Influence of external factors on the human epigenome

Olivia Alexandra Grant

Submitted in fulfillment of the requirements

of the Degree of Doctor of Philosophy

School of Life Sciences

University of Essex

January 2023

Abstract

Epigenetic mechanisms govern gene regulation and respond to both genetic and environmental influences. Epigenetic marks, including DNA methylation (DNAm) are dynamic across an individual's life and may be influenced by several environmental and genetic factors. In this thesis, I evaluated several factors which influence the human epigenome using a bioinformatics approach.

First, I explored the idea that DNAm changes may underlie the link between air pollution and poor health. I identified no significant link between background air pollution levels and DNAm, however, was able to identify a link between exposure to traffic related air pollution (TRAP) and DNAm. I identified 531 significant CpG sites related to TRAP which were enriched at regulatory sites and novel and previously reported transcription factor (TF) motifs and genes.

Secondly, I explored the role of DNA methylation in autosomal sex differences. Specifically, I identified and validated 396 CpG sites and 266 differentially methylated regions associated with sex. We found the majority of these to be female-biased CpGs (74%) and were enriched in CpG islands and located preferentially at 5/3'UTRs and enhancers. TF motif enrichment revealed enrichment of TF's related to critical developmental processes and sex determination such as SRY and ESR1.

Finally, I report a catalogue of loci across the human epigenome displaying either variable or stable inter- individual DNA methylation. I demonstrate that the majority of the VMPs are not controlled by age, sex or smoking status. I also report highly variable and stable CpG sites enriched at methylation sensitive TFs, highlighting that there is potentially a strong relationship between DNAm and TF binding, which will be important for gene regulation. In addition, I found that the VMPs were under higher genetic control than the SMPs and that this is in part directed by interindividual differences in 3D chromatin organisation. In summary, the results presented in this thesis give insights into the role that external factors play in influencing the human epigenome.

Acknowledgements

Firstly, my deepest gratitude goes to my supervisor Dr. Nicolae Radu Zabet who has guided me throughout my undergraduate to my PhD. His expertise and enthusiasm for the research we have conducted together over 4.5 years has motivated me in times I was struggling the most. He believed in my abilities at times I was unable to and for that I am eternally grateful. I could not have wished for a better advisor throughout this PhD.

I would also like to extend my gratitude to my supervisor, Professor Leonard Schalkwyk. His support, unending wisdom and computational knowledge have improved this thesis beyond expected. Besides his contribution to my academic progression, his personal generosity has made my time at the University of Essex highly enjoyable. He has always supported and encouraged me to explore the fun side of academia and science which has shaped me as a researcher and professional.

I am also sincerely grateful to my third supervisor, Professor Meena Kumari for her invaluable advice throughout my PhD, without her, many aspects of this PhD would have been impossible, and for that I am very grateful.

I thank all current and former members of the Genomics group at the University of Essex as well as my PhD cohort. Despite completing most of our PhD research in our bedrooms alone, we have all managed to support and encourage eachother endlessly. Special thanks goes to Eleanor Watson, who has truly made this experience as enjoyable as possible. From podcasting to pint of science, you made this experience unpredictable and exciting along every step of the way, I truly believe I would not have got through this as stress free - had you not had accompanied me on this journey.

Lastly, I extend my gratitude to my family, Mum, Harry, Granny and Grace who always have, and always will be a constant support and encouragement to me. This PhD would not have been possible without you all, truly.

Declarations

Chapter 2-5 Publicly available datasets were sourced from GEO and are the accession numbers are listed in the appropriate sections. Data collection and quality control for the Understanding Society discovery data set: UK Household Longitudinal Study was performed by Dr Melissa Smart, Dr Yanchun Bao and Professor Meena Kumari at the Institute of Socio- Economical Research at the University of Essex in collaboration with Professor Jon Mill, Dr Eilis Hannon and Dr Joe Burrage from the University of Exeter Medical School. Data quality control for the Understanding Society validation data set: UK Household Longitudinal Study was performed by myself, Dr.Tyler Gorrie Stone, Yucheng Wang and Professor Leonard Schalkwyk at the University of Essex. The ordinary kriging models used in Chapter 3 were developed by myself but due to data protection were performed by John Payne, Associate Director of data at Understanding Society.

