant association signals in the array genotype analysis. Overall, five out of the six CpGs showed no evidence of poor quality in the imputed genotype analysis. Only one of the CpGs with no associations in the array genotype analysis had moderate artefactual banding in the imputed genotype analysis. This suggested that the technical artefacts were less prominent throughout associations which had also been present in the array analysis, but perhaps were more abundant in methQTLs which were specific to the imputed genotype analysis.

To understand the source of this problem, I compared an alternative association analysis approach to ask if the technical artefacts were attributed to the Matrix eQTL method. In contrast to QTL-specific tools, I considered general GWAS methods to identify methQTLs. I used a robust GWAS analysis framework (developed by David Clark & Paul Timmers) which had been developed around the association tool RegScan (Haller et al., 2015) and which included additional options for sample-level and CpG-level quality filtering. I therefore asked if the artefactual methQTLs from Matrix eQTL were of comparably poor quality using this RegScan-based GWAS analysis. I focussed on a single CpG, cg15442702, which had shown artefactual banding in the previous analysis. As this analysis framework stores all association signals, I used both Manhattan plots and GIF to assess the quality of these analyses. The Manhattan plot of the methQTLs identified for cg15442702 using this RegScan analysis showed no artefactual banding in a Manhattan plot (Figure 14A). Additionally, these associations showed no genomic inflation (GIF = 0.98). This was further supported by a QQ-plot which showed a high correlation between the expected and observed significance distributions from the associations (Figure 14B).
Figure 14 – RegScan-based GWAS pipeline identifies non-significant methQTLs associated with cg15442702 methylation

A) Manhattan plot showing variants which associate with CpG methylation and their respective p-values. Blue vertical line indicates CpG position on chromosome 5. Blue horizontal line denotes nominal significance threshold of $5 \times 10^{-5}$. B) Quantile-Quantile plot showing observed significance relative to the expected significance. The software categorises variants by their minor allele frequency when plotting: Green (> 5%), orange (between 1-5%) and blue (< 1%). As MAF filtering has been applied, only common (Green) variants are plotted in this QQ plot. Red line shows where $y = x$. 
Finally, I also made sure that the high quality of the association analysis, from the Matrix eQTL analysis of cg02113055, were also of high quality when using the RegScan analysis. Indeed, these methQTLs produced a high-quality Manhattan plot which was identical to that from the Matrix eQTL analysis, and which showed no artefactual banding and a balanced GIF (0.98) (Figure 15). The RegScan analysis therefore accurately identifies methQTLs and effectively resolves the technical artefacts from the Matrix eQTL analyses. However, it remains unclear if this resolution was a consequence of the alternative association method or of the additional filtering which is established within this analysis.

3.12 – Further investigation of RegScan-based GWAS analysis identifies additional SNP/CpG quality control that resolves technical artefacts in methQTL analyses

I hypothesised that additional quality control steps included in the RegScan analysis would be the most likely explanation for the difference between these analyses. To investigate this hypothesis, I asked if there were any quality control steps which were present in the RegScan analysis which may have been overlooked in the design of my pre-processing inclusion criteria. I identified two additional filtering criteria in the RegScan analysis framework which were candidates for explaining the improvement in methQTL quality. The first was post-imputation filtering of SNPs by minor allele frequency (MAF) to eliminate rare imputed variant effects (MAF < 0.05), which resulted in a reduction from 15,005,778 to 5,441,805 imputed SNPs. The second was the exclusion of sample outliers who failed to meet a set of Z-score thresholds, defined as the number of standard deviations a sample is from the mean phenotypic value. In the RegScan analysis, Z-score filtering was applied for the raw methylation values (Z < 5) and then subsequently again for the covariate-adjusted raw methylation values (Z < 3) (Figure 16).

I asked if the exclusion of these two filtering steps in the RegScan analysis would recapitulate the artefactual banding seen from the Matrix eQTL analysis of cg15442702. I ran the RegScan analysis for cg15442702 after removing post-imputation MAF filtering and
Figure 15 - RegScan-based GWAS pipeline identifies a significant cis-methQTL associates with cg02113055 methylation

A) Manhattan plot showing variants which associate significantly with CpG methylation (P < 5x10^{-8}) and their respective p-values. Blue vertical line indicates CpG position on chromosome 10. Blue and red horizontal lines denote nominal significance thresholds of 5x10^{-5} and 5x10^{-8} respectively. B) Quantile-Quantile plot showing observed significance relative to the expected significance. The software categorises variants by their minor allele frequency when plotting: Green (> 5%), orange (between 1-5%) and blue (< 1%). As MAF filtering has been applied, only common (Green) variants are plotted in this QQ plot. Red line shows where y = x.
Figure 16 - Z-score filtering removes outlier samples and improves quantile distribution for methQTL analysis of cg15442702

Quantile-Quantile plots for cg15442702 before and after applying two steps of Z-score based outlier filtering. **A)** Before any Z-score filtering. **B)** After first filtering step based on sample Z-score from raw methylation (maximum Z-score threshold = 5). Red indicates seven samples marked as outliers for next step of filtering. **C)** After filtering based on sample Z-score from covariate-adjusted raw methylation (maximum z-score threshold = 3). The seven outlier samples have been removed.
Z-score thresholds from the quality control steps. The associations which were identified showed less observed significance than would be expected (GIF = 0.75) and showed similar artefactual banding as seen in the Matrix eQTL analysis of cg15442702 (Figure 17). Taken together, these results show that the Matrix eQTL and RegScan analyses both identify low quality methQTLs with technical artefacts when comprehensive filtering has not been applied. Whilst either method could therefore be used when sufficient quality control standards are applied, the RegScan analysis already incorporates these filtering steps. Furthermore, whilst a key advantage of Matrix eQTL would be its efficiency for high-speed analysis, this difference became negligible through parallelisation of the RegScan analysis to analyse smaller subsets of CpGs simultaneously. Thus, I proceeded with the RegScan analysis in my analysis.

These results highlight that post-imputation MAF and Z-score-based filtering is an essential quality control step for controlling genomic inflation and identifying high quality methQTLs. Rare variants have typically been removed from association analyses with low sample sizes as these lack sufficient power to detect associations with SNPs that are not common (Marees et al., 2018). While MAF filtering of genotype SNPs was part of my inclusion criteria, a second round of MAF filtering post-imputation was required as rare variants were reintroduced through the imputation process and therefore avoided pre-imputation quality control steps. In addition, outlier samples whose methylation levels diverged from a normal methylation distribution may interfere with the modelling of the association analysis and thus have been removed by Z-score filtering in methylation studies (Klajic et al., 2013). In the absence of Z-score filtering, there were seven outliers which had not been excluded. Consequently, these outliers showed disrupted a normal methylation distribution of their covariate-adjusted raw methylation levels (Figure 16).
Figure 17 - Additional variant/CpG quality filtering from Regscan-based GWAS method resolves artefact issues from Matrix eQTL analysis of cg15442702

Comparison of methQTLs identified with the exclusion of additional filtering (A-B) compared to inclusion of additional filtering (C-D). Manhattan plot showing significant associations between variant genotype and methylation at cg15442702 before (A) and after (C) additional quality filtering (based on Z-score and minor allele frequency). Blue vertical line indicates CpG position on chromosome 5. Red and blue horizontal lines mark nominal significance thresholds of $5 \times 10^{-5}$ and $5 \times 10^{-8}$ respectively. Quantile-quantile plot showing the expected and observed significance trend from associations before (B) and after (D) additional quality filtering. Legend describes three separate categories for minor allele frequency which correspond to colours: Green (> 5%), orange (between 1-5%) and blue (< 1%). Red line indicates diagonal line where $y=x$. 
3.13 - M-value transformation addresses statistical limitations of methylation beta values but has no practical effect on methQTL identification

After establishing a workflow for identifying methQTLs, I next considered that the results could be largely influenced by decisions regarding covariates and the statistical limitations regarding the input data. For example, while covariate adjustment will decide what biological effects are reflected in the results, the data used in analysis must be compatible with the assumptions of linear modelling which is fundamental to the RegScan analysis I have chosen to use. Therefore, I compared the results of methQTL analyses for various permutations of choices for which there are multiple options to establish and justify the decisions I make. I selected one of the six CpGs that I had selected to examine throughout the previous investigations, cg15442702. Following this, I asked if the results were consistent across all six of these CpGs.

CpG methylation is measured by the Infinium 450k array using probes which assess methylated (M) and unmethylated (U) CpG signals in a sample (Bibikova et al., 2011). These signals are commonly interpreted as a beta value, which describes the proportion of total probe signals that are methylated for any CpG. However, there are therefore two key limitations with beta values being used in association studies and which suggest data transformation would be required. Firstly, beta distributions for CpG probes in a sample are inherently bimodal (around zero and one). A normal distribution is an assumption of the linear modelling used by Matrix eQTL and RegScan and could consequently lead to false positives if not addressed. Secondly, beta values often become compressed around the extremes of zero and one where they are bounded. Both of these issues can be improved through the conversion to M-values, which are calculated as the log₂ of the M:U probe ratio (Irizarry et al., 2008). M-values are normally distributed and statistically more robust for association analyses, and have therefore been reported to address the limitations of beta values (Du et al., 2010).
As such, I asked if the transformation from beta values to M-values would improve the identification of high-quality methQTLs. I added the M-value transformation as a pre-processing step for the methylation data and ran the methQTL analysis so I could compare between the use of beta values and M-values. There were no clear differences from the conversion from beta values to M-values on the methQTL results when assessed visually by Manhattan plots and QQ-plots or indeed by their genomic inflation (Beta GIF = 0.991; M-value GIF = 0.986) (Figure 18A-D). M-values and beta values thus appear to have a comparable effect on the methQTLs identified for methylation change at cg15442702 when visually comparing their effects. However, this visual comparison of the effects of two different methylation data formats still does not directly compare them. I therefore compared the $-\log_{10}(P)$-values from the beta and M-value methQTL results for each SNP and also compared their quantile distributions. This showed directly that there were no practical differences in the effect of these data transformations (Figure 18E-F). When I compared this analysis to five other CpGs (analysed from the previous section), they similarly showed no practical difference when there were significant methQTL associations. While M-values therefore had no practical effect on the results, they do address the assumption of normality and represent the more statistically robust data for association studies compared with beta values. Given these statistical arguments, I chose to proceed with M-values in my methQTL analyses.

3.14 - Analysis using methylation-adjusted covariate data identifies different methQTLs in breast tumour tissue samples

It is necessary in association studies to incorporate genetic population structure as a covariate to account for any confounding correlations between ancestry and phenotype (Sul, Martin and Eskin, 2018). One common approach is to calculate principal components for genotype data, which use the features of the data which capture most of the dataset variation and which therefore represent genetic structure and ancestry (Bryc et al., 2010; Alhusain and Hafez, 2018). However, in contrast to how genetic structure in a cohort can represent ancestry, methylation structure is thought to represent heterogeneity in the cell content within tissues (Jaffe and Irizarry, 2014). In breast cancer, the methylation landscape
Figure 18 - Transforming methylation data from beta values to M-values has a comparable effect on the methQTL analysis of cg15442702

Comparison of methQTLs identified where the methylation data is in the format of beta values (blue) compared to M-values (orange). Manhattan plot shows comparable associations between variant genotype and methylation at cg15442702 with beta (A) and M-value (C) methylation data. Blue vertical line indicates CpG position on chromosome 5. Red and blue horizontal lines mark nominal significance thresholds of $5 \times 10^{-5}$ and $5 \times 10^{-8}$ respectively. Quantile-quantile plot showing the expected and observed significance trend from associations with beta (B) and M-value (D) methylation data. Legend describes three separate categories for minor allele frequency which correspond to colours: Green (> 5%), orange (between 1-5%) and blue (< 1%). Red line indicates diagonal line where y=x. E) QQ-plot of quantiles for $-\log_{10}(P$-values) from methQTLs identified using M-values and beta values. Red line shows reference trend if distributions are identical. F) Scatterplot of $-\log_{10}(P$-values) from methQTLs identified using M-values and beta values which were greater than 3 in either the beta values or the M-value analysis.
associates with different molecular subtypes; methylation patterns at enhancers and known breast cancer-associated transcription factor binding sites (e.g., ERα, FOXA1, and GATA3) have been sufficient to identify distinct clusters for different subtypes (Holm et al., 2010; Fleischer et al., 2017). Indeed, principal components of methylation data has been shown to differentiate between cell types; for example, the tissue composition of brain and blood samples can be resolved using the top five principal components of methylation data (Farré et al., 2015). By calculating principal components from the methylation data as covariate data, the confounding associations of SNPs with tumour cell heterogeneity can be teased apart from those associations with the CpGs themselves. Accounting for methylation structure as a covariate in association studies, however, is less consistently implemented in methQTL analyses.

I therefore asked if the methQTL association signals were different when methylation structure had been included as a covariate in the analysis. I once again began by analysing cg15442702, before comparing against the five other CpGs (from the previous sections). I compared two methQTL analyses whereby the methylation structure was either adjusted for as a modelling covariate (alongside genotype data) or unadjusted for (where the only covariate is genotype data). For the methylation-adjusted analysis, I calculated principal components from the methylation data and included the first two as covariate variables to represent the variation in methylation structure. I reasoned that a methQTL which is significant in the methylation-adjusted analysis, but not significant in the unadjusted analysis, would be associated specifically with CpG methylation changes and not with the confounding tumour cell heterogeneity. Both analyses produced Manhattan plots that were largely similar in their distribution of significance (mainly in the range of $5 \times 10^{-4}$ and $5 \times 10^{-7}$), whereby there were only a small number of peaks which did not reach Bonferroni-corrected significance threshold (Figure 19A, C). Interestingly though, there was a significant peak in chromosome 4 ($P < 5 \times 10^{-6}$) which was present in the methylation-adjusted analysis but which was insignificant in the unadjusted analysis. When considering the genomic inflation between these analyses, there was an effectively negligible difference between the methylated-adjusted (GIF = 0.981) and unadjusted (GIF = 0.986) analyses, suggesting they have similarly low inflation. In addition, the QQ-plots assessing their distributions were comparably uniform (Figure 19B, D). Finally, when I directly compared
Manhattan/QQ plots of cg15442702: **unadjusted**

A

B

Manhattan/QQ plots of cg15442702: **Methylation-adjusted**

C

D

E

F

Quantiles for methylation-adjusted against unadjusted

Scatterplot of −log_{10}(P) for methylation-adjusted against unadjusted

Unadjusted methylation analysis (−log_{10} p-value)
Figure 19 - Inclusion of methylation structure as covariate identifies unique set of methQTLs for cg15442702

Comparison of methQTLs identified where two principal components (PCs) for methylation data is included (blue) or excluded (orange) as a covariate in the Regscan-based GWAS method. Manhattan plot shows different associations between variant genotype and methylation at cg15442702 with the inclusion (A) and exclusion (C) of the methylation covariate. Blue vertical line indicates CpG position on chromosome 5. Red and blue horizontal lines mark nominal significance thresholds of 5x10^{-5} and 5x10^{-8} respectively. Quantile-quantile plot showing the expected and observed significance trend from associations including (B) and excluding (D) the methylation covariate.

Legend describes three separate categories for minor allele frequency which correspond to colours: Green (> 5%), orange (between 1-5%) and blue (< 1%). Red line indicates diagonal line where y=x. E) QQ-plot of quantiles for -log_{10}(P-values) from methQTLs identified using methylation-adjusted and unadjusted covariate data. Red line shows reference trend if distributions are identical. F) Scatterplot of -log_{10}(P-values) from methQTLs identified using methylation-adjusted and unadjusted covariate data which were greater than 3 in either the methylation-adjusted and unadjusted analysis.
both the distribution and the quantiles of the \(-\log_{10}(P\text{-values})\) between these two analyses, there were no notable differences in their distributions (Figure 19E-F). When I examined these results from the analysis of the five other CpGs, the results were consistent for CpGs when there were significant methQTL associations.

Overall, these results show that tumour heterogeneity is indeed a confounder for methQTL associations in breast tumours and that adjusting for methylation structure as a covariate addresses this effect. While the results of an unadjusted analysis are interesting for what they can infer about the overall methylation changes in the tumour tissue cell population, an interesting comparison would be to have a comparison of the key findings both with and without the adjustment for methylation structure to see how tumour cell composition associates with certain methylation patterns. However, as there will undoubtedly be structure and confounders which I cannot control for, it would be more appropriate (in the context of my aims) to be able to make more conclusive statements about the important overall tumour-associated methylation effects I see, rather than focus on some individual aspects of these methylation changes (like those driven by cell heterogeneity) and ignore others as unimportant. As such, I proceeded without adjusting for methylation structure as a covariate to address my key aims. However, given more time, I would have repeated these experiments with methylation structure as a covariate for additional comparison.

3.15 - A stricter imputation quality score threshold for SNPs has no practical effect on the SNP set which passes quality control due to redundancy with other filtering steps

A major feature of quality control for genotype imputation is that imputed SNPs are highly correlated with the reference haplotype. As such, imputation services identify SNPs with do not meet a minimum correlation score (\(R^2\)) threshold with the reference haplotype sequence and remove them (Charon et al., 2021). Across literature, this threshold has varied; whilst some studies have justified a more lenient nominal threshold of 0.4, which prioritises a high number of SNPs, other studies have chosen a stricter threshold of 0.9 to
ensure imputed SNPs are of high accuracy (Kreiner-Møller et al., 2015; Charon et al., 2021). This depends on balancing the greater power from a large genotype dataset against maintaining a high standard of accuracy for these imputed SNPs.

I therefore asked if two different imputation score thresholds would affect which significant methQTLs were identified from the RegScan analysis. Once again, I continued to analyse cg15442702 for consistency when analysing modelling decisions, before comparing with the five other CpGs (mentioned in previous sections). I used a nominal threshold of 0.4 and a strict threshold of 0.9 to compare; there were 5.4 million SNPs which passed quality control at the nominal threshold but only 4.8 million at the strict threshold. When I compared these analyses, there was no notable change in the methQTLs identified when assessing the Manhattan plots or indeed the genomic inflation (0.4 GIF = 0.986; 0.9 GIF = 0.978) (Figure 20A, C). Moreover, QQ plots were used to assess both the uniformity of each p-value significance and also to directly compare these distributions, but neither analysis showed any significant difference between the two analyses. Additionally, I compared the uniformity (Figure 20B, D) alongside the raw and quantile -log10(P-values) distributions (Figure 20E-F), and saw no appreciable difference between the two analyses. Moreover, when I compared against the other five CpGs, there was similarly no change in effect when a significant methQTL association was found. This shows that the two different imputation score thresholds have no additional effect on methQTL identification.