Chapter 4 This chapter has been published by the author and is provided in the appendix of this thesis. Grant, O.A., Wang, Y., Kumari, M. et al. Characterising sex differences of autosomal DNA methylation in whole blood using the Illumina EPIC array. Clin Epigenet 14, 62 (2022). https://doi.org/10.1186/s13148-022-01279-7.

Impact of COVID-19 statement

The COVID-19 pandemic impacted this thesis in a significant way. I commenced the research for this thesis in October 2019, and in March 2020, Essex University suspended all face to face research and teaching. The original plan for this thesis, was to focus solely on exploring the research aims raised in chapter 3. However, due to data access issues experienced as a result of COVID-19, it was decided to expand the aims of my thesis which resulted in chapters 4 and 5. This meant redefining the thesis aims and learning new skill sets and moving to full time remote study.

List of abbreviations

DNA	Deoxyribonucleic acid
DNAm	DNA methylation
CpG	Cytosine-Phosphate-Guanine
5mC	5-methyl cytosine
CpGI	CpG islands
TF	Transcription factor
DNMT	DNA methyl transferases
ESCC	Esophageal squamous cell cancer
HESC	Human embryonic stem cells
SAM	S-Adenosyl methionine
TET	Ten eleven translocation
ESC	Embryonic stem cells
5hmC	5 hydroxymethylcytosine
5fC	5 formylcytosine
5caC	5 carboxylcytosine
BER	Base excision repair

NGS	Next Generation sequencing
HPLC	High performance liquid chromatography
GWAS	Genome wide association study
EWAS	Epigenome wide association study
DMP	Differentially methylated position
DMR	Differentially methylated region
LUR	Land use regression
$\mathbf{PM_{10}}$	Particulate matter 10
$\mathbf{PM_{2.5}}$	Particulate matter 2.5
$\mathbf{NO_2}$	Nitrogen dioxide
$\mathbf{O_3}$	Ozone
PAH	Polyaromatic hydrocarbon
AURN	Automatic urban ad rural network
IDW	Inverse distance weighting
OK	Ordinary Kriging
AQUM	Air quality unified model
TRAP	Traffic related air pollution
saDMP	Sex associated differentially methylated probe
saDMR	Sex associated differentially methylated region
FDR	False discovery rate
SD	Standard deviation

TE	Transposable elements
VMP	Variability methylated region
SMP	Stably methylated region
GO	Gene ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
MCC	Maximum clique centrality
mQTL	Methylation quantitative trait loci
TAD	Topologically associated domain
RSV	Respiratory syncytial virus
CR	Control regions
NSCLC	Non small cell lung cancer

Contents

Abstract 1

Acknowledgements 3

Declarations 4

Impact of COVID-19 statement 5

List of abbreviations 6

List of Figures **13**

List of Tables **16**

1 Introduction **17**

1.1 Epigenetics 17

1.2 DNA methylation 18

1.3 Mechanisms of DNA methylation 19

1.4 DNA methylation throughout life 21

1.5 DNA Demethylation 22

1.6 Assessing DNA methylation 23

1.7 Epigenome wide association studies 25

1.8 Considerations for epigenome wide association studies 27

1.9 Aims of this thesis 29

2 Materials and methodology **30**

2.1 Materials 30

2.1.1 Understanding Society 30

2.1.2 Publicly available data sets used 31

2.2 Epigenome wide association studies 32

2.2.1 DNA methylation measurements and quality control . . . 32

2.2.2 Identifying sex associated autosomal differential methylation 33

2.2.3 Identifying pollution associated differential methylation . . 34

2.3 Identifying variable and stable sites on the Illumina EPIC array . 35

2.4 Genomic annotation of CpG sites 36

2.5 Enrichment of CpGs in transcription factor binding motifs 36

2.6 Gene ontology analyses 37

2.7 Protein-protein network visualisation and hub gene identification . 37

2.8 Integration with gene expression 37

2.9 Overlap of CpGs with chromatin loops 38

2.10 Annotation of VMPs and SMPs to methylation quantitative trait loci and their occupancy in topologically associated domains . . . 39

2.11 Characterising epialleles 40

3 How does the environment influence the human epigenome? 41

3.1 Introduction 41

3.1.1 Air pollution 42

3.1.2 Assessing air pollution 44

3.1.3 The implications of air pollution exposure preconception and in utero 46

3.1.4 Air pollution exposure in children 50

3.1.5 Air pollution induced DNAm changes in adults 54

3.1.6 Additional factors affecting DNA methylation and air pollution 59

3.1.7 Air pollution and accelerated ageing 62

3.1.8 Limitations of the current research 63

3.1.9 Conclusions 65

3.2 Methods 66

3.3 Development of air pollution exposure measures 66

3.3.1 Introduction of current methods and limitations 66

3.3.2 Air pollution data sources 68

3.3.3 Bayesian model-based estimates 69

3.3.4 Exposure estimates at postcode level 71

3.3.5 Distance to road measures 75

3.4 Results 76

3.4.1 Statistical analysis for models to estimate air pollution exposure 76

3.4.2 Inverse distance weighting model 77

3.4.3 Ordinary Kriging model 78

3.4.4 Mapping in R 79

3.5 Epigenome wide association study of air pollution 81

3.5.1 There are no significant detectable DNA methylation changes in response to exposure to background levels of ambient air pollution 81

3.5.2 A distance to road measure gives insight into how exposure to traffic related air pollution may modify DNA methylation signatures across the genome 86

3.5.3 Characterisation and location of road associated DMPs . . 88

3.5.4 TFs enriched at raDMPs 94

3.5.5 Identifying road associated differentially methylated regions 97

3.6 Discussion 98

4 The role of DNA methylation in autosomal sex differences 106

4.1 Introduction 106

4.2 Results 109

4.2.1 Females show higher methylation at a subset of autosomal loci 109

4.2.2 Characterisation of sex associated DMPs 113

4.2.3 Enrichment of saDMPs in transcription factor binding sites 117

4.2.4 Relationship with gene expression 122

4.2.5 Sex associated differentially methylated regions 123

4.3 Summary and discussion 125

5 Identifying interindividual variation and stability of DNA methylation 135

5.1 Introduction 135

5.2 Results 139

5.2.1 Variably methylated probes and stably methylated probes are widespread across the genome and enriched in regulatory regions 139

5.2.2 VMPs and SMPs located at promoters and enhancers are enriched for transcription factor motifs 143

5.2.3 VMPs are under higher genetic control than SMPs 148

5.2.4 Identifying putative epialleles in human whole blood . . . 155

5.3 Summary and Discussion 160

6 General discussion 167

6.0.1 Key findings and their implications in the field 168

6.0.2 Limitations and future research directions 172

6.0.3 Final words 175

A Appendix 176

Bibliography 201

List of Figures

1.1 Waddington's epigenetic landscape 18

1.2 Overview of the mechanisms of DNA methylation and DNA demethylation 23

3.1 UK map showing locations of the 144 AURN monitoring stations 68

3.2 Inverse distance weighting interpolation illustration 72

3.3 Study area of interest demonstrating known locations and unknown locations for pollution interpolation 74

3.4 Results from inverse distance weighting models 78

3.5 Results from ordinary kriging models 80

3.6 Interpolated values by Ordinary Kriging of nitrogen dioxide and $PM_{2.5}$ 80

3.7 Manhattan plot for EWAS analysis of exposure to $PM_{2.5}$. 82

3.8 Scatterplot showing relationship between methylation value and $PM_{2.5}$ exposure at CpG site cg21553490 in gene RNF6. 84

3.9 Histogram showing range of $PM_{2.5}$ exposure values across our study sample for the study period of 2007-2011. 85

3.10 Manhattan plot for EWAS analysis of exposure to the four main air pollutants 86

3.11 Manhattan plot for EWAS analysis of exposure to TRAP. 87

3.12 Annotation of raDMPs relative to CpG islands 90

3.13 Functional annotations of raDMPs. 91

4.1 Venn diagram showing overlap of differentially methylated positions identified in our validation and discovery data set before and after filtering of the list of saDMPs 110