One possible explanation for the stricter $R^2$ showing no impact is that other post-imputation filtering steps (including MAF filtering) already increases the mean $R^2$ closer to 0.9 in both genotype sets, and indeed this appeared the case as the number of samples filtered out when the MAF $R^2$ filter was removed remained a similar size. As such, I made the practical decision to retain the maximum number of SNPs possible whilst still maintaining high power with a relatively low sample size. Therefore, I proceeded to use the nominal threshold in my analysis.

After evaluating additional steps for post-imputation quality control and how modelling decisions may affect which association signals are identified, I had additional
Manhattan/QQ plots of cg1542702: Nominal threshold (0.4)

Manhattan/QQ plots of cg1542702: Strict threshold (0.9)

Quantiles for nominal against strict imputation score

Scatterplot of $-\log_{10}(P)$ for nominal against strict imputation score
Figure 20 - A strict imputation quality score threshold identifies comparable methQTLs with a nominal imputation score when analysing cg15442702

Comparison of methQTLs identified where imputation score threshold is strict ($R^2 > 0.9$; blue) or nominal ($R^2 > 0.4$; orange) in the Regscan-based GWAS method. Manhattan plot shows different associations between variant genotype and methylation at cg15442702 with the Strict (A) and nominal (C) imputation quality scores. Blue vertical line indicates CpG position on chromosome 5. Red and blue horizontal lines mark nominal significance thresholds of $5 \times 10^{-5}$ and $5 \times 10^{-8}$ respectively. Quantile-quantile plot showing the expected and observed significance trend from associations with strict (B) and nominal (D) imputation quality scores. Legend describes three separate categories for minor allele frequency which correspond to colours: Green (> 5%), orange (between 1-5%) and blue (< 1%). Red line indicates diagonal line where $y=x$. E) QQ-plot of quantiles for p-values from methQTLs identified using nominal and strict imputation score thresholds. Red line shows reference trend if distributions are identical. F) Scatterplot of -log10(P-values) from methQTLs identified using nominal and strict imputation quality score threshold which were greater than 3 in either the nominal and strict analysis.
filtering steps to incorporate into my methQTL analysis process (Table 5). The methylation data used in the analysis was in the format of M-values, whilst methylation structure was not included as a covariate. The first twenty principal components from the imputed genotype data were included as a covariate in the analysis to adjust for population structure. Additionally, the imputed genotypes were filtered to remove SNPs with poor correlation scores with the reference haplotype sequence ($R^2 > 0.4$). Imputed SNPs were also removed if they had monomorphic alleles, if their minor allele frequency was $< 0.05$ and finally if the sample missing call rate for any SNP was greater than 0.03. Z-score filtering of samples was also integrated to filter out sample outliers and was applied first to raw methylation ($Z < 5$) and then subsequently to covariate-adjusted raw methylation ($Z < 3$). Finally, any methQTLs identified would only be significant if the p-value reaches a Bonferroni-corrected threshold of $P < 5 \times 10^{-14}$, as this adjusted p-value accounts for a multiple testing correction. The final inclusion cohort was therefore 333 breast tumour subjects with matched data from 5,441,805 million imputed SNPs in normal blood and 398,230 CpG probes in primary tumours.

3.16 - Discussion

As new methods are made available for performing association studies, greater judgement and evaluation is required to deduce which of these approaches is most effective with the data being analysed. I established an analytical framework which effectively identifies methQTLs in breast tumours using comparative evaluations to justify the inclusion and importance of these steps. I evaluated pre-processing steps, including an effective normalisation method for both genotype and methylation data, and established inclusion criteria which maintained a high standard of quality control. I selected a RegScan analysis for identifying associations between these datasets and identified post-imputation filtering and outlier removal as two sources of technical artefacts that required additional filtering. Finally, I compared options for data transformations, covariates and post-imputation filtering parameters to justify my selections for an optimal framework.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Modelling choice</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPG</strong></td>
<td>Methylation data format</td>
</tr>
<tr>
<td></td>
<td>Methylation PCs</td>
</tr>
<tr>
<td><strong>SNP</strong></td>
<td>Genotype PCs</td>
</tr>
<tr>
<td></td>
<td>Imputation R2</td>
</tr>
<tr>
<td></td>
<td>MAF</td>
</tr>
<tr>
<td></td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SAMPLE</strong></td>
<td>Z-score filtering</td>
</tr>
<tr>
<td><strong>METH QTL</strong></td>
<td>P-value filtering</td>
</tr>
</tbody>
</table>

Table 5 – Post imputation filtering and modelling decisions to optimise methQTL analysis of breast tumours

Modelling and post-imputation filtering decisions across the levels of CpG (orange), SNP (blue) and samples (green). MAF; Minor allele frequency. PCs; Principal components.
While methQTL studies are common, there are a wide selection of methods to choose from. Many studies select an established set of analysis steps for methQTL identification based on popular choices which have been justified elsewhere in the literature by papers comparing the numerous options for each of these steps (such as how Fortin, Triche and Hansen (2017) compared methods for normalising methylation data). Indeed, there are even bespoke analyses for methQTL identification now that include pre-processing, quality control and calling associations between SNPs and CpGs (Scherer et al., 2021). However, these studies do not themselves evaluate which choices may be most appropriate for the various steps involved in a methQTL analysis, nor do they verify that their chosen approach is the optimal method with respect to the datasets they are analysing. As such, while there have been reviews of each of these steps, this process of evaluation and comparison is often absent from methQTL studies. When included however, this process generates a robust analysis for methQTL discovery that is also optimised for the data being analysed.

I chose to proceed with an established RegScan-based GWAS analysis, which integrated comprehensive quality control measures, as opposed to a method designed specifically for QTLs, such as Matrix eQTL (Shabalin, 2012), which would have also required these additional filtering steps to be integrated in to the analysis. MethQTL studies have commonly used QTL-specific tools which were designed for expression QTL analyses but which have been adapted effectively for various molecular phenotypes, such as methylation. For example, Matrix eQTL has been adapted for both simple methQTL analyses (McClay et al., 2015; Hannon, Gorrie-Stone, et al., 2018), meta-analyses of methQTL studies (Min et al., 2021) and also in an EWAS suite called GEM for analysing associations between methylation patterns and combinations of genotype-environment data (Pan et al., 2016). Other examples of QTL tools used for methylation QTL discovery include Fast QTL (Ongen et al., 2016; Scherer et al., 2021; Oliva et al., 2022) and QTL2 (Broman et al., 2019; Mozhui et al., 2023). However, as these QTL-specific methods were designed primarily for expression data, and the assumptions of a normal distribution, these methods do not account for the specific properties of methylation data which violate this assumption of normality.
For example, the correlated methylation levels of neighbouring CpGs has been recently addressed by Scherer et al. (2021) with an R package called MAGAR, which firstly groups local CpGs together into clusters of highly correlated CpGs and then finds SNP associations for those clusters. Another example is the assumptions for linear regression models (such as normality and homoskedasticity), which are not satisfied by methylation data and are not commonly corrected for in modelling. An example is to choose a beta regression model, although this still has limitations when data is skewed toward 0 or 1 (Saadati and Benner, 2014). As these QTL-specific tools remain limited for methQTL analysis, other general association tools – such as Plink (Purcell et al., 2007; Houlahan et al., 2019), RegScan (Haller et al., 2015; Pan et al., 2023) and MERLIN (Burdick et al., 2006; McRae et al., 2018) – have been used for methQTL identification, but these methods all still operate by the assumptions of linear modelling and have the same limitations. Therefore, if time permitted, this analysis could be improved by the integration of association methods which further address these limitations for association tools based on linear models. For example, new methods have arisen to address these problems with using methylation data, such as using a methylation random field which is designed to perform dimension reduction specifically on the beta distribution of methylation (Saadati and Benner, 2014; Lyu et al., 2021).

To adjust for methylation structure, I calculated principal components (PCs) and incorporated them as a covariate in a methQTL analysis to investigate if methylation structure had an effect on methQTL identification. Adjusting for methylation structure primarily accounts for the cell-type composition of a tissue sample in an association analysis. This cell content is variable between individuals and may underpin spurious associations between a phenotype and cell-type-specific methylation patterns if unaccounted for (Jaffe and Irizarry, 2014). PCs have been used in methQTL studies to correct for spurious associations with cell-composition by decomposing cell-type-specific methylation patterns (McClay et al., 2015; Huan et al., 2019). However, there are limitations to this approach as adjusting for methylation structure as a covariate primarily corrects for cell-composition differences between individuals, but disregards the variability of cell-types across individuals. Thus, a strong phenotype association from a single cell type may still be
masked by the noise of other cell-type association signals, even when general cell composition has been adjusted for between individuals.

While it may one day be practical and affordable to study DNA methylation at a single-cell resolution in large populations, this data is currently particularly limited by the high cost of single-cell methylation data generation and consequently sample sizes which are too small for population studies. These limitations have been addressed by studies which attempt to model for cell-type-specific signals from; for example, Rahmani et al. (2019) has used a novel decomposition method called Tensor Composition Analysis where it uses bulk data and known cell-type proportions of individuals to detect single-cell signals (per individual, as opposed to per population) to address the current inadequacy of current single-cell data available and required for profiling DNA methylation at this resolution and specifically for use in association studies. There are also studies which use machine learning to study cell composition heterogeneity, such as MethylNet (Levy et al., 2020) and DeepCpG (Angermueller et al., 2017). In breast cancer specifically, Chen et al. (2019) leverage the findings that there are sub-type specific methylation patterns and that proportions of subtypes are population-specific, to describe improved epi-profiles for molecular subtypes across populations using supervised clustering approaches. Thus, a comparison of different methods for adjusting a methQTL analysis for methylation structure could be one step of the analytical framework which could be improved with more comprehensive analysis approaches.

One finding was that the effects of tumour cell heterogeneity on methylation changes could be accounted for by comparing the results of methQTL analyses whereby methylation structure has been adjusted or unadjusted for as covariate. This supports the idea that methylation structure can be used to represent the heterogenous cell composition of tissues, which is a well-established and significant confounder for methylation changes in cancer. On one hand, this would be an interesting analysis to identify which methQTLs were specifically associated with the methylation effects of cell heterogeneity, and indeed, which methQTLs were associated with methylation effects which were independent of this confounder. On the other hand, however, there is an advantage to considering tumour-associated methylation effects as a whole instead of trying to separate confounding effects.
Without methylation structure as a covariate, we find genetic variants associated with changes in the entire tumour-associated methylome, which will comprise of both tumour-driven changes (what I’m primarily interested in) and confounders (such as cell heterogeneity driven changes). But there will undoubtedly be more complex structure and unknown confounders I can’t possibly account for, and these should be considered equally interesting aspects of tumour-associated methylation change. So, by not correcting for structure, and by considering all methylation changes as one tumour-associated effect, I can have more conclusive findings, as opposed to analysing individual parts of this effect and pretending I have a full understanding of what those parts are. The unadjusted analysis identifies methylation changes that are significant and biologically informative, despite having not accounted for every possible confounding factor.

Whilst there was no difference between the methQTLs identified using methylation data in the format of beta values or M-values, these have been directly compared for their performance and statistical suitability to association studies which use linear regression models (Du et al., 2010). Whilst beta values are simpler and intuitive for biological interpretations, it has been well established that M-values represent a more statistically robust choice as they address assumptions of normality and homoskedasticity required for linear models which are violated by the beta distribution of methylation (Du et al., 2010). As such, there are some methQTL studies that have used the conversion beta values in to M-values (Vito et al., 2021; Shang et al., 2023). Meanwhile, some methQTL studies persisted with beta values with no mention of M-values perhaps due to the downstream analyses being more interpretable using beta values (Pierce et al., 2018; Huan et al., 2019; Lu et al., 2019; Zhang et al., 2021). Indeed, there are some alternative regression models based on alternative distributions, such as beta regression (Ferrari and Cribari-Neto, 2004), that may be where further improvement could be found; one study has even suggested that since methylation does not perfectly satisfy the assumptions of either beta- or M- distributions, a bivariate gamma distribution for beta values may be more accurate still (Weinhold et al., 2016). At this point however, M-value transformation is a sufficient step towards addressing the limitations of beta values.
For comparing between different modelling, filtering and data transformations choices, I primarily focussed on comparing analyses using a single CpG, while also considering if these conclusions were present in the five other CpGs which I had analysed during throughout the chapter. This allowed me to evaluate the effects of the specific modelling changes by presenting the analysis of one, consistent CpG. While the CpG cg15442702 had no methQTL associations which were significant, this was irrelevant to the purpose of the comparisons I was making, which were to ask how the collective association signals changed with different modelling choices. However, six CpGs cannot represent all loci as methQTL identification for any subset of CpGs may differ if the CpGs are affected differently by these modelling decisions. These modelling decisions should have a generally consistent effects on all CpGs and I am unaware of any studies which contradict this. However, there are CpG characteristics that may suggest different loci could hypothetically be affected variably by these decisions. For example, due to the heteroskedasticity of beta values in the extreme boundaries, the magnitude effect of the M-value transformation may be greater for a CpG which is in the boundary percentiles of the beta distribution (near 0 or 1) than for any other CpGs (Du et al., 2010). As such, this analysis would benefit from additional evaluation involving repeating these analyses using a larger selection of CpGs with a focus on how CpGs with different beta value distributions affect these results, to conclusively show the generality of my conclusions.

I effectively compared the performance of three normalisation approaches to justify the selection of ssNoob as a method of controlling for technical variation without confounding biological variation. However, there remained the possibility of additional confounders which I had not directly examined, such as batch effects. Whilst normalisation may generally reduce this technical variation between samples, this does not specifically normalise for different sample processing times and conditions, which may remain a confounder following normalisation. A further analysis I could do to investigate this would be to re-run my pipeline for some of the most significantly associated SNPs, but including batch annotations for samples (by their batch chip number and spatial position on the chip). This would have confirmed if batch effects remained after normalisation, or if they were sufficiently controlled for.
Chapter 4 – Mechanisms which drive methQTL associations in breast tumours are different from those driving methQTLs in normal blood tissue

4.1 – Introduction

Having now established an analytical framework for the identification of SNP-CpG associations, I now aimed to analyse the pre-processed methylation and genotype data for methQTLs. While many methQTL studies have focussed on implicating candidate loci with causal roles in specific phenotypes of interest (such as disease), few have considered them as tools which represent how the genetically programmed changes across the methylome of a tissue. Additionally, the use of methQTLs to interrogate the underlying genetic mechanisms which specifically characterise different tissue methylomes has been difficult when certain tissues have limited data resources. I therefore first asked how methQTLs are distributed across the breast tumour methylome, and how this distribution changes depending on whether the CpG associations are local (cis) or distal (trans) to a SNP/QTL. I then considered what this could imply about the mechanisms which are most prevalent across the breast tumour genome to drive methQTL associations. Next, I asked if the effect sizes of these cis- and trans-methQTLs in breast tumours were different from one another. Finally, I wanted to compare this methQTL characterisation in breast tumours against methQTLs in normal blood to infer if the prevalence of these mechanisms changes between the tissue types.
To date, the majority of methQTL analyses have been studied in the context of normal non-cancer tissue (primarily in blood), as cohorts of other tissues can be limited by smaller sizes. However, the expansion of cohort methylation measurements to include non-blood tissues, including tumours, has facilitated methQTL studies in a greater range of tissues. In cancer studies, methQTLs may be characterised by whether or not they are preserved between normal and tumour tissues or if they are tumour-specific. By identifying tumour-specific methQTLs, studies may better understand how genetic loci may promote methylation changes found exclusively in the cancer methylome.

MethQTL studies have shown that the detection rate and effect size of local cis-associations are much greater than long-range trans-associations in normal tissue (Hannon, Gorrie-Stone, et al., 2018). In line with these results, cis-methQTL mechanisms are hypothesised to largely involve the disruption of local binding sites for transcription factors which mediate local methylation changes (Banovich et al., 2014; Wu et al., 2018). As such, these inherently direct mechanisms result in greater effect sizes. In contrast, trans-methQTL mechanisms are hypothesised to be less direct, whereby a SNP regulates gene expression at one locus for a protein which then affects methylation CpGs at a distal locus, and thus require greater power to detect smaller effect sizes (Pai, Pritchard and Gilad, 2015; Bonder et al., 2017; Huan et al., 2019). As such, while literature comprising studies primarily detecting cis-methQTLs in normal tissues is vast, only more recently have sufficiently powered studies characterised trans-methQTLs (Min et al., 2021). Moreover, comprehensive tumour-specific analysis detecting both cis- and trans-methQTLs have still been limited. As such, my overall aims for this chapter are to identify methQTLs in breast tumour tissue and begin to characterise the SNPs and CpGs of these associations respectively, while also comparing against how methQTLs have been characterised in normal blood tissue.
4.2 – There are significant SNP associations with CpG methylation patterns in breast tumours

I first asked if the RegScan-based methQTL association analysis I have established could identify methQTLs in breast tumours. I analysed associations between 5,441,805 million imputed SNPs from normal blood samples and the methylation levels of 398,230 CpG probes from primary breast tumour samples across 333 matched subjects. I found 446,482 SNP associations with primary tumour CpG methylation at a Bonferroni corrected significance threshold of P < 5x10^{-14}. Thus, methQTL associations are present in breast tumours.

4.3 – Clumping SNPs by linkage disequilibrium identifies independent methQTLs in breast tumours

I then asked if any of these breast tumour methQTL-associated CpGs which had multiple SNP associations across the genome. I looked for CpGs from the 446,482 breast tumour methQTLs which had more than one SNP associated. Out of 5,779 CpGs which were associated with these methQTLs, the majority had multiple SNP associations; only 333 had a single SNP association (Figure 21A).

These results suggested that most CpGs had multiple SNP associations, but did not discern whether these additional SNPs were having independent associations. It was therefore necessary to separate independently-associated breast tumour methQTLs from the secondary SNP associations confounded by genetic linkage. I used LD clumping to identify primary SNPs with the most significant associations, so that each LD block has one representative lead SNP. This identified 13,195 methQTLs which represented genetic clumps (i.e. LD blocks) which had independently associated with methylation at a CpG site. Within 9,534 genetic clumps, there were 6,725 distinct SNPs which associated with 5,779 distinct CpGs across the genome. Only 550 clumps contained more than 200 secondary SNPs, but the largest clump contained 2,348 other secondary SNPs (Figure 22A-B). However, most
Figure 21 – Frequency of SNPs and CpGs in methQTLs identified before LD clumping had been applied

Histograms show the occurrence of CpGs (A) and SNPs (B) in 446,482 significant methQTLs.
Figure 22 - Description of clumps and their associations with CpGs and secondary SNPs

A) Histogram showing the number of clump occurrences for any methQTL-CpG. B) Histogram showing the number of secondary SNPs present in clumps.
methQTLs therefore have an independently associated lead SNP and a relatively small number of secondary SNPs.

4.4 – MethQTLs in breast tumours share loci with methQTLs in normal blood tissue

As tumour-specific methQTLs may directly describe tumour-specific methylation changes, I asked how many breast-tumour methQTLs were unique to tumour tissue and how many overlapped with normal tissue methQTLs. I compared the breast tumour methQTLs against the methQTLs identified in a large meta-analysis of normal blood tissue by Min et al., (2021), which used the same LD clumping method to identify independent loci. I also considered the possibility that if separate studies used different parameters for clumping, then the clumped methQTLs may consequently have different lead SNPs for a single common LD block. I addressed this by assigning each of my methQTL-SNPs with a clump ID (a concatenate of the lead SNP and CpG it associated with) in the breast tumour methQTLs, which then allowed for a comparison of clumps across different methQTL studies irrespective of which lead SNP was reported for a clump. By this method, even if another study had selected an alternative secondary SNP as a lead SNP, this could still be compared with my study.

I compared my 13,195 breast tumour methQTLs against 19.5 million methQTLs from the normal blood meta-analysis. There were 8,767 methQTLs which were shared between these datasets (Figure 23). This finding suggests that the 4,429 (33.6%) of the tumour methQTLs which did not overlap could potentially be tumour-specific. Therefore, whilst most of the genetic variation associating with CpG methylation in tumours is also observed in normal tissues, some of these associations may also be tumour-specific.
Overlapping QTLs between breast tumour methQTLs and normal blood methQTLs

![Venn diagram showing the overlapping loci between breast tumour methylation QTLs (blue) and normal blood methylation QTLs (red). N is the total number of methQTLs in each dataset. Shared QTL number is the subset which overlap.]

**Figure 23 – Breast tumour methQTL-SNPs share loci with normal blood methylation QTLs**

Venn diagram showing the overlapping loci between breast tumour methylation QTLs (blue) and normal blood methylation QTLs (red). N is the total number of methQTLs in each dataset. Shared QTL number is the subset which overlap.
4.5 – The ratio of cis- and trans-methQTLs is more balanced in breast tumours than in normal blood tissue

As cis- and trans-methQTLs have distinct genomic profiles and regulatory mechanisms, the ratio of cis:trans methQTLs may suggest how mechanisms of methylation regulation may be different between tissues (Villicaña and Bell, 2021). I therefore asked what the ratio of cis:trans methQTLs was in breast tumours. I classified methQTLs as cis when the SNP association occurred within 1Mb of the corresponding CpG site, a threshold suggested to be optimal by Fauman and Hyde (2022). Using this threshold, there were 10,500 cis-methQTLs and 2,695 trans-methQTLs. Therefore, I observed a cis:trans ratio of around 4:1 in the breast tumour methQTLs. There is therefore a greater proportion of local SNP-CpG associations in breast tumours than there are distal, long-range associations.

I then compared the ratio of cis:trans methQTLs in breast tumours against this ratio in normal blood tissue. I referred to the Min et al., (2021) meta-analysis for comparison, who had described cis/trans-methQTLs in normal blood. They reported that the frequency of cis-methQTLs was ten-fold higher than trans-methQTLs in normal blood (Figure 24). There was also a significant difference between these cis:trans ratios in normal and tumour methQTLs ($\chi^2$ test; $P < 5x10^{-16}$). The proportion of cis-methQTLs is therefore significantly higher than trans-methQTLs in both tumour and normal methQTLs, but this ratio is more balanced in tumours.

4.6 – The effect size of trans-methQTLs is lower than cis-methQTLs in breast tumours

The effect size of trans-acting QTLs has been reported to be generally weaker than cis-acting QTL effects in normal tissue (Gaunt et al., 2016), likely reflecting the long-range, indirect mechanisms through which a distant SNP can affect a local molecular phenotype. I therefore asked if the effect size of trans-methQTLs was therefore also lower than cis-
Figure 24 – Tumour methylome shows more balanced distributional *cis:trans* ratio across the tumour genome compared with normal methylome

Bar plot showing the ratio of methQTLs which associate with a CpG in *cis-* and *trans*-methQTLs. The *cis:trans* ratio in normal blood was 91.49% to 8.51%. The *cis:trans* ratio in breast tumours was 79.58% and 20.42%. MethQTLs were detected to a conservative Bonferroni corrected threshold of $P < 5 \times 10^{-14}$. Total number of independent methQTLs shown as $n$. Difference is statistically significant by means of chi-squared test ($\chi^2 P < 1 \times 10^{-16}$).
methQTLs in breast tumours. Using the effect slope (as a beta value) of the methQTL association to measure effect size, I found that this trend was common to breast tumour methylation as well; the mean absolute effect slope of trans-methQTLs was significantly smaller than that of cis-methQTLs (Wilcoxon-Sign Ranked test, \( P < 5 \times 10^{-16} \)) (Figure 25A). I also extended the analysis to consider directionality in addition to effect size magnitude. In contrast with the negatively skewed unimodal distribution of the absolute effect slopes, the beta distributions for cis- and the trans-methQTLs all followed a similar bimodal shape around zero with comparable beta densities (Figure 25B). However, there was an observable skew in the histogram towards trans-methQTLs over cis-methQTLs around lower magnitude effect slopes (between -1 and 1).

Finally, I reasoned that a lower magnitude effect size would also result in greater methQTL detection at less stringent thresholds for significance. I assessed whether the cis:trans ratio was sensitive to the significance threshold selected for methQTL detection by repeating the analysis using a series of more lenient thresholds (Figure 26), from nominal association (\( P < 5 \times 10^{-8} \)) to Bonferroni-adjusted association (\( P < 5 \times 10^{-14} \)). At a nominal threshold, the ratio was just below 3:1 (73:27). This was significantly lower than the ratio of 4:1 when the Bonferroni corrected threshold was applied (\( \chi^2 < 5 \times 10^{-16} \)). Overall, these results suggest the effect size of trans-methQTLs in breast tumours is indeed lower than that of cis-methQTLs, as is reported in normal blood.

4.7 – Discussion

A robust analysis of both local and distal SNP-CpG associations in breast tumours, and how these associations might compare against those found in normal tissue, is missing from literature but would improve our understanding of the breast tumour methylome. Indeed, there were SNP associations with CpG methylation in breast tumours representing methQTLs which were both significant and independent of other neighbouring SNPs within an LD block. Most of the breast tumour methQTLs were local cis associations with greater effect sizes compared with long range trans associations, an observation also reported in
Figure 25 – The effect size of *trans*-methQTLs is smaller than that of *cis*-methQTLs in breast tumours

Orange corresponds to *cis*; blue corresponds to *trans*. A) Violin plot conveying summary statistics of the absolute effect slopes of each methQTL alongside the density distribution. Statistical significance of difference between distributions calculated using a Wilcoxon-Signed Ranked Test (*P* < 0.05). B) Histogram of directional beta values for all breast tumour methQTLs. *Cis*-methQTL and *trans*-methQTL bars are overlayed and thus brown both co-occurring. Positive effect slope refers to effect of minor allele, negative refers to effect of alternative allele.
Figure 26 – The ratio of cis:trans increases with respect to stricter significance thresholds

Analysing independent methQTLs at different significance values ranging from a nominal correction (5x10^{-8}) to a Bonferroni correction (5x10^{-14}) applied across both cis- (orange) and trans-methQTLs (blue). Proportion shown as bar plots. Total number of independent methQTLs at each threshold shown as n. The difference between nominal (1x10^{-8}) and Bonferroni corrected (1x10^{-14}) is statistically significant ($\chi^2 P < 1.7x10^{-16}$).
normal tissues. However, the overall ratio of cis:trans in these breast tumour methQTLs was less biased towards local associations compared to those found in normal tissue.

Comparing the genomic distribution and ratio of cis- and trans-methQTLs SNPs and CpGs may indicate certain mechanisms which could be driving methylation changes in different tissues. For example, assuming cis-methQTLs are largely associated with a single mechanism (local disruption in transcription factor binding sites), then a relatively cis-biased ratio would suggest the genetic effect on the methylome in this tissue acts predominantly through these local mechanisms. Indeed, the cis-proportion of ~90% that has been observed in multiple normal blood studies (one of which was a relatively well powered meta-analysis) suggests that these local methylation mechanisms dominate the normal methylome (Hannon, Gorrie-Stone, et al., 2018; Min et al., 2021). Alternatively, assuming trans-methQTLs represent a broader range of indirect mechanisms (eg. distal cis-eQTL, complex transcription factor interactions), then a comparatively less biased cis:trans ratio would suggest a more diverse range in the mechanisms underpinning methylation change in this tissue. Indeed, while fewer studies have made comparisons in cancers, Gong et al. (2019) recently showed the trans proportion from a methQTL analysis across 23 different tumour tissues (using genotype and methylation data from TCGA) to be 10.72% of all methQTLs identified (specifically 13.27% in breast carcinoma), compared to the 20.42% in the breast tumour methQTLs identified in this analysis. While this may suggest that there are proportionally more long-range mechanisms of DNA methylation change in the tumour methylome compared with normal blood tissue, it is likely the large sample size of the normal blood cohort enables a sufficiently powered meta-analysis to detect more trans-methQTLs but also simultaneously detect larger numbers of cis-methQTLs. Therefore, an ideal next step in exploring the cis:trans ratio would be to develop larger datasets of methylation measurements in breast tumours for a more powered analysis of the true scale of trans effects, but this will require tissue-specific methylation cohorts which are sufficiently large enough.

As non-normal tissue cohorts with methylation measurements expand, studies will be better powered to detect and further characterise these trans-acting SNP associations with methylation changes, which will particularly benefit tumour-specific studies. Indeed,
the lower effect size in trans-methQTLs (observed in both normal tissue and in breast
tumour tissue) is evidence that long-range interactions are inherently less direct and require
greater statistical power to detect. However, improved power would also allow for a better
understanding of the true proportion of global CpG methylation which have trans-
associations. Studies which have been sufficiently powered have agreed that trans-
methQTLs are generally thought to affect no more than 5% of CpGs in normal tissues, while
there still remains relatively few studies in specific tumours which have had the power for
such analysis (Villicaña and Bell, 2021). However, one recent study of methQTLs across 23
different cancers has reported a lower proportion of CpGs affected by trans-methQTLs;
Gong et al. (2019) showed the median trans-CpG proportion from their pan-cancer analysis
was 2.93%, but only 0.47% in breast cancer specifically (N = 664 samples). In my breast
tumour methQTLs, the percentage of CpGs which had significant trans associations were
1.45%, albeit with limited power. Greater power will greatly improve accuracy with which
we can understand the scope of long-range methylation associations.

Some of the methQTLs identified were exclusively significant for tumour-specific
methylation changes, showing that some SNPs may associate with methylation patterns
which exclusively characterise the tumour methylome. As previously discussed, tumour
tissues have distinct methylation patterns dependent on their heterogenous cell
composition, but tumour-specific methQTL analyses rely on sufficient data resources which
are not yet available for all cancer types. In one tumour-specific example where there was
sufficient resources, Houihan et al. (2019) identified tumour-specific methQTLs in prostate
cancer and also used colocalisation of these methQTLs with tumour-specific eQTLs (tumour
methylation-expression QTLs; meth-eQTL), which overall allowed them to infer candidate
prognostic cancer genes which may also associate with tumour-specific changes in both
methylation and expression patterns. Alternatively, Zhang et al. (2021) firstly identified
melanocyte-specific methQTLs and then compared against malignant melanoma tumours to
investigate methQTL-CpG preservation (which was found to be 45%), but also found that
60% of melanoma methQTL-CpGs were tumour-specific. Thus, tumour-specific methQTL
analyses can develop our understanding of the role of methylation changes in tumour
progression.
Clumping genetically linked SNPs (using linkage disequilibrium) was used to identify secondary SNP associations within the breast tumour methQTLs and to focus on a subset of independent SNP associations. LD clumping (and a similar process called LD pruning) have been commonly used in methQTL analyses to account for redundant secondary associations for a methQTL (Shi et al., 2014; Hannon, Gorrie-Stone, et al., 2018; Min et al., 2021). However, some methQTL studies have used alternative approaches available for addressing secondary SNPs which may potentially answer slightly different questions about these secondary associations. For example, one popular method for addressing linkage is a conditional analysis, which determines if any secondary SNP for a given genetically linked lead SNP is affecting methylation when conditioned against the independent effects of the lead SNP (by regressing the genotype of a secondary variant against the residuals from the lead variant association with a CpG) (Yang et al., 2012). However, a conditional analysis is specifically testing the statistical independence of each secondary SNP as an additional independent association (i.e. one LD block may have multiple independently associated SNPs) whereas clumping aims to identify the lead SNP to represent an independently associated LD block. Additionally, LD clumping is less concerned with whether there is one or multiple independently associated SNPs within the LD block. As such, a conditional analysis answers a slightly different question to LD clumping; a conditional analysis may generally test the statistical independence of an association, whereas LD clumping does this in a fast and simple way by assuming that the cause is likely due to LD structure. Indeed, there are examples where both clumping and conditional analysis have been incorporated as separate steps in a methQTL analysis (Min et al., 2021). Overall, LD clumping was therefore the most appropriate approach for this analysis for two reasons. First, it addresses secondary associations within the aims of my analysis, as this project was less interested in resolving candidate causal SNPs from an LD block. Second, it is a faster and computationally less intense approach compared with alternative methods, such as conditional analyses.

One limitation in how I have selected significance thresholds for this methQTL analysis is that I have not accounted for local CpG methylation correlations to the same extent that I have accounted for local SNP genotype correlations (i.e. through LD clumping). As neighbouring SNPs may be within linkage disequilibrium with one another, CpGs have been shown to have correlated methylation levels, whereby neighbouring CpGs (such as in
CpG islands) have a greater chance of correlated states of co-methylation due to their proximity (Schatz, Dietrich and Schuster, 2004; Haerter et al., 2014; Sun and Sun, 2019). Indeed, CpGs in normal breast tissue have been shown to have significantly more methylation correlation over short regions (< 50bp) in breast tumours than CpGs do in normal tissue (Sun et al., 2019). This correlation between methylation levels of CpGs can therefore effectively reduce the number of independent tests being made within an analysis. Thus, one solution would be to group highly-correlated CpGs into independent “CpG clumps” before calculating a Bonferroni correction, and this has been incorporated in some studies already (Hannon, Gorrie-Stone, et al., 2018; McRae et al., 2018; Scherer et al., 2021).

There are statistical limitations to these interpretations especially with detecting trans-methQTLs. As discussed, high power is required to detect trans associations with such low effect sizes. Additionally, trans detection may also be affected by the degree of tissue heterogeneity. Another limitation is how data has been filtered differently between this study and others which I have made comparisons with. Different quality control steps and thresholds for association analyses may result in slight differences in cis and trans methQTL detection. For example, whilst a single Bonferroni-corrected significance threshold across all methQTLs may be calculated to address multiple testing burden, there are arguments that cis and trans should have a separate correction. Given that cis-methQTL identification only tests for association with CpGs within a 1Mb window, it could be argued that a less conservative significance threshold correction may be necessary for the lower multiple testing burden of cis-methQTL detection relative to trans-methQTLs. There are some methQTL studies which have used alternative cis/trans analyses thresholds to reflect the different number of independent tests and subsequent Bonferroni-correction (Huan et al., 2019; Min et al., 2021). However, as my aim is to compare across cis- and trans-methQTLs, alternative thresholds for cis/trans analyses may make this comparison less appropriate and so a single conservative threshold allows for consistency.

Taken together, these results show that the balance of cis and trans acting methQTLs in breast tumours are different from that in normal blood tissue, but do share some similar
characteristics, such as a lower *trans* effect size. Overall, this suggests that methQTL mechanisms may act differently in breast tumours than they do in normal blood tissue.
Chapter 5 – The functional context of cis-methQTLs in breast tumours shows minor differences from normal tissue

5.1 – Introduction

Cis-methQTLs represent genetic variation that is associated with local methylation patterns (in a typically tissue-specific manner) and so characterising these cis-methQTLs with genomic information offers further depth to our understanding of the underlying methylation mechanisms between SNPs and local CpG methylation changes. However, while cis-methQTLs have been characterised in normal blood tissue (Min et al., 2021), a similar characterisation is lacking in breast tumours. I therefore asked how the SNPs and CpGs from cis-methQTLs are distributed across genomic features in breast tumours, and if additionally, there are any significant enrichments/depletions; CpG methylation changes would be expected to have a higher effect within regulatory elements of genes (such as in CGIs, promoters and enhancers) but less so in introns or intergenic loci. I next compared this against our understanding of the genomic distribution of cis-methQTLs in normal issue to see if there are any striking differences, or indeed, similarities. Finally, I wanted to interpret these enrichments/depletions in breast tumours by asking what this may allow us to infer about the underlying cis-acting mechanisms and how these fit in to our current understanding of the tumour methylome.

While there has been few studies focussing on their genomic profile in tumours, the genomic profile for cis-methQTLs has been well characterised in normal blood tissue (Villicaña and Bell, 2021). The QTL-SNPs are often located in active regions, including promoters and enhancers in CpG islands. However, the nearby CpGs they associate with are
enriched in enhancers and inactive regions but depleted in promoters and CpG islands. As such, cis-methQTLs typically follow a model whereby the variant disrupts transcription factor binding which either actively or passively results in changing local CpG methylation (Banovich et al., 2014; Wu et al., 2018). We can therefore study how cis-methQTLs are distributed and enriched in different genomic features in tumours to compare against cis-methQTLs in normal tissue, and possibly infer which mechanisms may be driving these local associations change.

5.2 – SNP genotypes associate with local CpG methylation in breast tumour cis-methQTLs

I wanted to characterise the genomic profile of SNPs and their associated CpGs in breast tumour cis-methQTLs. I first examined the complement of breast tumour methQTL-SNPs which were associating with local CpGs (within 1Mb of one another). Out of these 13,195 independent methQTLs, 10,500 (79.77%) were cis-methQTLs which comprised of 6,725 independent SNP loci associated with 4,702 distinct CpGs.