4.2 QQ plots and cell type composition plots for discovery and validation data sets. 112

4.3 Characterisation of saDMPs 114

4.4 Location of saDMPs 115

4.5 Chromatin loops connecting saDMPs to additional genes 116

4.6 Subnetworks of the top 30 genes annotated to saDMPs 118

4.7 Transcription factor motif enrichment analysis 119

4.8 GO enrichment analyses for enriched TF motifs at saDMPs 120

4.9 Protein protein interaction networks for TFs enriched at saDMPs 121

4.10 Differentially expressed genes between males and females 123

4.11 Plots of sex-associated differentially methylated regions (saDMR) 126

5.1 Venn diagram showing the number of VMPs and SMPs identified in discovery and validation data sets 141

5.2 Characteristics of VMPs and SMPs 142

5.3 Identification and location of variably methylated probes (VMPs) and stably methylated probes (SMPs) 144

5.4 Transcription factor motif enrichment analysis for VMPs and SMPs at promoters 146

5.5 Transcription factor motif enrichment analysis for VMPs and SMPs at enhancers 147

5.6 Methylation quantitative trait loci connect VMPs and SMPs with SNPs. 149

5.7 Barplot representing the percentage of trans mQTL pairs connected by chromatin loops 150

5.8 Barplot representing the percentage of trans mQTL pairs in the same topological associated domain 151

5.9 mQTL pairs occupy the same topologically associated domains . . 152

5.10 Enriched GO and KEGG terms for mQTLs. 154

5.11 Epiallele like sites identified in whole human blood 157

5.12 Functional annotations of epiallele like sites in human whole blood 158

5.13 Relationship between DNA methylation variation and target gene expression variation in microarray data 159

List of Tables

2.1 Datasets used throughout this thesis 30

3.1 Missing data (percent) for the four pollutants of interest from the period of 2007-2011. 69

3.2 Parameters collected from the AURN monitoring stations across the UK 70

3.3 Top 20 significantly associated raDMPs identified by EWAS analysis. 92

3.4 Additional genes annotated to raDMPs via chromatin loops 93

3.5 Enriched transcription factor motifs at raDMPs. 94

3.5 Enriched transcription factor motifs at raDMPs. 95

3.5 Enriched transcription factor motifs at raDMPs. 96

3.5 Enriched transcription factor motifs at raDMPs. 97

3.6 3 road associated differentially methylated regions identified . . . 98

4.1 Enriched GO terms among the genes containing the 54,261 CpGs identified to be significantly associated with sex 110

4.2 Overlap of autosomal sex-associated differentially methylated positions reported in this study with previous literature in various tissues 128

A.1 Enriched KEGG terms for VMPs and SMPs 177

Chapter 1

Introduction

1.1 Epigenetics

The field of epigenetics emerged from the work of iconic scientists such as Waddington and Hadorn and their work aiming to combine genetics and developmental studies. Epigenetics, a term first coined by Waddington in 1942 means 'above' or 'on top of' genetics. Waddington developed a metaphorical epigenetic landscape to provide modern day scientists with a powerful, yet simple way of thinking (Figure 1.1). Epigenetics is defined as mitotically heritable changes which correlate with gene expression and consequently protein expression [Jaenisch and Bird, 2003]. This ultimately leads to genes being 'switched on and off', a process which is crucial for cellular activities such as differentiation. Epigenetic changes do not alter the primary DNA sequence itself, and there are three main epigenetic regulators, DNA methylation, histone modifications and non-coding RNAs. Exploring these markers enables investigations into the relationship between epigenetics and a range of disorders. These markers are flexible and can be altered in response to several environmental factors such as air pollution, smoking and physical activity level [Alegría-Torres et al., 2011]. These epigenetic hallmarks have potential to provide biomarkers and act as indicators of exposure, allowing identification of vulnerable individuals.