Before continuing with profiling, I next wanted to validate the cis-methQTLs to verify that the analysis was accurately identifying significant genetic variation which associated with local CpG methylation. I performed a methQTL analysis on an individual cis-methQTL-CpG (cg00004073, located on chromosome 2) and then visually checked that these associations were local. I used a Manhattan plot to visualise variants which had significantly associated (P < 5x10^{-8}) with cg00004073 methylation (Figure 27A). This showed a single peak of significant cis-signals (P < 5x10^{-8}) within 1Mb of the CpG, as would be expected for a cis association. The variant with the strongest methylation association from the peak was rs3768921 (P = 5x10^{-120}).

I then asked if there was a linear relationship between the SNP genotype and mean CpG methylation between rs3768921 (genotype G/A) and cg00004073. There was a clear decrease in the mean methylation of cg00004073 in a dosage-dependent association with
Figure 27 - rs3768921 is a significant **cis**-methQTL for methylation at cg00004073

**A)** Manhattan Plot of SNPs across genome which show significant association with methylation at cg00004073. Shows peak of genetically linked SNPs in chromosome 2. Red/blue lines mark nominal significance thresholds ($5 \times 10^{-5}$/$5 \times 10^{-8}$).

**B)** QQ-plot showing inflated significance after $5 \times 10^{-4}$, representing the SNP peak in chromosome 2. Only SNPs with MAF > 5% are shown.

**C)** Boxplot of rs3768921 genotype associating with differential methylation at cg00004073 ($P = 5 \times 10^{-120}$).

**D)** Manhattan Plot showing only significant SNPs from the chromosome 2 peak. This represents multiple SNPs within high genetic linkage in a methQTL. Red line is significance ($P < 5 \times 10^{-14}$).
the alternative allele of rs3768921 (Figure 27C). This confirmed that genetic variation associates with local methylation change in an allele-dependent manner in breast tumours.

While cis-methQTLs in normal blood are commonly found in active genes/chromatin, where they have high potential for a role in transcriptional regulation of local genes (which I will refer to as regulatory potential), it is not clear if this is the case in tumour cis-methQTLs. However, understanding the genomic context and positioning of cis-methQTL-CpGs and -SNPs respectively may reveal insights into their regulatory potential. I therefore asked if rs3768921 aligned with the genomic context and regulatory potential associated with cis-methQTL-SNPs in normal blood. I used the LDProxy online tool (Machiela and Chanock, 2015b), which annotates SNPs based on their regulatory potential using data from RegulomeDB (Boyle et al., 2012) and whether the variant is coding or non-coding.

RegulomeDB consolidates multiomic datasets into a single score which ranges from high (1) to low (7) regulatory potential, depending on the volume and strength of separate data evidence that has been reported for a SNP. For example, a score of 5 indicates that the SNP has data evidence for overlapping a single transcription factor binding site motif or a DNase peak. Meanwhile, a score of 1 indicates that a SNP has data evidence for overlapping an eQTL as well as overlapping motifs and binding signatures for both DNase and transcription factors. The cis-methQTL-SNP rs3768921 was assigned moderate regulatory potential (4), indicating data evidence from RegulomeDB for overlapping a transcription factor binding site and a DNase peak (Figure 28). Furthermore, rs3768921 was located in the promoter of the AGAP1 gene. AGAP1 regulates GTPase proteins in membrane trafficking and cytoskeleton dynamics, and has been suggested to control cell migration and invasion in cancer (Tsutsumi et al., 2020). Thus, rs3768921 appears to fit the genomic positioning and regulatory context of a cis-methQTL-SNP in a tumour-associated gene.

Finally, I asked if there were any local secondary SNPs around rs3768921 which may also have fit the correct profile for a tumour-associated cis-methQTL-SNP, in case there were multiple candidate SNPs for this cis-methQTL association. In order to consider neighbouring SNPs with high genetic linkage, I also examined the broader LD block of SNPs which were genetically linked ($R^2 \sim 1$) with rs3768921 (Figure 29). There were ten SNPs
Figure 28 – Haploblock of genetically linked SNPs around rs3768921 colocalises with AGAP1 gene promoter

A) SNPs from the haploblock are plotted by their chromosome position and their linkage disequilibrium ($R^2$) with the lead query variant rs3768921 (blue) in the top panel. SNPs are coloured by coding (red) and non-coding (orange) variants. High (1) or low (7) regulatory potential indicated numerically and according to SNP annotations from RegulomeDB. Grey lines/peaks represent the combined recombination rate at chromosome positions. B) Genes and their chromosomal positions are shown.
Figure 29 – SNPs with greatest genetic linkage within rs3768921 haploblock show higher regulatory potential

SNPs from the haploblock are plotted by their chromosome position and their linkage disequilibrium ($R^2$) with the lead query variant rs3768921 (blue) in the top panel. SNPs are coloured by coding (red) and non-coding (orange) variants. High (1) or low (7) regulatory potential indicated numerically and according to SNP annotations from RegulomeDB. Table showing variant annotations for top ten variants with highest $R^2$ with rs3768921. Table ordered by $R^2$ and distance from rs3768921.

<table>
<thead>
<tr>
<th>Query Variant</th>
<th>Functional Class</th>
<th>Minor Allele Frequency</th>
<th>RegulomeDB</th>
<th>Regulatory Potential</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>non-coding</td>
<td>rare</td>
<td>common</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td>coding</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

145
within a 30kb window with $R^2 > 0.85$ of rs3768921 which were annotated as having a variable degree of moderate/low regulatory potential (ranging between 4 and 7 on the RegulomeDB score of regulatory potential) (Table 6). This individual CpG/SNP analysis confirms that the cis-methQTL-SNPs identified fit the characteristics of a cis-acting SNP by having an allele-dependent effect on local CpG methylation and residing in active regulatory regions.

5.3 – Cis-methQTL-CpGs are depleted in promoters and CGIs in breast tumours

After briefly assessing the positioning of these cis-methQTL-SNPs and -CpGs, I then more broadly examined their functional context in genomic elements. In normal blood, cis-methQTL-CpGs have been shown to be enriched in enhancers but depleted in CpG islands and promoters (Shi et al., 2014; Min et al., 2021). This aligns with a model whereby cis-methQTL-SNPs may disrupt binding sequences for transcription factors (in CGIs and promoters) which then reprogram methylation in local enhancers (Banovich et al., 2014; Wu et al., 2018). Understanding if this enrichment is also seen in breast tumours would indicate if a similar mechanism is behind cis-methQTL-CpG methylation changes in breast cancer.

I asked if cis-methQTL-CpGs were therefore enriched in any genomic features (such as CpG islands, promoters or enhancers). I used the Infinium 450k array probe genomic feature annotations from the Infinium array annotations which were largely from ENCODE (promoter and enhancers) and the UCSC genome browser project (CpG islands). ENCODE regulatory elements are typically from DNA hypersensitivity or CHIP-seq assays (ENCODE Project Consortium, 2012). Meanwhile, CpG islands were classed as sequences with: GC content > 50%; length > 500bp; expected:observed CpG count per segment > 0.6 (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002). The surrounding 2kb sequence describes CpG island shores, whilst the surrounding 2kb sequence from these shores is defined as CpG island shelves (Irizarry et al., 2009).
Table 6 - Regulatory potential of SNPs with highest $R^2$ and within LD block of rs3768921

Table showing variant annotations for top ten variants with highest $R^2$ with rs3768921. Table ordered by $R^2$ and distance from rs3768921. Regulatory potential is represented as numbers (from 1-7) representing the extent of data evidence for a SNP having a regulatory role (1 is high evidence and 7 is low evidence). Column names are: SNP rsIDs; Chromosome number; Chromosome position on the Human Genome Build 38; Allele complement; Minor allele frequency; Distance from rs3768921; $D'$ is a linkage measurement similar to $R^2$; $R^2$ correlation; Correlation patterns observed between the alleles of these 2 SNPs; Regulatory potential according to the database FORGE DB; Regulatory potential according to the database Regulome DB.
I overlapped these cis-methQTL-CpGs with functional annotations and then calculated OR scores for cis-methQTL-CpGs to ask if they were observed more or less often than expected by chance in any genomic functional elements of interest. Using the Illumina annotations, there was a significant depletion of cis-methQTL-CpGs in promoters (OR = 0.84; \( P = 8 \times 10^{-6} \)), CpG islands (OR = 0.85; \( P = 5 \times 10^{-7} \)) and the surrounding CpG island shores (OR = 0.79; \( P = 8 \times 10^{-11} \)) (Figure 30). In line with cis-methQTL-CpGs being depleted in promoters/CGIs and local to SNPs which have been reported to be enriched in these regions, cis-methQTL-CpGs are enriched in CpG island shelves (OR = 1.19; \( P = 3 \times 10^{-4} \)). Meanwhile there was notably no significant enrichment or depletion in enhancers. Overall, this enhancer feature is the only striking difference between my breast tumour cis-methQTL-CpGs and the genomic context which has been reported in normal cis-methQTL-CpGs (Min et al., 2021).

5.4 – Cis-methQTL-CpGs are depleted in chromatin-based annotations for promoters and enhancers in breast tumours

I next wanted to verify that these enrichments/depletions are robust to the Illumina annotations. I reasoned there should be similar enrichments/depletions in genomic feature annotations which are derived from characteristic chromatin marks. Thus, I asked if these cis-methQTL-CpG enrichments/depletions in breast tumours were also present in these features when the genomic annotations were based on their associated chromatin signatures in cell lines. Furthermore, I used the chromatin mark-based annotations to assess enrichments in other interesting functional elements, such as transcription factor binding sites or inactive chromatin. For this, Taberlay et al. (2014) generated chromatin-based annotations using Hidden Markov models. Hidden Markov models work by binarizing signal input, and this can be used on various molecular measurements, including chromatin marks (Yoon, 2009; Ernst and Kellis, 2012). Taken simultaneously, histone modifications can be aligned to a profile.
Figure 30 – Breast tumour cis-methQTL-CpGs are depleted in CpG islands and promoters across Illumina reference genome annotations

Enrichment represented as the odds ratio for cis-methQTL-CpGs shown as dots (including 95% confidence intervals as bars). Significant enrichments/depletions (Fisher’s test; $P < 0.05$) denoted with their $P$-value in bold. Genome annotation taken from Illumina reference genome.
which corresponds with genomic elements, such as enhancers or promoters. Thus, *chromHMM* genome annotations are derived directly from the sample chromatin, rather than from a reference sample which may be a different tissue type (Vu and Ernst, 2022). I also compared across annotations from two cell lines: MCF7 as a broad representation of breast tumour chromatin and HMEC as a broad representation of normal breast chromatin. The chromatin mark combinations used to profile different functional elements are shown in Table 1.

I overlapped the *cis*-methQTL-CpG coordinates with these annotations and calculated a Fisher’s test to calculate the odds ratio (OR) score for these CpGs, which reflects the proportional abundance of *cis*-methQTL-CpGs in a given feature compared to all other CpGs. There was a significant depletion in promoter regions in both MCF7 (OR = 0.79, P = 2x10^{-9}) and HMEC (OR = 0.76, P = 2x10^{-13}) annotations (Figure 31). However, there was also a significant depletion in MCF7 enhancers (OR = 0.87, P = 2x10^{-2}), but not in HMEC. In both cell lines, there was also an enrichment in heterochromatin regions (MCF7 OR = 1.31, P = 9x10^{-20}; HMEC OR = 1.46, P = 9x10^{-38}). Finally, across both cell lines, there was a depletion in repressed regions (MCF7 OR = 0.48, P = 6x10^{-12}; HMEC OR = 0.36, P = 7x10^{-28}). Overall, the significance and direction of enrichment scores in *cis*-methQTL-CpGs between HMEC and MCF7 cell lines were concordant in their enrichment or depletion.

5.5 – *Cis*-methQTL-CpGs are depleted in PMDs in breast tumours

The observation of *cis*-methQTL-CpG depletion in enhancers, promoter and CGIs in breast tumours could suggest they are therefore more frequent in the intergenic partially methylated domains (PMDs) that are primarily responsible for global hypomethylation in the tumour methylome. This is also suggested by the enrichment I observed in the *chromHMM*-annotation for heterochromatin, where PMDs are most common. In normal blood however, *cis*-methQTL-CpGs were equally distributed across both regions of high and low methylation (Min et al., 2021). I therefore asked if breast tumour *cis*-methQTLs-CpGs were enriched in genomic annotations for highly methylated domains (HMDs) or PMDs.
Figure 31 – Breast tumour cis-methQTL-CpGs are depleted in promoters and CpG islands in chromHMM genome annotations across tumour and normal mammary tissue

Enrichment represented as the odds ratio for methQTL-CpGs shown as dots (including 95% confidence intervals as bars). Significant enrichments/depletions (Fisher’s test; P < 0.05) denoted with their P-value in bold. Tumour (red) and normal (blue) annotations are colour coded.
used annotations defined by Zhou et al. (2018) for a set of 417 shared tumour PMDs (PMD present in at least 21/30 different tumour samples, from eight tumour types, and with 13% genomic coverage) and a set of 830 shared regions with no PMD presence termed highly methylated domains (HMDs in 30/30 tumour samples; 32% genomic coverage). I overlapped cis-methQTL-CpG coordinates with PMD/HMD annotations and used a Fisher’s test to calculate OR scores for HMDs and PMDs, representing methylation gain and loss respectively. Cis-methQTL-CpGs were significantly and comparably depleted in both HMDs (OR = 0.91; P = 2x10⁻³) and PMDs (OR = 0.85; P = 1x10⁻⁵) (Figure 32A).

As PMDs may be different across cancer types (Salhab et al., 2018), I next considered if these depletions were consistent using PMDs annotations specifically from breast tumour samples. I used annotations from Brinkman et al. (2019) who measured PMDs across 30 breast tumour samples. I merged the PMD annotations for 30 samples in to a single aggregated annotation of genomic regions which contained a PMD in at least one of the 30 samples. I then overlapped cis-methQTL-CpGs with these annotations and calculate an OR score by using a Fisher’s test. However, there was no significant enrichment or depletion in the aggregated breast tumour PMD annotations (Figure 32B).

Merging the PMDs in to an aggregate annotation from 30 samples does not take in to account the proportion of those samples in which a CpG occurs in a PMD, and whether or not any enrichment/depletion is necessarily consistent across all 30 of these samples. For ease of reference, I will use the term PMD occurrence to be defined as the number of these 30 samples in which a single CpG is enriched in a PMD. I therefore finally asked what the PMD occurrence was across all cis-methQTL-CpGs in comparison with non-methQTL-CpGs. A PMD occurrence of zero would mean a CpG does not occur in any PMDs, whilst 30 would mean a CpG is consistently in a PMD. More than 40% of the cis-methQTL-CpGs were not in a PMD in any sample (Figure 33A). When the PMD occurrence was below 10, cis-methQTL-CpGs were relatively more enriched in PMDs than other CpGs (Figure 33B). However, above 10, cis-methQTL-CpGs were relatively less enriched in PMDs than other CpGs (Figure 33B). Using a chi-squared test, the PMD occurrence distribution for cis-methQTL-CpGs against CpGs which were not methQTLs were significantly different (χ² = 93.2; P = 2x10⁻⁸). In other words, whilst cis-methQTL-CpGs have a significantly different distribution of PMID
Figure 32 - Breast tumour cis-methQTL-CpGs are depleted in common tumour, but not breast tumour, PMDs and HMDs

Enrichment represented as the odds ratio for methQTL-CpGs shown as dots (including 95% confidence intervals as bars). Significant enrichments/depletions (P < 0.05) denoted with their P-value in bold. A) Zhou annotations are PMDs/HMDs which are common across tumours. B) Brinkman annotations are PMDs described in breast tumours.
Figure 33 - Cis-methQTL-CpGs are relatively enriched in less conserved PMDS and depleted in more conserved PMDs across 30 breast tumour samples compared with non-methQTL-CpGs

Shows the distribution of the frequency by which CpGs occur in a PMD across 30 breast tumour samples (A) and focused with a PMD occurrence of zero removed (B). Orange represents cis-methQTL-CpGs. Black represents non-methQTL-CpGs which are on the microarray background. Statistically significant difference between distributions calculated with chi-squared ($\chi^2$); P < 0.05.
occurrence compared with non-methQTL-CpGs, a more nuanced analysis reveals that they are enriched in breast tumour PMDs when PMD occurrence is rare (< 10) and depleted when PMD occurrence is common (> 10). Hence, cis-methQTL-CpGs are enriched specifically in poorly conserved PMDs in breast tumours.

5.6 – Cis-methQTL-CpGs are enriched in CTCF sites outside of promoters in breast tumours

CTCF is a protein which controls chromatin compaction and structure to mediate long-range regulation of gene expression (Dehingia et al., 2022). Occupancy of CTCF at binding regions is associated with local hypomethylation, and protection against hypermethylation, which is lost when these binding sequences are disrupted (Stadler et al., 2011; Damaschke et al., 2020). As hypermethylation is a hallmark of the tumour methylome, I asked if there was any enrichment for cis-methQTL-CpGs in CTCF sites. I used the CTCF chromHMM annotations from Taberlay et al. (2014) for both the MCF7 and HMEC cell lines, (which are based on ChIP-seq data). There was strong enrichment in CTCF sites in cis-methQTL-CpGs across both MCF7 (OR = 1.35; P = 4x10^{-4}) and HMEC (OR = 1.26; P = 1x10^{-2}) cell lines (Figure 31). There was also significant depletion in CTCF sites that overlapped with a promoter (but not enhancer) histone mark profiles for MCF7 (OR = 0.83; P = 2x10^{-2}) and HMEC (OR = 0.83; P = 2x10^{-2}) cell lines. Thus, cis-methQTL-CpGs are proportionally more abundant at CTCF sites in breast tumours, but not at CTCF sites within promoters.

5.7 – Cis-methQTL-CpGs are depleted in bivalent promoters in breast tumours

Bivalent promoters are often developmental genes that are primed with both active and repressive histone marks, and which are primarily regulated by polycomb repressive complexes (Blanco et al., 2020). However, polycomb-marked bivalent promoters have a
mutually exclusive relationship with DNA methylation, and the abnormal tumour methylome has been shown to replace polycomb-mediated repression of bivalent promoters with repression by DNA methylation (Hahn et al., 2014). I therefore asked if cis-methQTL-CpGs are enriched in these bivalent promoters in the altered tumour methylome. I used the Taberlay chromHMM annotations for MCF7 and HMEC whereby bivalent promoters were marked with both H3K4me3 and H3K27me3. There was a significant depletion of cis-methQTL-CpGs in both the MCF7 (OR = 0.26; P = 2x10^{-14}) and HMEC (OR = 0.32; P = 2x10^{-24}) cell lines (Figure 31). As such, cis-methQTL-CpGs are less abundant at bivalent promoters than expected.

### 5.8 – A minority of breast tumour cis-methQTL-SNPs also associate with gene expression

Finally, cis-methQTL-SNPs may also associate with local gene activity if the associated CpG is located in a gene promoter/enhancer and these methylation changes affect regulation of gene transcription. To investigate this overlap in breast tumours, I asked if any cis-methQTL-SNPs are in cis-acting expression QTLs (cis-eQTLs) in normal breast tissue. While there may be non-tissue-specific eQTLs I could compare with, I focussed on breast tumour eQTLs which can be related to breast tissue biology. I compared the 13,195 methQTLs in breast tumours against 1,441,05 cis-eQTLs reported in normal breast tissue from the GtEX repository (Carithers et al., 2015) of expression and genotype data (Section 2.1). There were 485 breast tumour cis-methQTLs which were also reported to be cis-eQTLs (Figure 34). This meant that 12,710 had no significant association with normal breast tissue gene expression. Thus, a minority of cis-methQTLs associate locally with both CpG methylation and gene expression in across both normal and tumour breast tissues.
Overlapping QTLs between breast tumour cis-methQTLs and normal blood cis-eQTLs

Normal breast *cis*-eQTLs (GtEX)  
(n = 1,441,105)

Breast tumour *cis*-methQTLs  
(n = 10,500)

454  
Shared QTL loci

Figure 34 - Breast tumour *cis*-methQTL-SNPs share loci with normal breast expression *cis*-QTLs

Venn diagram showing the overlapping loci between breast tumour methylation QTLs (yellow) and normal breast *cis*-expression QTLs (green). N is the total number of *cis*-QTLs in each dataset. Shared QTL number is the subset which overlap.
5.9 – Discussion

By analysing cis associations between genetic variation and local methylation patterns in breast tumours, I have analysed the genomic distribution and functional context of cis-methQTLs. I have also related these findings to the mechanisms which may drive these tumour changes, and compared how these differ from the mechanisms reported for normal tissue cis associations. Cis-methQTL-CpGs are similarly depleted in both CpG islands and promoters in normal and tumour tissues from the Illumina ENCODE-derived annotations, but while they are not enriched in enhancers in tumours (as was observed in normal tissue), there is significant depletion in enhancers annotated from breast tumour cell chromatin marks. Moreover, while cis-methQTL-CpGs are depleted in general tumour PMD annotations, cis-methQTL-CpGs are enriched in poorly conserved PMDs and depleted in highly conserved PMDs across breast tumours. Finally, cis-associating methylation also reveals the relationship between features of the tumour methylation and other molecular mechanisms associated with tumour progression, including enrichment in CTCF sites, depletion in bivalent promoters and a minority overlap with cis-eQTLs. Taken together, these results show that the cis-methQTL-CpG profile described in normal blood is not entirely consistent with cis-methQTLs in breast tumours, as methQTLs are distributed differently across genomic features in both genomic-derived or chromatin-derived annotations.

The overrepresentation of normal blood cis-methQTLs in enhancers was notably absent from breast tumour cis-methQTLs, which were instead underrepresented in chromHMM-defined MCF7 enhancers. Enhancers have been shown to be more differentially methylated between normal and tumour tissue than the relatively conserved low-variance methylation observed in promoters or CpG islands, and the majority (63%) of these differentially methylated enhancers are hypomethylated in tumours, which is thought to promote active chromatin and cancer progression (Bell et al., 2016). However, there are also technical limitations which may explain this depletion, including the low enhancer coverage bias of the Illumina 450k array. However, this issue will be a less significant in studies which primarily used the more extensive Illumina EPIC array (which has ~350,000
additional enhancer probes) (Moran, Arribas and Esteller, 2016). Another possible technical explanation of enhancer depletion could also be that H3K4me1 marks could hypothetically be lost from the chromatin in breast tumour tissue, as these marks determine chromatin-based enhancer annotations; an epigenetic redistribution and overall reduction of H3K4me1 has been shown during ageing and targets cancer/development-associated pathways in hematopoietic stem cells and which consequently may predispose to acute-myeloid leukaemia (Adelman et al., 2019).

CpG islands and promoters were similarly depleted across normal and breast tumour cis-methQTL-CpGs. Promoters and CpG islands have a typically conserved pattern of constitutive hypomethylation in normal tissues to conserve gene transcription in active genes (Long et al., 2013). This is one explanation for why CpG islands are already generally depleted for cis-methQTL-CpGs in such promoters in normal blood tissue (Min et al., 2021). However, another explanation for the promoter and enhancer depletion could be the consequences of the CGI hypermethylation observed in tumours (Nishiyama and Nakanishi, 2021). For example, if methQTLs rely on an allele-associated change in methylation state for a CpG, then hypermethylation at these CpG islands may disrupt CpG methylation variance and these associations would therefore appear less often than expected in tumour CpG islands. Indeed, for the small proportion of CGIs/promoters which are differentially methylated between normal and tumour tissue, 63% of differentially methylated promoters and 94% of differentially methylated CGIs showed hypermethylation and silenced (Bell et al., 2016). Regardless of whether low methylation variance is due to the hypomethylation in normal tissues or the hypermethylation in tumours, low CGI methylation variance would therefore align with low allele-dependent methylation patterns.

This suggests that cis-associated promoter depletion could be related to CGI hypermethylation in tumours. Indeed, promoter depletion was even greater in bivalent promoters of heterochromatic polycomb-repressed developmental genes in breast tumour cis-methQTL-CpGs, which are characterised by both active and repressive histone marks (Blanco et al., 2020). Tumour CGI hypermethylation occurs primarily in cell-cycle/developmental gene promoters which are associated with polycomb-mediated repression, which is typically mutually exclusive with DNA methylation as two separate
repressive mechanisms (Easwaran et al., 2012). However, polycomb normally promotes and maintains hypomethylation in normal tissue through the TET family of demethylation proteins (Li et al., 2018). As such, one explanation may be that hypermethylation at these promoters in tumours could impede the PRC2-driven deposition of the H3K27me3 repressive histone mark. This would subsequently manifest as depletion in cis-methQTL-CpGs in the chromHMM-based annotation (Figure 35).

However, an alternative explanation may be that depletion may indicate hypomethylation at these bivalent promoters in tumours as well as in normal tissue. Indeed, in colorectal cancer, these promoters were susceptible to hypermethylation when both H3K27me3 and H3K4me3 marks were lost whilst loss of H3K27me3 alone was shown to correlate with hypomethylation of these promoters (considered a manner of gene activation in cancer) (Hahn et al., 2014). After studying H3K27me3 redistribution upon global hypomethylation, Duncan et al showed that a high ratio of H3K27me3 to H3K4me3 at bivalent promoters is positively associated with canonical polycomb-repression, negatively associated with both DNA methylation and transcriptional activity and that these bivalent promoters are highly susceptible to hypermethylation in breast tumour (Dunican et al., 2020). As such, both instances of hypomethylation and hypermethylation can be observed in bivalent promoters in cancer; while the former may drive cancer progression through gene activation, the latter likely has no effect on gene expression when DNA methylation replaces polycomb as a more stable form of repression.

Relative to other enrichments seen in this analysis however, these conclusions are supported by depletion in regions with repressive histone marks (which more closely represents facultative heterochromatin where these developmental genes are found) but not in the heterochromatin annotation (which more closely resembles constitutive heterochromatin) (Penagos-Puig and Furlan-Magaril, 2020). However, this chromHMM heterochromatin annotation is limited as H3K9me3 ChIP-seq data, the primary defining mark of constitutive heterochromatin, is not used for classification but rather heterochromatin is classed in the Taberlay study by the absence of chromatin mark profiles which define other genomic features (Taberlay et al., 2014).
Figure 35 – Bivalent promoters may undergo stable DNA methylation through CGI hypermethylation or oncogene activation through CGI hypomethylation in tumours

Shows model mechanism for how tumour-associated CGI promoter methylation changes could affect repression of bivalent promoters. **A)** Natural hypomethylation in bivalent promoters means that they are repressed by polycomb which is targeted to H3K27me3. **B)** Tumour-associated hypomethylation causes redistribution of H3K27me3 away from bivalent promoters and accumulation of H3K4me3, which activates the gene. **C)** Tumour-associated hypermethylation of bivalent promoters which have lost both H3K4me3 and H3K27me3 marks interrupts polycomb-mediated repression by inhibiting deposition of H3K27me3, and taking over as the stable repressive mechanisms at these promoters.
Instead, cis-methQTLs appeared more often than expected in heterochromatin. Since tumour heterochromatin is associated with high PMD occurrence, PMD enrichment might have also been expected, but cis-methQTL-CpGs are depleted in the heterochromatin annotation. However, this may be explained by a significant caveat to the methylation array used; Infinium arrays were designed to primarily study gene expression, and thus there is CpG probe bias towards genic regions (Pidsley et al., 2016). Thus, intergenic PMDs are generally underrepresented in such arrays. While this limitation could be addressed with the improved genome coverage that whole genome bisulfite sequencing may offer, this approach still has challenges beyond affordability. Amongst the technical bias and regions which are difficult to sequence, sequencing depth must balance costs with accuracy and has been recently proposed to require 100x coverage to match the precision achieved with methylation arrays (Zhou et al., 2019). However, cis-methQTLs also may also be simply less abundant outside of active CpG islands where this mechanism of local reprogramming of methylation by transcription factors may be common.

Interestingly, there was also an enrichment for cis-methQTL-CpGs in non-promoter CTCF sites. Studies have established a relationship between CTCF binding sites and DNA methylation in normal tissue whereby CTCF occupation of these sites protects against local sequence hypermethylation (Damaschke et al., 2020). Indeed, loss of CTCF in tumours drives hypermethylation at differentially methylated promoter-CTCF sites in cancer-associated genes (Figure 36) (Damaschke et al., 2020; Segueni and Noordermeer, 2022). However, this mechanism does not explain local cis-methQTLs where there is an allele-dependent association with methylation change in a CTCF site, which may be more likely to be explained by SNPs which disrupt CTCF binding sites. While these enrichments are for CpGs rather than the QTLs, they still are a reliable assessment of methylation change around CTCF sites, but this analysis falls short of asking specifically if any cis-methQTL-SNPs or CpGs are directly disrupting the binding site for other established CTCF annotations. As a preliminary analysis, I checked how many SNPs or CpGs had a significant cis- or trans- methQTL which overlapped with an CTCF region in mammary epithelial tissue from ENCODE (ENCODE Project Consortium, 2012; Experiment: ENCSR697YIN); there were 56 SNPs and 333 CpGs which overlapped these CTCF regions. Furthermore, a similar analysis showed that 68 SNPs...
Figure 36 – CTCF local protection against hypermethylation is lost when CTCF binding region is disrupted by genetic variant

A) CTCF occupancy of binding region represses methylation enzymes (eg. DNA methyltransferases, or DNMTs), either sterically or through other proteins which interact with DNMTs (eg. PARP-1), when the SNP allele matches the CTCF binding motif and thus protects local CpGs against hypermethylation. B) In the absence of CTCF when the alternative SNP allele disrupts the binding motif, methylation proteins can bind and methylate local CpGs.
and 96 CpGs were overlapping with CTCF binding motifs (from JASPAR; Mathelier et al., 2016). As such, CTCF binding sites may be disrupted by SNPs and local CpG methylation patterns are consequently being altered in tumours. Future work could elaborate further on this mechanism by investigating if, for a given genetic variant in the CTCF binding region from the breast tumour cis-methQTLs, there was simultaneous CTCF occupancy and hypomethylation which exclusively associated with one allele (i.e. no occupancy and hypermethylation associated with the other allele).

Finally, some tumour cis-methQTLs in breast tumours were also eQTLs in normal blood. These examples support a mechanism whereby a cis-methQTL is commonly in a binding region for a transcription factor, but which then may subsequently affect local CpG methylation in the regulatory elements of a gene and therefore may possibly also associate with expression as a cis-eQTL (Pierce et al., 2018; Zhao et al., 2019). On the other hand, the majority of tumour cis-methQTLs which do not overlap must therefore affect CpG methylation without affecting gene expression. These likely follow the more canonical model of variants within protein binding sequences that disrupt the activity of transcription factors or chromatin regulators, such as the mechanism discussed for CTCF (Damaschke et al., 2020). However, as these are normal blood eQTLs overlapping with breast tumour cis-methQTLs, there will also inevitably be a cell-specific difference between expression and methylation patterns in these tissue types. An improved analysis could therefore be to examine the overlap between methQTLs and eQTLs which are both tumour-specific, as performed by Houlahan et al. (2019) in a methQTL analysis in prostate cancer.

Overall, one important caveat in an analysis of methQTL enrichment in genomic regions is that the CpGs are limited to the Illumina 450k microarray which has substantial biases. For example, whilst the 450k array was designed for higher coverage of promoters and genes, it has sparse coverage of distal regulatory elements (i.e. enhancers) and indeed with intergenic regions in general (Morris and Beck, 2015). There have been improvements addressing this lack of enhancer coverage, such as the Illumina EPIC (850k) array which added ~350,000 CpG sites for enhancers on top of most of the 450k array probes, but even this array panel still only represents a small portion of CpGs sites in the genome (Moran, Arribas and Esteller, 2016). Therefore, as the most recent EPIC array has taken steps to
address the coverage bias, they still offer a more affordable option which is comparable in both accuracy and genomic element coverage when compared with alternative approaches to measuring CpG methylation (Heiss et al., 2019), such as whole-genome (McClay et al., 2015; Perzel Mandell et al., 2021) or reduced-representation (Fan et al., 2019; Szymczak et al., 2020) bisulfite sequencing. As such, Illumina’s 450k arrays (McRae et al., 2018) and EPIC arrays (Hannon, Gorrie-Stone, et al., 2018; Min et al., 2021; Oliva et al., 2022) are currently common choices for methQTL analyses, although this study would perhaps have benefited from the additional enhancer probes from the EPIC array if it were repeated. However, one strength of the enrichment analyses which addresses this limitation is that the odds scores are calculated against the proportion of CpGs absent/present in a given annotation in the array background, rather than genome wide, and so this bias is accounted for.

Taken together, these results show that cis acting methQTL-SNPs associate with CpGs in enhancers and poorly conserved PMDs in breast cancer, and may disrupt protein binding sites and repression systems around developmental genes. Overall, this shows that the functional context of cis-methQTL SNPs/CpGs in breast tumours show similarities with those of cis-methQTLs in normal blood tissues, but suggests SNP-associated changes in breast tumour methylation patterns also show some distinct differences.
Chapter 6 – The functional context of trans-methQTLs in breast tumours show major similarities with normal tissue

6.1 – Introduction

While our understanding of local cis interactions between genetic variation and methylation patterns is relatively comprehensive, and their mechanisms well studied, long-range trans associations have been more difficult to characterise due to the relatively indirect mechanisms driving these associations and by the limited power to detect these associations in typically smaller cohort sizes. However, characterisations of trans associations and their potential mechanisms would reveal if there is a predominant mechanism (as in cis associations) or indeed a range of mechanisms driving these associations. Using a similar analysis plan for the characterisation of cis-methQTLs in the previous chapter, I therefore first asked how methylation at trans-methQTL associated CpGs were distributed across various functional genomic elements (such as CpG islands, promoters, enhancers, CTCF sites, and heterochromatin). I then asked if this distribution of trans-associated methylation alterations were different from those in normal blood trans-methQTLs. Finally, I asked what this distribution of trans-methQTLs across genomic elements suggests about the mechanisms underpinning these long-range associations. Together, these analyses will develop our understanding of breast tumour methylation changes that, like many non-blood tissues, have been thus far relatively more limited in comparison with cis association studies.

In normal blood, trans-methQTL-SNPs have been found primarily in inactive regions, such as heterochromatin (Villicaña and Bell, 2021). In contrast, trans-methQTL-CpGs are
enriched in CGIs and promoters, but less so in enhancers. From these observations, the simplest model for trans-methQTL mechanisms has been that trans-methQTLs are primarily cis-eQTLs in an inactive region of the genome for a transcription factors with long-range effects on CpG methylation - specifically in promoter CGIs for other genes (Bonder et al., 2017; Huan et al., 2019). For example, Bonder et al. (2019) found that SNPs with multiple trans-methQTL CpG correlations often had cis effects on the occupancy of transcription factors, such as for NFKB1 and CTCF, which affected methylation consistently at various CpG sites across the genome; One SNP had 380 trans CpG effects which associated with decreased DNA methylation (38% of which were enriched for the NFKB1 binding site) and increased cis NFKB1 gene expression. They also noted a SNP with 779 trans CpG effects which associated with increased DNA methylation and reduced cis CTCF occupancy. Additionally, increased promoter methylation of breast cancer-associated TF Oestrogen Receptor 1 (ER1) has been shown to correlate with tumour progression (Wei et al., 2012). Furthermore, a recent study in cervical cancer showed that ten different cervical tumour risk genes had high promoter methylation and correlated with tumour growth has promising biomarker potential (El Aliani et al., 2021).

There are some alternative models which have been less well studied, such as trans-methQTL mechanisms involving polygenicity or 3D chromatin organisation/looping bringing distal regions in close proximity and consequently enabling typically cis-methQTL mechanisms (Rao et al., 2014; Gaunt et al., 2016). As with cis-methQTLs however, these observations are primarily in normal blood tissues, and few studies have used trans-methQTLs to interrogate the diversity of long-range mechanisms in breast tumours.

6.2 – SNP genotypes associate with both distal and interchromosomal CpG methylation in breast tumour trans-methQTLs

I proceeded by first profiling the subset of breast tumour methQTLs which were associating in trans with CpG methylation. Any methQTL where the SNP-CpG distance was greater than 1Mb, either on the same or a different chromosome, was considered a trans-
methQTL. Of the 13,195 methQTLs I identified in breast tumours, 2,695 were trans-methQTLs. These comprised 1,139 unique CpGs showing significant associations with 1,712 unique independent SNPs. Furthermore, of these trans-associated unique CpGs, 570 were interchromosomal associations.

I selected a single trans-methQTL-CpG for further characterisation to observe the allele-dependent association with methylation. I chose the trans-methQTL-CpG cg00160777 (on chromosome 16) to validate in an individual CpG methQTL analysis, which was part of a six CpG panel I had examined in chapter 3. I used a Manhattan plot to visualise SNPs which associated with cg00160777. There was a significant peak (P < 5x10^-8) of trans-signals on chromosome 4. The variant which showed the strongest methylation association (P = 5x10^-70) was rs2008600 (Figure 37A-B).

I then verified that the mean methylation of cg00160777 had a linear correlation with the genotype of rs2008600 in a dosage dependent relationship with the alternative allele. I found a clear loss in mean methylation at cg00160777 associated with the minor allele (C) of rs2008600 (Figure 37C). This confirmed that genetic variation is associating with distal intrachromosomal and interchromosomal CpGs in breast tumours in an allele-dependent manner.

Variants in trans-methQTLs in normal tissue are often in intergenic or repressed regions, whereby any mediation of methylation or gene expression is likely to be through long-range or complex regulatory mechanisms (Min et al., 2021). As such, I evaluated the genomic profile of rs2008600 to see if this aligned with trans-methQTLs characteristics described in normal tissue. I used LDProxy to analyse the regulatory potential and genomic location of rs2008600 on chromosome 4; rs2008600 was assigned the lowest regulatory potential by RegulomeDB and was located within the intron of CCSER1 (Figure 38). CCSER1 overexpression causes proliferation and it has been a source of chromosomal instability when deleted or mutated (Patel et al., 2013; Santoliquido et al., 2021). Interestingly however, this does not fit the Min characterisation of trans-methQTL-SNPs in normal blood, which were depleted in introns (Min et al., 2021).
**Figure 37** - rs2008600 is a significant *trans*-methQTL for methylation at cg00160777

**A**) Manhattan Plot of SNPs across genome which show significant association with methylation at cg00160777. Shows peak of genetically linked SNPs in chromosome 4 with the CpG on chromosome 16. Red/blue lines mark nominal significance thresholds ($5 \times 10^{-5}$/$5 \times 10^{-8}$). **B**) QQ-plot showing inflated significance after $5 \times 10^{-4}$, representing the SNP peak in chromosome 4. Only SNPs with MAF > 5% are shown. **C**) Boxplot of rs2008600 genotype associating with differential methylation at cg00160777 ($P = 5 \times 10^{-70}$). **D**) Manhattan Plot of only significant SNPs from chromosome 4 peak. This represents multiple SNPs within high genetic linkage in a methQTL. Red line is significance ($P < 5 \times 10^{-14}$).
Figure 38 – Haploblock of genetically linked SNPs around rs2008600 colocalises with CCSER1 intron

A) SNPs from the haploblock are plotted by their chromosome position and their linkage disequilibrium ($R^2$) with the lead query variant rs2008600 (blue) in the top panel. SNPs are coloured by coding (red) and non-coding (orange) variants. High (1) or low (7) regulatory potential indicated numerically and according to SNP annotations from RegulomeDB. Grey lines/peaks represent the combined recombination rate at chromosome positions. B) Genes and their chromosomal positions are shown.
To account for the potential confounding effect of significant associations with multiple neighbouring SNPs (due to genetic linkage), I also extended this analysis to the surrounding SNPs which showed high LD ($R^2 \sim 1$) with rs2008600. The 10 surrounding SNPs which showed the highest linkage with rs2008600 also showed low regulatory potential (Figure 39, Table 7). Overall, this example suggests that breast tumour trans-methQTL-SNPs are not necessarily in intergenic inactive regions, but are unlikely to have a role in transcriptional regulation.

6.3 – Trans-methQTL-CpGs are enriched in CpG islands and promoters but depleted in enhancers in breast tumours

In normal blood, trans-methQTL-CpGs are enriched in regions of low methylation, such as promoters and CpG islands, but depleted in enhancers. I asked if trans-methQTL-CpGs in tumours had a similar genomic distribution. I first overlapped the trans-methQTL-CpGs coordinates with genomic annotations for functional elements, and then calculated an enrichment or depletion as an Odds ratio score, using a Fisher’s test. I used the Infinium 450k array annotations (as used in the cis-methQTL section 5.3) for regulatory elements (enhancers, promoters) and CpG features (CGIs, shores and shelves) based on data from ENCODE project and UCSC genome browser. There was trans-methQTL enrichment in promoters (OR = 1.17; $P = 0.03$), CpG islands (OR = 1.31; $P = 1 \times 10^{-5}$) and CpG shelves (OR = 1.43; $1 \times 10^{-4}$), but depletion in enhancer regions (OR = 0.53; $1 \times 10^{-14}$) and CpG shores (OR = 0.83; 0.01) (Figure 40). These enrichments/depletions were also very similar to the patterns of enrichment/depletion that have been reported in normal blood trans-methQTL-CpGs (Min et al., 2021). Overall, this analysis shows that trans-methQTL-CpGs are similarly depleted in enhancers and enriched in CpG islands in both breast tumours and normal blood.
Figure 39 – SNPs with highest genetic linkage within rs2008600 haploblock have low regulatory potential

SNPs from the haploblock are plotted by their chromosome position and their linkage disequilibrium ($R^2$) with the lead query variant rs2008600 (blue) in the top panel. SNPs are coloured by coding (red) and non-coding (orange) variants. High (1) or low (7) regulatory potential indicated numerically and according to SNP annotations from RegulomeDB.
### Table 7 – Regulatory potential of SNPs with the highest R² within LD block of rs2008600

Table showing variant annotations for top ten variants with highest R² with rs2008600. Table ordered by R² and distance from rs2008600. Regulatory potential is represented as numbers (from 1-7) representing the extent of data evidence for a SNP having a regulatory role (1 is high evidence and 7 is low evidence). Column names are: SNP rsIDs; Chromosome number; Chromosome position on the Human Genome Build 38; Allele complement; Minor allele frequency; Distance from rs2008600; D’ is a linkage measurement similar to R²; R² correlation; Correlation patterns observed between the alleles of these 2 SNPs; Regulatory potential according to the database FORGE DB; Regulatory potential according to the database Regulome DB.

<table>
<thead>
<tr>
<th>RS Number</th>
<th>Chr</th>
<th>Position (GRCh38)</th>
<th>Alleles</th>
<th>MAF</th>
<th>Distance</th>
<th>D’</th>
<th>R²</th>
<th>Correlated Allele</th>
<th>FORGEdb</th>
<th>RegulomeDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs72887446</td>
<td>4</td>
<td>90590817</td>
<td>(G/A)</td>
<td>0.3132</td>
<td>11484</td>
<td>1.0</td>
<td>1.0</td>
<td>A=G,G=A</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>rs72887445</td>
<td>4</td>
<td>90590534</td>
<td>(C/T)</td>
<td>0.3132</td>
<td>11401</td>
<td>1.0</td>
<td>1.0</td>
<td>A=C,G=T</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>rs28591382</td>
<td>4</td>
<td>90589810</td>
<td>(A/T)</td>
<td>0.3132</td>
<td>10677</td>
<td>1.0</td>
<td>1.0</td>
<td>A=A,G=T</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>rs28523954</td>
<td>4</td>
<td>90579665</td>
<td>(C/A)</td>
<td>0.3132</td>
<td>532</td>
<td>1.0</td>
<td>1.0</td>
<td>A=C,G=A</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>rs11734079</td>
<td>4</td>
<td>90564383</td>
<td>(A/C)</td>
<td>0.3187</td>
<td>14750</td>
<td>1.0</td>
<td>0.9749</td>
<td>A=A,G=C</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>rs77297031</td>
<td>4</td>
<td>90564501</td>
<td>(T/A)</td>
<td>0.3187</td>
<td>14632</td>
<td>1.0</td>
<td>0.9749</td>
<td>A=T,G=A</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>rs10006668</td>
<td>4</td>
<td>90565954</td>
<td>(G/A)</td>
<td>0.3187</td>
<td>13179</td>
<td>1.0</td>
<td>0.9749</td>
<td>A=G,G=A</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>rs2045998</td>
<td>4</td>
<td>90560928</td>
<td>(T/C)</td>
<td>0.3187</td>
<td>12205</td>
<td>1.0</td>
<td>0.9749</td>
<td>A=T,G=C</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>rs58996221</td>
<td>4</td>
<td>90567495</td>
<td>(G/C)</td>
<td>0.3187</td>
<td>11638</td>
<td>1.0</td>
<td>0.9749</td>
<td>A=G,G=C</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>rs5632235</td>
<td>4</td>
<td>90566814</td>
<td>(A/G)</td>
<td>0.3187</td>
<td>10319</td>
<td>1.0</td>
<td>0.9749</td>
<td>A=A,G=G</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>
Figure 40 – Breast tumour trans-methQTL-CpGs are enriched in CpG islands and promoters but depleted in enhancers

Enrichment represented as the odds ratio for methQTL-CpGs shown as dots (including 95% confidence intervals as bars). Significant enrichments/depletions (P < 0.05) denoted with their p-value in bold. Genome annotation taken from Illumina reference genome.
6.4 – Trans-methQTL-CpGs are depleted in chromatin-based annotations for enhancers in breast tumours

I next wanted to validate these genomic features by assessing enrichments in annotations derived from characteristic chromatin states. I overlapped the trans-methQTL-CpG positions with the chromHMM state annotations and used odds ratio (OR) scores (calculated from a Fisher’s test) to identify any enrichments or depletions for trans-methQTL-CpGs in any of these chromatin-derived genomic annotations. I used chromHMM state annotations for both normal breast (HMEC) and breast tumour (MCF7) cell lines that were used in the cis-methQTL-CpG section. Trans-methQTLs showed a significant depletion (OR = 0.76; P = 0.03) in normal HMEC enhancers (Figure 41). In contrast, there was significant enrichment in promoters in MCF7 chromatin (OR = 1.18; P = 0.02). There was also depletion in HMEC repressed regions (OR = 0.5; P = 9x10^{-5}). The most contrasting OR scores between HMEC and MCF7 was for heterochromatin, where there was significant enrichment in normal HMEC (OR = 1.26; P = 1x10^{-4}) but depletion in the tumour MCF7 (OR = 0.79; P = 7x10^{-5}) cell lines. Comparing between the cell lines, the OR scores overall showed strong signals that were not always consistent across both HMEC and MCF7, which contrasts with how the cis-methQTL-CpGs were often concordant.

6.5 – Breast tumour trans-methQTL-CpGs are depleted in breast-specific tumour PMDs

The CpGs in normal blood trans-methQTLs have been found more abundantly in CGIs and enhancers, where methylation is mainly low in normal tissue (Min et al., 2021). Additionally, low methylation occurs at PMDs and heterochromatin, the latter of which appeared to be enriched in the normal HMEC annotation but depleted in the tumour MCF7 annotation. I asked if trans-methQTL-CpGs were enriched in HMDs and PMDs in breast tumours. I overlapped trans-methQTL-CpGs with HMD/PMD annotation coordinates and used a Fisher’s test to identify significant enrichments in trans-methQTL-CpGs compared with all other CpGs. I used the Zhou et al. (2018) annotation for common tumour
Figure 41 – Breast tumour trans-methQTL-CpGs are enriched in CpG islands and promoters but depleted in enhancers for chromHMM chromatin annotations across tumour and normal mammary tissue

Enrichment represented as the odds ratio for methQTL-CpGs shown as dots (including 95% confidence intervals as bars). Significant enrichments/depletions (p < 0.05) denoted with their p-value in bold. Tumour (red) and normal (blue) annotations are colour coded.
PMD/HMDs described previously. The \textit{trans}-methQTL-CpGs were significantly enriched in HMDs (OR = 1.34; P = 6x10^{-7}) but showed depletion in PMDs (OR = 0.84; P = 3x10^{-2}) (Figure 42A). I then repeated this analysis specifically using breast tumour PMD annotations, as PMDs may be cell-specific. I used an aggregated PMD annotation which included all PMDs identified across 30 different breast tumour subjects by Brinkman \textit{et al.} (2019) (Figure 42B). Indeed, this annotation also showed a significant depletion for \textit{trans}-methQTL-CpGs in PMDs (OR = 0.73; P = 2x10^{-7}).

This method of merging PMDs does not assess the proportional occurrence of CpGs in PMDs, and the possibility that CpGs may occur in relatively well or poorly conserved PMDs. Therefore, I finally investigated the number of PMDs which each \textit{trans}-methQTL-CpG occurs in against non-methQTL-CpGs, which I termed the PMD occurrence. I analysed the PMD occurrence for \textit{trans}-methQTLs-CpGs (as described in the \textit{cis}-methQTL analysis; Section 5.5). For almost all (25 out of 30) of the PMD occurrences (between 0 and 30 samples), \textit{trans}-methQTL-CpGs were more depleted in PMDs than non-methQTL-CpGs (Figure 43A). The distributions followed a similar shape but were significantly different by means of a chi-squared test to the array background CpGs ($\chi^2 = 69.5; P = 6x10^{-5}$) (Figure 43B). Thus, a significant depletion is observed regardless of whether PMDs are well or poorly conserved.

6.6 – \textit{Trans}-methQTL-CpGs share some similar mechanisms of methylation change with \textit{cis}-methQTLs, through CTCF enrichment, bivalent promoter depletion and eQTL overlap

An allele-dependent association with hypermethylation at CTCF sites would be expected to be a less prominent mechanism for \textit{trans}-methQTL associated methylation changes since CTCF disruption is primarily associated with local methylation gain rather than distal (Damaschke \textit{et al.}, 2020). I therefore asked if there were any comparable CTCF enrichments observed in both \textit{cis}- and \textit{trans}-methQTLs. The breast tumour \textit{trans}-methQTL-CpGs showed a significant enrichment in CTCF sites in MCF7 chromatin.
Figure 42 – Breast tumour trans-methQTL-CpGs are depleted in PMDs but enriched in HMDs in common tumour annotations

Enrichment represented as the odds ratio for trans-methQTL-CpGs shown as dots (including 95% confidence intervals as bars). Significant enrichments/depletions (p < 0.05) denoted with their p-value in bold. Brinkman annotations described in breast tumours. **A) Zhou annotations** are PMDs/HMDs which are common across tumours. **B) Brinkman annotations** are PMDs described in breast tumours.
Figure 43 – *Trans*-methQTL-CpGs are relatively depleted in both well and poorly conserved PMDs across 30 breast tumour samples compared with non-methQTL-CpGs

Shows the distribution of the frequency by which CpGs occur in a PMD across 30 breast tumour samples (A) and also a more focussed look excluding a PMD occurrence of zero (B). Blue represents *trans*-methQTL-CpGs. Black represents non-methQTL-CpGs which are on the microarray background. Statistically significant difference between distributions calculated with chi-squared ($\chi^2$); $P < 0.05$. 
(OR = 1.49; P = 1 \times 10^{-2})$, but not in the normal breast HMEC chromatin (Figure 41). This was a modest enrichment in comparison with the cis-methQTL enrichment, which was significant in both cell lines, but does show that a CTCF-based mechanism of methylation change does show some consistency (albeit less significant) between trans-methQTLs and cis-methQTLs.

Furthermore, as trans-methQTLs are shown to associate with methylation changes in inactive regions in normal blood, I also asked if the polycomb-based repression of bivalent promoters observed in cis-methQTLs may also be a mechanism influencing trans-methQTLs. Indeed, there was significant depletion in bivalent promoters in both HMEC (OR = 0.4; P = 2 \times 10^{-5}) and MCF7 (OR = 0.32; P = 1 \times 10^{-3}) (Figure 41), suggesting that this mechanism also has a common role in both trans methylation associations as well as cis associations.

It has also been proposed that trans-methQTLs may be cis-eQTLs for genes with a role in transcriptional regulation. I therefore asked if any breast tumour trans-methQTLs also associated with local gene expression (termed eQTLs) in normal breast tissue. I compared against 1,441,105 cis-eQTLs reported in normal breast tissue (from GtEX). There were 85 breast tumour trans-methQTLs which also associated with expression and thus may be involved in gene regulation in normal breast tissue (Figure 44). This therefore suggests that 2,610 trans-methQTLs do not associate with QTL-dependent gene expression changes observed in normal tissue. Overall, there are some key mechanisms of methylation change which are common between cis- and trans-methQTLs in breast tumours.

6.7 – Discussion

I have analysed long-range genetic associations with methylation changes at distal CpGs, and used these results to infer the trans mechanisms that may be driving these changes in breast tumours. Breast tumour trans-methQTL-CpGs are enriched in promoters and CpG islands, but not in enhancers, which aligns with the profile of trans-methQTLs described in normal blood trans-methQTLs (Min et al., 2021). However, these observations show inconsistencies across chromatin-derived genomic annotations from both normal and tumour breast cell lines. Furthermore, these trans-methQTLs are depleted in both common
Figure 44 – Breast tumour trans-methQTL-SNPs share loci with normal breast expression cis-eQTLs

Venn diagram showing the overlapping loci between breast tumour trans-methylation QTLs (blue) and normal breast cis-expression QTLs (green). N is the total number of QTLs in each dataset. Shared QTL number is the subset which overlap.
tumour PMDs and breast tumour PMDs. Finally, *trans*-methQTLs have some mechanisms for methylation alterations which are consistent with *cis*-methQTLs (such as CTCF enrichment, bivalent promoter depletion and eQTL overlap) which highlights features of breast tumour methylation changes that are common to all methQTLs regardless of the genomic position of the corresponding CpGs.

Overall, low sample size remains the primary challenge in identifying *trans* associations between genetic variation and methylation changes as this limits how reliably the interpretation of the results of this study can be discussed in the wider context of literature. Indeed, this study is relatively underpowered to detect *trans* associations with a sample size of 333 subjects, compared with some of the largest methQTL studies which have to analysed cohorts of as many as ~4000 individuals in whole blood (Bonder et al., 2017; Huan et al., 2019). For this reason, the normal blood methQTL functional profiles I compared against were taken from a meta-analysis study which had a substantially larger cohort size (N = 32,851) and greater power to detect *trans* associations than any other individual study could manage (Min et al., 2021). However, beyond the large cohorts of whole blood, studies have been able to analyse *trans* associations using smaller sample sizes (typically below 1000 samples); some of these have been a comparable sample size to this study (below 500 samples), and found similar findings to the larger studies (Shi et al., 2014; Scherer et al., 2021; Zhang et al., 2021). However, these smaller sample sizes are typically limited due to these studies analysing non-blood tissue-specific samples. Therefore, tissue-specific *trans*-methQTL analyses can be made with lower sample sizes as long as this limitation is acknowledged. As a consequence of these limitations, there is therefore a limited number of *trans*-methQTL studies to compare against and to contextualise these findings; the few that do almost all focus on normal tissue, and fewer still have aimed to characterise the mechanisms of *trans*-methQTLs.

In breast tumours, *trans*-methQTL associated CpGs are enriched in CGIs and promoters but depleted in enhancers from the Illumina *ENCOD/UCSC* annotations, and these patterns are further observed in chromatin-derived *chromHMM* annotations for normal breast HMEC cells (promoter enrichment) and breast tumour MCF7 cells (enhancer depletion). This aligns with the hypothesis that the most common *trans*-mechanism is
where trans-methQTL-CpGs are primarily cis-eQTLs (most likely enhancers in inactive intergenic regions) associated with transcription or DNA binding factor which subsequently alters (either directly or indirectly) CpG methylation at a distal promoter region (Bonder et al., 2017; Huan et al., 2019). The breast tumour trans-methQTL genomic profile is therefore similar to the normal blood trans-methQTL genomic profile (Min et al., 2021). This suggests that the cis-eQTL/trans-methQTL mechanism for long-range CpG associations is likely to be independent of tumour or normal blood tissue. This is in contrast to what has been examined in cis-associations, which showed some unique differences in the tumour genomic profile, suggesting that trans mechanisms underpinning tumour methylation changes are more conserved across normal/tumour tissues than the cis mechanisms are. Indeed, methQTL-CpGs in CGIs may not be expected to differ greatly between normal and tumour breast tissues as methylation at CGIs has low variance regardless of tumour-associated hypermethylation or normal-associated constitutive hypomethylation. Furthermore, an explanation for the dissimilar OR scores for trans-methQTL chromatin-derived features could be that histone marks used to profile these promoters/enhancers are lost in tumours, as has been reported (Adelman et al., 2019).

The trans-methQTL-CpGs were significantly depleted in PMDs, both in annotations for breast cancer and common tumour PMDs. Given PMDs are found mostly in inactive heterochromatin, this PMD depletion is further supported by trans-methQTL-CpG depletion in chromHMM annotations for breast tumour MCF7 heterochromatin. As trans-methQTL-CpGs are similarly found in active regions (specifically in promoter CGIs) in breast tumour as in normal blood, a simple explanation is that heterochromatin and PMDs are gene poor and so there may be less trans associated methylation changes in these regions. This would be further supported by the contrasting enrichment observed in common tumour HMDs (i.e. regions which were never within a PMD across multiple, different cancer types). Furthermore, this proposed bias against identifying trans-methQTLs in PMD-rich inactive regions would be exaggerated as CpGs from the Infinium arrays are also typically gene poor and thus already have limited coverage in these PMD regions.

However, while trans-methQTL-CpGs were depleted in the breast tumour cell line (MCF7) heterochromatin, they were enriched in normal breast (HMEC) heterochromatin.
This difference between MCF7 and HMEC may be a consequence of the tumour-associated global hypomethylation that forms PMDs primarily in heterochromatin, whereby large sections of reduced methylation (at least a 70% reduction) may eliminate the allele-dependent CpG methylation associations and thus appear as a depletion in normal chromatin and enrichment in tumour chromatin (Lister et al., 2009; Schroeder and LaSalle, 2013). Moreover, as cis- and trans-methQTL-CpGs are associated with active and inactive regions respectively, this loss of allele-dependent associations in PMDs may be more prominent in gene-poor/intergenic trans-methQTL-CpGs compared with relatively active/genic cis-methQTL-CpGs which are less affected by tumour hypomethylation. Indeed, this is supported by cis-methQTL-CpGs which were enriched in heterochromatin in both MCF7 and HMEC chromHMM annotations. Thus, trans-methQTLs show PMD depletion and tumour-specific heterochromatin enrichment, with only minor differences from those observations in cis-methQTLs.

There was also a significant enrichment in the MCF7 tumour chromatin for CTCF sites (but not in the normal breast HMEC). This suggests trans-methQTL-SNPs have an association with CpG methylation patterns in CTCF sites in the MCF7 breast tumour cell line. The mechanism proposed to explain cis-methQTL-CpGs enrichment (whereby SNP variants may disrupt CTCF binding sites in an allele-dependent manner) should be expected to have a relatively minor role in the long-range regulation of CpG methylation as it has a primarily local effect on CpG methylation (Damaschke et al., 2020). However, trans-methQTLs may also operate through cis-methQTL mechanisms when 3D chromatin structure has enabled a cis-methQTL to be in proximity with distal CpGs (Do et al., 2017). For example, looping could explain how a disrupted CTCF site could be in close proximity to distal CpGs and thus repress DNMT binding at these long-range loci as well (Rao et al., 2014). As such, allele-associated methylation changes in CTCF sites may act by a similar mechanism in driving both cis and trans genetic associations in breast tumours.

Indeed, a number of these results also suggest some mechanisms underpin methylation changes with both cis and trans associations in some genomic features. The chromatin annotations for bivalent promoters, for example, are similarly depleted in trans-methQTLs they are in cis-methQTLs. The mechanism proposed for depletion in cis-methQTLs
(whereby hypermethylation of polycomb-associated bivalent promoters removes bivalent histone marks) should also hold true for trans-methQTLs as they should both be equally affected by this hypermethylation. Also similar to cis-methQTLs, there were some trans-methQTLs which overlapped with cis-eQTLs in normal blood, which further supports the model that trans-methQTLs may primarily be cis-eQTLs. Another analysis could be to further characterise those overlapping cis-eQTLs with ontological gene information to see if they may have a potential role in regulating methylation patterns. It is clear however that our understanding of both cis and trans mechanisms would equally benefit from the incorporation of additional multiomic data which informs more detailed descriptions of how methylation interacts with other molecular processes in tumours, such as proteomics and breast cancer risk genes.

Finally, there are a few alternative mechanisms that have been proposed for how trans-methQTLs may interact with methylation at distal CpGs. One alternative mechanism is that there may be long-range linkage disequilibrium between cis-methQTLs and distal SNPs. While Long range LD is typically associated with genetic drift, gene interactions or population bottlenecks, it may occur whenever the sample size is sufficiently small and two variants have an extremely rare minor allele frequency, as this may form a spurious and random association rather than an association from a biological relationship (Park, 2019). While the small sample size of this study therefore could be susceptible to such effects, removing SNPs with rare MAFs as a quality control step will have helped to reduce this to some degree. However, long-range LD can be simply addressed by conditioning for the corresponding cis-methQTL and removing trans-methQTLs which are therefore dependent on the cis association (Shi et al., 2014). Another mechanism that has been less explored in literature is that there may be some polygenicity, whereby complex regulatory networks of CpG methylation could explain the low representation of trans genetic variation explained by trans-methQTLs identified (less than 1% compared to over 50% of cis genetic variation explained by cis-methQTLs) and thus the high power required for detection (Gaunt et al., 2016). Overall, there may be more variation in the range of mechanisms underpinning trans-methQTL associations than there is in cis-methQTL associations, and this would be an interesting future direction to continue to develop our understanding of this mechanisms which drive trans-methQTL associations.
Chapter 7 – Discussion and conclusions

7.1 - Main findings of the project

We know that the widespread dysregulation of the methylome in cancer has been detailed, but is still not fully understood. We also know that population methods have been used to investigate SNP-associated methylation changes in normal tissue, but less so in tumour tissue. As such, I utilised methQTLs as a tool to represent the genomic redistribution of genetically-programmed methylation changes in breast cancer. I then used these methQTLs to study and contrast the genomic distribution and functional context of SNP-associated methylation changes between normal breast and breast tumour tissues. I have shown that the genomic and functional profile underpinning SNP-associated methylation changes are relatively consistent between normal breast and breast tumour tissues (more so in trans-methQTLs than in cis-methQTLs). However, the genomic distribution of these methylation changes show some distinct differences in breast cancer.

I identified methQTLs in the TCGA breast tumour data, as has been shown in breast tumour literature (Gong et al., 2019). Gong established an association framework for a cancer methQTL database (across 23 cancer types from The Cancer Genome Atlas), and identified 203,391 cis-methQTLs and 31,118 trans-methQTLs in breast cancer (between 2,721,411 imputed SNPs and 384,084 CpG probes across 664 matched samples). This analytical framework was very similar to the one I established, but additionally included methylation structure as a covariate to address possible batch effects and confounding methylation sub-structure (such as from heterogeneous cell content), a step which was omitted from my analysis. However, as the cis:trans ratio of 13.27% was lower than the 20.42% I observed, the inclusion of methylation structure as a covariate could be a possible explanation for this difference. For example, my analysis could additionally be detecting confounding cell type-specific trans-methQTLs which are actually associating with the different cell compositions of breast tissue samples.
There are not many examples of studies which have identified cis- and trans-methQTLs in cancer and compared the ratio between them, as methQTL studies in cancers have often taken differing approaches and asked different questions. For example, an analysis of tumour-specific methQTLs in prostate cancer selected a limited number of candidate SNPs from prior studies to replicate trans-methQTL identification, rather than using genome-wide imputed SNP genotypes in their associations analysis as I have. Thus, they do not have a cis:trans ratio to compare against (Houlahan et al., 2019). Alternatively, some studies which have identified cis- and trans-methQTLs have done so with incomparable conditions. For example, a study in melanocytes reported that there was 1,497,502 cis-methQTLs and 15,179 trans-methQTLs (a cis:trans ratio of 99.6:1), but they defined trans-methQTLs whereby the SNP-CpG distance was greater than 5Mb, which was a larger than the 1Mb distance threshold I had used to define trans-methQTLs (Zhang et al., 2021). Therefore, it is difficult to directly compare to other cancer studies which may have different analysis conditions, especially as there are not many which have focussed as strongly on trans-methQTL identification as there have for cis-methQTL identification.

Overall, my breast tumour methQTLs showed a different genomic distribution from those observed in normal blood tissue (Min et al., 2021), showing there are relatively more trans-methQTLs than cis-methQTLs in breast tumours. Additionally, there was a greater effect size in cis-methQTLs than in trans-methQTLs, as has been reported in normal tissue (Hannon, Gorrie-Stone, et al., 2018). MethQTLs also showed a respective functional context for their corresponding SNPs and CpGs respectively which has largely been similar between normal and tumour breast tissues. Overall, in both normal and tumour breast tissue methQTLs, cis-methQTLs likely operate with a relatively direct mechanism of SNP-associated disruption of local methylation change while the mechanisms underpinning trans-methQTL methylation change are likely to be more indirect.

These results support the hypothesis that breast tumour cis-methQTLs may largely operate under primarily one mechanism which is shared with cis-methQTLs in normal blood tissue; SNPs in transcription factor binding sites disrupt the binding of factors which are responsible for reprogramming methylation at local CpGs (particularly in local enhancers).
However, breast tumour cis-methQTLs do show some minor changes in their SNP/CpG functional context compared with the normal tissue, such as CpG depletion in enhancers. This depletion agrees with studies showing that enhancers are undergoing tumour-driven gain or loss of methylation that could consequently protect them from local, allele-dependent methylation changes (Batra et al., 2021). I could therefore characterise this further by considering the enhancers which lost cis-methQTLs in tumour tissues, validating that these methylation changes were independent of the cis-methQTL-SNP allele, and finally investigating if these enhancers had any functional association with breast cancer progression which correlated with the methylation state; for example, the enhancer would hypothetically be methylated if it had a role in oncogenic gene activity, or unmethylated if it had a role in tumour suppressor gene activity (Figure 45).

In contrast, breast tumour trans-methQTLs align with the hypothesis that these are indirect, long-range mechanisms, as trans-methQTL-SNPs reside in regulatory regions which associate with the expression of a transcription factor which in turn may mediate SNP-associated methylation change at distal CpGs found in promoters. Unlike cis-methQTLs, trans-methQTLs show little divergence from the functional context of those in normal blood methQTLs, suggesting the long-range methylation changes seen are not changed by carcinogenesis in contrast to some small differences observed in cis mechanisms (Figure 45).

These OR score analyses for enrichment/depletion of methQTL-CpGs in different functional contexts also suggest that some SNP-associated methylation alterations in breast tumours are driven by relatively consistent mechanisms across both cis- and trans-methQTLs (eg. CTCF enrichment, bivalent promoter depletion, eQTL overlap, CGI hypermethylation). SNP-associated CTCF disruption leads to an allele-dependent gain of methylation as CTCF occupancy protects against this hypermethylation, and CTCF loss has been shown in breast cancer to mediate hypermethylated CTCF-promoters to drive tissue-specific cancer progression (Damaschke et al., 2020). On the other hand, studies of bivalent promoters have shown that a redistribution of repressive histone marks may be driven by changes in methylation that impede polycomb-mediated silencing to enable tumorigenesis (Dunican et al., 2020).
CpG island hypermethylation (inherent to all cancer methylomes) may also interfere with allele-specific associations. Indeed, if trans-methQTLs are primarily cis-eQTLs for methylation factors which drive long-range methylation changes in active regions (such as CGIs), an enrichment of trans-methQTL-CpGs in CGIs may be observed as CGI hypermethylation would not interfere with the trans-methQTLs (which are in inactive regions) and their mechanisms of directing methylation change at CGIs. Meanwhile, the depletion of cis-methQTL-CpGs in CGIs may be observed because CGI hypermethylation could disguise the cis mechanism of disrupting methylation factor binding. There would therefore be proportionally less local allele-dependent methylation changes detectable in these CGIs.

Figure 45 – Cis and trans-methQTL mechanisms are largely similar across normal and tumour breast tissue methylome

When comparing normal and tumour breast tissue methylomes, cis-methQTL mechanisms shows minor change in methylation mechanism (enhancer depletion), whereas trans-methQTLs show a similar mechanism.
Finally, there are also a number of *cis*- and *trans*-methQTLs which colocalise with normal breast eQTLs. This builds on both the *cis*-methQTL mechanism model that methylation changes local to a gene promoter may affect regulatory elements for a gene, and also the *trans*-methQTL model of primarily being *cis*-eQTLs for long-distance methylation factors. Indeed, studies of eQTLs in matched normal and tumour have shown that most tumour eQTLs are tumour-specific (Ongen et al., 2014).

### 7.2 - Implications and relevance to the broader field

These results inform a better understanding of how methQTLs are distributed and functionally annotated across the tumour methylome, and how this compares with the normal blood methylome. While studies have looked at *cis*- and *trans*-methQTLs in normal blood (McRae et al., 2018; Huan et al., 2019) or tumour breast tissue (Fleischer et al., 2017; Gong et al., 2019) for their individual roles in disease phenotypes, I have used the distribution and ratio of *cis* and *trans* methQTLs to represent the balance of mechanisms which characterise the abnormal tumour methylome in comparison with the normal methylome. Also, while some functional methQTL characterisations have been made in normal tissue (Min et al., 2021), I created a comprehensive characterisation of the functional context of *cis*- and *trans*-methQTLs respectively, based on SNP-CpG distance, genomic positioning and functional enrichments to describe the distinct mechanisms underpinning the genetic programming of local and distal locus methylation.

I have taken a population-based analytical framework that has been applied in normal tissue (Min et al., 2021), and used that to study the relatively uncharacterised breast tumour methylome. This would have allowed for further analyses to have been repeated using this framework, but under different conditions, had there been more time available. For example, I could have stratified samples by their molecular subtype of breast cancer prior to this analysis to explore the effects of breast tumour subtypes on SNP-associated methylation changes. In addition, this framework was also capable of addressing the more complex biological challenges of tumour methylation studies, by accounting for tumour
heterogeneity though the incorporation of methylation principal components in to the analysis framework. This analytical framework therefore has the potential for further adaption and additional analyses of interest.

These findings will benefit future studies of the breast cancer methylome through a clearer understanding of how local and distal SNP-associated methylation changes are genetically programmed in tumours and of the relatively more balanced ratio of cis/trans-methQTLs responsible for these changes. Considering there are mainly similarities between the cis and trans mechanisms in the breast tumour and normal breast methylomes, future studies can assume that there are only minor differences in the genomic distribution of the loci where local methylation changes are genetically programmed. However, they may also account for minor tumour-specific differences, such as that genetically programmed local methylation changes are less likely to be in enhancers. Moreover, this understanding may also be beneficial for the clinical implications of these findings; as epigenetic treatments have already been shown to have some synergistic potential alongside traditional anticancer therapeutics for breast tumours. However, the clinical benefits will inevitably improve as we continue to develop this understanding (Vietri et al., 2021; Arnold et al., 2022).

For example, consider the observation that cis-methQTLs in breast tumours are less likely to affect methylation in nearby enhancers in tumours than in normal tissue. This cis-methQTL-CpG enhancer depletion can be linked to breast tumour biology, as it is been observed in ESR1 (Oestrogen sensitive receptor 1) positive breast tumours when enhancers are associated with expression of ESR1-regulatory factors (Stone et al., 2015). If hypermethylation also impedes allele-dependent associations with methylation changes (i.e. both alleles associate with the methylated CpG), then this model would align with the tumour depletion of cis-methQTL-CpGs in enhancers. Similarly, the finding that the mechanism of CTCF-occupancy-dependent protection against hypermethylation may have a role equally in trans- as in cis-methQTLs, and that the chromHMM OR score results suggest this not to be the case in normal methylome, could suggest loss of CTCF occupancy as a mechanism to consider when developing new treatment options. Indeed, CTCF loss in breast and prostate cancer has already been shown to cause downregulation of tumour-
associated CTCF-regulated genes and drive cancer progression, but this can be reversed with DNA methylation inhibition (Damaschke et al., 2020). This suggests epigenetic therapies could have a role in reversing CTCF-occupancy-associated cancer progression. Understanding genetically programmed methylation changes in breast tumours may therefore have important clinical and academic value.

7.3 - Limitations of project

The normal tissue selected to make comparisons against the differential methylation observed in the breast tumour methylome was normal blood tissue methylation data. However, a significant limitation to this study is that this comparison was not made with normal breast tissue. This would have removed any tissue-specific biological differences from possibly confounding the methylation changes. Instead, the inherent biological differences between the blood and breast will inevitably have some effect on their distinct methylomes. As such, this influence may appear as tumour-specific effects in this study; Sehl et al., (2017) reported that the DNA methylation-derived epigenetic age of healthy breast tissue is higher than observed in matched peripheral blood tissue. Normal breast tissue studies, however, have been relatively limited and underpowered thus far, making them difficult to use in larger population studies or for identifying methQTLs. As such, in the absence of normal breast methylation resources, the breadth of studies and the large sizes of data cohorts available for normal blood analyses made for a practical comparison when highlighting tumour-specific effects while also maintaining the necessary statistical power.

One solution may have been to use the limited number of matched tumour-adjacent “normal-like” breast tissue samples available from the TCGA breast cancer project (73 out of 783 samples). However, beyond the low sample size making this analysis underpowered, a key limitation of this “normal-like” data is that the idea of “normal” cells may not accurately represent the tumour microenvironment of cells which consists of a mix of normal cells and cancer cells in a tumour tissue. Indeed, tumour-adjacent “normal-like” tissues are an unreliable representation of normal tissue due to cancer field effects, whereby the
immediately surrounding cells may still be under some influence of carcinogenic changes (Seton-Rogers, 2012). However, despite the limitations of the matched normal-like data, it could be leveraged to exclude the possibility of tumour-specific DNA methylation or somatic mutations as confounders in the analysis of normal blood. By validating my top SNP hits with these matched samples, I could highlight broadly if the trends and enrichments observed in these methQTLs aligned closer with normal blood or breast tumour methQTLs. If these trends were closer to the normal blood methylome profile (similar cis:trans ratios; similar enrichment profiles etc.), this would strengthen the validity of the conclusions from this the comparisons made in this project. While I was unable to find any papers which had performed cancer methQTL analyses and accounted for cancer field effects, there are examples of papers which have analysed methQTLs in tumour and tumour-adjacent normal samples. For example, one study mapped eQTLs and methQTLs with prostate cancer risk SNPs between histologically benign tumour-adjacent tissue and tumour tissue, showing that eQTL regulatory mechanisms are maintained in prostate tumours and that DNA methylation may have an intermediary role in these mechanisms (Dai et al., 2020). While this study was relatively unique for considering tumour-adjacent tissue, they also acknowledged that cancer field effects could indeed alter the gene expression profiles of tumour-adjacent tissue.

Another solution to this normal tissue comparison limitation would be to access a repository of normal breast tissue methylation data for comparison, an approach I had considered which would identify a more accurate set of tumour-specific methQTLs. For example, one such repository is the Susan G. Komen Tissue Bank which contains DNA methylation data for healthy women (48,872 samples across 8,694 normal breast donors and 5,022 normal blood donors), which I would have analysed given more time (Sehl et al., 2017). While this would lose the benefit of tumour/normal subjects being matched, it would be a sufficiently powered methQTL analysis to more accurately control for variables other than tumour progression, such as tissue type.

The significance of the methQTLs identified in this analysis are largely subject to the statistical power which is dependent on the study sample size (as are the OR enrichments and depletions). Relative to other methQTL studies, the sample size of this study has been
small, but not uncommonly so for studies which are studying tissues where data resources are still scarce (Shi et al., 2014; Scherer et al., 2021; Zhang et al., 2021). However, the limited statistical power may have inevitably affected long range trans-methQTL detection more so than local cis-methQTLs; this aligns with the lower effect size observed for trans-methQTLs. This may also be a potentially confounding factor when analysing the cis:trans ratio as a representation of the how genetic programming drives the cancer methylome. While efforts are being made to expand datasets with greater data, such as the Enhancing GTEx project (Stranger et al., 2017), it is important to continue analysing available data while acknowledging the caveat of statistical limitations when interpreting underpowered analyses.

One of the most significant challenges when analysing methQTLs in tumours has been accounting for the highly heterogenous cell composition of tumours, whereby samples will have varying proportions of cancer cells in their cell content (Scherer et al., 2020). This is because association signals may correlate with cell-specific methylation patterns rather than germline SNP-associated alterations. While I did identify different methQTL associations after correcting for methylation structure with principal components, I made the decision to exclude this step from my analysis framework and instead I focussed on exploring all SNP-associated methylation changes regardless of the confounding effects of cell heterogeneity. There are some examples of methQTL studies which have accounted for cell heterogeneity, (primarily through the inclusion of methylation data principal components as a covariate) although these are more common in studies of normal tissues (to address less tumour-specific issues, such as batch effects). (Shi et al., 2014; McClay et al., 2015; Huan et al., 2019; Scherer et al., 2021). Heterogenous tumour cell composition has therefore not been widely addressed in methQTLs analyses of tumours. As such, it is difficult to directly compare my analysis to the outcome of similar studies which have chosen to correct for methylation structure, as the conditions of methQTL analyses can vary so much in their modelling decisions. However, given that there were differences changes (albeit minor ones) in the association signals that were identified with and without the methylation principal components as a covariate, my results are likely to be confounded to a small degree by the cell heterogeneity that I decided not to correct for.
One study analysed expression-methylation-QTLs (emQTLs) in breast cancer and found that CpG methylation patterns at enhancers and Oestrogen receptor signalling-associated transcription factors can be used to discern molecular subtypes of breast tumour cells (Fleischer et al., 2017). Furthermore, they identified a cluster of associations signals (through unsupervised clustering of emQTL p-values) which correlated with high lymphocyte infiltration, high intratumor heterogeneity and thus lower cancer proportion of cancer cells in the tumour content (Fleischer et al., 2017). Thus, methylation patterns can be confounded by intratumor cell heterogeneity, but also by molecular subtype-specific methylation patterns and indeed by the variability of the tumour microenvironment.

One important consideration of the confounding effects of tumour heterogeneity is therefore also deciding which cell line annotations will most accurately represent my breast tumour samples. Indeed, a caveat when inferring mechanisms from genomic enrichments is that annotations for these genomic features must be selected and used, but there is never complete certainty that these annotations are accurate representations of the exact cell type in the sample you are studying. For example, annotations from a breast tumour cell line (such as MCF7) may be selected to represent breast tumours, but the nuanced cell-specific biology of breast cancer means that some cell line annotations may better represent some molecular subtypes compared with others. This would be another benefit if I had accounted for cell type heterogeneity in my analysis, as I could use genomic annotations for cell lines which more closely represent specific cell types. For this reason, these functional genomic feature enrichments are valuable, but limited in how precisely they represent cell type heterogeneity.

One factor which was not addressed in this study but may have a significant influence on unexplained tumour methylation changes (and when describing the functional context of methQTLs) is the impact of environmental exposures on methylation, and indeed methQTLs. One prominent environmental modifier is smoking (Philibert, Beach and Brody, 2014). For example, Shenker et al found that a SNP in the smoking-associated gene ATTR also had prognostic value as a biomarker for smoking-associated methylation changes (Shenker et al., 2013). The different proportion of cis/trans methQTLs may be confounded by interactions between smoking and methylation if a SNP associates with smoking
behaviour, which consequently affects genomic methylation (Joehanes et al., 2016). Indeed, \textit{cis}-methQTL-SNPs were shown to associate with CpGs which also associate with smoking, but these \textit{cis}-methQTL-SNPs have themselves no association with smoking exposure or mortality, and so this suggests they could confound \textit{cis}-methQTLs (Gao et al., 2017). As environmental exposures are inconsistent across cohorts, this could also be one explanation for \textit{trans}-methQTLs that are less replicated across studies. One way to address this would be to ask if any of these methQTLs colocalised with GWAS hits for smoking-related phenotypes, such as smoking exposure or smoking-related cardiovascular disease. Other environmental exposures with significant methylation associations which could also have been studied includes BMI (Reed et al., 2020) and alcohol (Liu et al., 2018).

Another consideration has been ensuring that the genetic changes which are driving these methQTLs are from germline changes, and not confounded by the widespread somatic changes which are also inherent to carcinogenesis (Greenman et al., 2007; Oh and Sung, 2021). This genotype alteration could arise from somatic mutation, or through loss of heterozygosity, which is an early genetic event in half of breast cancer cases (Cleton-Jansen et al., 2000). As these somatic changes are a significant and complex confounder, I focussed on methylation changes with associate with germline SNP genotype data to ensure we were not considering broad methylation changes which may associate with somatic mutations. However, other studies have more directly considered the distinction between these two; for example, Pattee et al. (2020) adapted a TWAS analysis by incorporating both somatic and germline cancer genotype data to identify germline eQTLs in lung cancer and describe significant germline effects after correcting for somatic alterations. Additionally, Pereira et al. (2022) first performed a global eQTL analysis before also determining somatic eQTLs in order to explicitly identify germline effects on gene expression. They found that somatic alterations were the dominant source of expression variation across different tumours (Pereira et al., 2022). Thus, one improvement to my analytical framework could be to adapt a methQTL “screening” approach to find germline-specific by removing any methQTLs which were also identified using somatic tissue genotype data.

One specific cancer-associated somatic chromosomal rearrangement which may be a confounder for this methQTL analysis is cancer-associated chromosomal translocations.
Faulty double strand break repair (which is common in cancer) can lead to proximal, non-homologous chromosome ends merging incorrectly, resulting in translocations of genetic sequence (Canoy et al., 2022). This could mean that some of the trans-methQTL signals I detect may have originally been cis-methQTL signals before chromosomal translocation. However, the methylation effects around these translocations are unclear; differential methylation around translocation break points have been shown in a chromosomal translocation (between chromosomes 1 and 11) associated with psychiatric illness (such as schizophrenia) (McCartney et al., 2018). However, a study in leukemia suggested that a BRC-ABL translocation (between chromosomes 9 and 22) showed no association with the abnormal methylation effects in cancer (Byun et al., 2017). As such, it would be interesting to ask if any of the methylation changes observed in breast cancer are associated with chromosomal translocation sites. This could be addressed by DNA break point analysis, which assesses significant copy number alterations in DNA to detect where chromosomal translocation are likely to have occurred (Ritz et al., 2011). Another way to address this could be to account for these associations during quality control steps. For example, we might expect that if a CpG showed a trans-association with only one SNP, and not with any neighbouring SNPs, then this is uncharacteristic of a normal association and could suggest somatic mutations or translocations.

One interesting analysis I do on multiple occasions is to consider if the QTLs I identify for methylation patterns in breast tumour patients are also QTLs for other molecular phenotypes in different tissues. However, given the relatively small number of methQTLs I identified (13,195), the sample size differential is substantial in comparison with the larger datasets which are in the millions of QTLs (the normal breast eQTL and the GoDMC normal methQTL datasets). As such, a degree of these overlapping QTLs would be expected to be overlapping by chance, rather than by statistical significance. Therefore, this overlap analysis requires further validation to address this statistical caveat; in future, I would perform a fisher’s test to see if the overlapping QTLs are statistically significant against the background of non-overlapping QTLs between the various datasets.
7.4 - Future Work

The mechanisms which these results have suggested may be common across both cis- and trans-methQTL-CpGs (such as CTCF enrichment and bivalent promoter depletion) could all be further validated by mechanistic experimental methods or further computational analysis to match SNP-associated methylation changes with an allele within a cohort. For example, I could explore this relationship further by stratifying the cohort by their alleles for a SNP within a CTCF state, and then leveraging data on CTCF occupancy (ChIP-seq data from UCSC) and CTCF expression (from GTEx) alongside methylation data for an individual CTCF binding site. According to the hypothesised model of CTCF site disruption, one allele would align with the low CTCF occupancy, low CTCF expression and hypermethylated CTCF sites, whereas the alternative allele would associate with high CTCF occupancy, high CTCF expression and hypomethylated CTCF sites. As an alternative in vitro experiment, I could also introduce a SNP in to a CTCF binding site in in a breast cancer cell line and measure any local allele-dependent methylation changes.

Using a similar logic, I could also use these experimental approaches with polycomb-associated histone marks. Under the hypothesized model for bivalent promoter hypermethylation disrupting PRC2 complex, I could leverage expression data for proteins of the PRC1/2 complexes (from GTEx), repressive H3K27me3 histone marks and bivalent promoter occupancy (both using ChIP-seq data from UCSC) alongside DNA methylation data. I would expect to see one allele align with high expression of PRC2 complex-associated proteins, PRC-associated protein binding at a bivalent promoter, H3K27me3 histone modifications and hypomethylation, while the alternative allele would associate with low PRC1/2 complex-associated protein occupancy and expression, an absence of H3K27me3 modifications and hypermethylation. Similarly, if I downregulated PRC1/2 complex-associated proteins in vitro, I would expect to see subsequent hypermethylation of the bivalent promoter.

As mentioned, one future development would have been to repeat this analysis in some alternative datasets that would shed greater light on the underlying breast tumour
biology and consider the complexity of tumour tissues. For example, rather than normal blood, I could use unmatched normal breast tissue samples and matched breast normal-like samples to develop a series of observations which represent the temporal progression of methylation changes in breast cancer. This would begin with normal breast (available from the Komen databank of healthy breast samples mentioned previously), progress through tumour-adjacent normal-like samples (accessed from TCGA) and finally to the tumour samples (Sehl et al., 2017). Another direction for further analysis of the underlying breast cancer biology would be to repeat the study whilst accounting for the breast tumour heterogeneity in the association analysis. As discussed, this could be done by integrating the methylation data principal components as a covariate in the association analysis, and would disseminate methQTLs which were truly SNP-associated methylation changes apart from associations which were instead associating with cancer-associated widespread methylation change, which are cell-type specific within the breast tumour cell content. Finally, another approach could have been to first stratify samples by their molecular subtypes and then to find subtype-specific methQTLs. This could be achieved using a combination of expression or methylation markers for estimating molecular subtypes (Yersal and Barutca, 2014; Chen et al., 2019). However, this would have answered a distinct and separate set of aims from my own which would have focussed more on studying breast cancer biology, and so this analysis was not pursued. This analysis could answer questions such as how methylation defines molecular subtypes and the spatio-temporal patterns of methylation throughout cancer progression.

Alternatively, I could have expanded on the novel approach taken in this project to consider the wider distribution of methQTLs to represent breast cancer methylome, rather than focus on individual methQTLs to represent a specific gene model. I would focus more on how the hallmarks of the cancer methylome, namely PMD global hypomethylation and polycomb-marked promoter CGI (poly-CGI) hypermethylation, change individually between normal and tumour breast samples. For this, I would define and measure these features (i.e. hypomethylated PMDs and hypermethylated poly-CGIs) across each sample to create an aggregate score that represents the feature phenotype metric for each individual. I would then calculate genetic scores to see if there were any SNPs which associated with genome-wide PMD hypomethylation or poly-CGI hypermethylation respectively, and which may
therefore suggest SNPs or genes which may have a driving role in one specific hallmark feature rather than simply of methylation changes at a few CpGs. This could discover distinct epigenetic regulators of these hallmark features, or more interestingly, could also discover overlapping mechanisms if these two hallmark processes show some dependency on one another.

Finally, I could have also developed these functional characterisations of the methQTLs by asking what other molecular phenotypes of interest may also associate with the SNPs in these breast tumour methQTLs. I would first perform a colocalisation analysis with GWAS hits for characteristic phenotypes of breast tumours; this could be for mortality or tumour aggression in breast tumour samples, as was performed by Houlahan for prostate cancer (Houlahan et al., 2019). This may reveal methQTLs where the SNPs underpinning methylation changes also have some association with tumour severity. Alternatively, this could be for prognosis by examining risk SNPs in normal samples. For example, Yang et al. (2019) looked at associations between methylated CpGs and GWAS hits for breast cancer risk, and found novel CpG markers for risk. I could expand on the GtEX eQTL overlap analysis by accessing RNA-seq data from TCGA for the samples I have analysed, as this would therefore provide a set of eQTLs which are more accurately representative of my samples. Another interesting colocalisation analysis would be with QTLs which associate with protein levels (pQTLs), which would reveal SNPs which may implicate abnormal protein levels with these methylation changes. For example, Shu et al. (2022) used pQTL database and breast cancer molecular subtype data to infer proteins which were subtype-specific, and used mendelian randomisation to show a causal direction whereby the pQTL drives subtype risk. Another example could be to study how methylation may associate directly with gene expression independently of SNP data; quantitative trait methylation (eQTMs), which represent an association between a CpG site and gene expression, could be used to further characterise the role of methylation changes in breast tumour (Bonder et al., 2017). Finally, I could have used ontological enrichment analyses to ask which genes and functional pathways these cis/trans-methQTL-SNPs may have been associated with, which would have given more characterisation to the individual methQTLs.
7.5 - Concluding Statement

While population approaches have previously interrogated the relationship between genetic sequence and methylation changes in normal breast tissues, the limited resources and molecular complexity of studying tumour tissues has meant that comparisons between the normal and tumour methylome of breast tissue (using a comprehensive characterisation of the genomic and functional profile of methylation QTLs) have been limited. Overall, I found that breast tumour methQTLs indicate similarities in how SNP-associated methylation changes are genetically programmed in the tumour methylome compared to the normal breast methylome. However, some key differences are suggested in where methylation changes occur and in the genomic balance of long-range and short-range SNP-CpG associations. This will inform future studies of the similarities and minor differences by which tumour-associated methylation changes are genetically driven in breast tumours.
References


Aran, D. et al. (2017) ‘Comprehensive analysis of normal adjacent to tumor transcriptomes’, Nature Communications, 8, p. 1077. Available at: https://doi.org/10.1038/s41467-017-01027-z.


Banovich, N.E. et al. (2014) ‘Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels’, PLOS Genetics, 10(9), p. e1004663. Available at: https://doi.org/10.1371/journal.pgen.1004663.


Core, L.J. et al. (2014) ‘Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers’, *Nature genetics*, 46(12), pp. 1311–1320. Available at: https://doi.org/10.1038/ng.3142.


Fleischer, T. et al. (2017) ‘DNA methylation at enhancers identifies distinct breast cancer lineages’, Nature Communications, 8(1), p. 1379. Available at: https://doi.org/10.1038/s41467-017-00510-x.


Fujiki, K. et al. (2013) ‘PPARy-induced PARylation promotes local DNA demethylation by production of 5-hydroxymethylcytosine’, Nature Communications, 4, p. 2262. Available at: https://doi.org/10.1038/ncomms3262.


Han, D. et al. (2017) ‘CTCF participates in DNA damage response via poly(ADP-ribosylation)’, Scientific Reports, 7(1), p. 43530. Available at: https://doi.org/10.1038/srep43530.


Min, J.L. et al. (2021) ‘Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation’, Nature Genetics, 53(9), pp. 1311–1321. Available at: https://doi.org/10.1038/s41588-021-00923-x.

Minimac Info File - Genome Analysis Wiki (no date). Available at: https://genome.sph.umich.edu/wiki/Minimac3_Info_File#Rsq (Accessed: 9 February 2024).


Oliva, M. et al. (2022) ‘DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits’, *Nature Genetics*, pp. 1–11. Available at: https://doi.org/10.1038/s41588-022-01248-z.


Pan, L. et al. (2023) ‘Hidden Genetic Regulation of Human Complex Traits via Brain Isoforms’, *Phenomics* [Preprint]. Available at: https://doi.org/10.1007/s43657-023-00100-6.


Stranger, B.E. et al. (2017) ‘Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease’, Nature Genetics, 49(12), pp. 1664–1670. Available at: https://doi.org/10.1038/ng.3969.


Wu, Y. *et al.* (2018) ‘Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits’, *Nature Communications*, 9, p. 918. Available at: https://doi.org/10.1038/s41467-018-03371-0.