1.2 DNA methylation

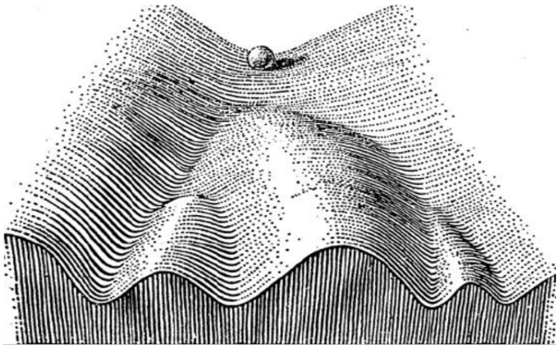


Figure 1.1: Waddington's epigenetic landscape, the position of the ball represents different cell fates.

DNA methylation (DNAm) is the quintessential epigenetic modification and it involves the addition of a methyl group to the fifth carbon of a cytosine base. A modified version of cytosine was first found in 1948 by Hotchkiss during his preparation of bovine thymus when using paper chromatography [Hotchkiss, 1948]. During his experiments he concluded that these modified cytosines i.e. methylated cytosine existed naturally in DNA. Despite this, the role of methylated cytosine in mechanisms governing gene function was not well established until 1980 [Razin and Riggs, 1980]. DNA methylation occurs predominantly on cytosines proximal to a guanine base, commonly referred to as CpG sites. The majority of CpG sites within the genome are methylated and generally thought to induce transcriptional repression upon interaction with histone modifications and non-coding RNAs, and is essential for silencing retroviral elements, genomic imprinting and X chromosome inactivation [Moore et al., 2013]. However, levels of modified cytosine vary largely across cell types and genomic locations and are not always located within CpG sites, non CpG methylation has previously been identified in plants, brain tissues and embryonic stem cells [Xie et al., 1999],[Catoni et al., 2018],[Lamadema et al., 2019].

The relationship between DNA methylation and gene expression remains poorly understood, with elevated gene body DNAm being attributed to active expression [Wagner et al., 2014] and hypermethylation within enhancers and promoters leading to stable transcriptional silencing [Ehrlich and Lacey, 2013]. This relationship between DNA methylation and gene expression was first coined by Bird in 1987 through his work demonstrating that methylated and non-methylated CG sequences coexist in many animal genomes [Bird, 1987]. The majority of the genome is thought to be depleted of CpG's, possibly due to spontaneous deamination of 5mC to Thymine leading to CpG suppression in vertebrates.

Despite the lack of genome wide CpGs, most genes are known to contain CpG islands (CpGIs). These are short CpG rich regions approximately 1kb long with a GC content greater than 50%. These regions tend to be unmethylated, especially when they are located at transcription start sites [Bird et al., 1985]. Therefore, it is not surprising that CpGIs are commonly located within promoter regions of housekeeping genes and avoid spontaneous deamination since they are rarely methylated in germline [Smallwood et al., 2011]. CpGIs are therefore described as being highly conserved in humans and mice, highlighting their functionality in gene regulation [Vinson and Chatterjee, 2012]. Dynamic features of CpGIs enable enhanced gene expression, including the lack of nucleosomes which brands DNA permissive for transcription. Secondly, the presence of histone modifications which encourage transcription factor (TF) binding and amplified gene expression [Wang et al., 2012].

When methylation of CpGIs is observed, it normally functions to serve long term silencing (X chromosome inactivation for example). CpGIs are contiguous with CpG shores followed by CpG shelves, which are located 2kb from an island and 2kb outside of a shore, respectively [Asmar et al., 2015]. CpGs shores have been reported to be related to tissue- and cancer-related methylation and age-related hypomethylation change have also been reported to be strongly related to gene expression [Jaffe and Irizarry, 2014].

1.3 Mechanisms of DNA methylation

CpG methylation is catalysed by a family of DNA methyltransferases (DNMT1, DNMT3A, DNMT3B and DNMT3L) which facilitate the covalent linkage of a methyl group to the fifth carbon in the pyrimidine ring of a cytosine residue, subsequently forming 5mC (5-methylcytosine). DNMT1 is constitutively expressed and known as the maintenance methyltransferase due to its preference to methylate hemimethylated DNA, ensuring the maintenance of the methylation pattern to daughter strands during mitosis. DNMT1 localizes to the replication fork during S phase to detect imbalances in methylation amongst CpGs between a parent and newly synthesised strand of DNA. Therefore, adding a methyl group to the newly synthesised strand ensuring identical methylation patterns. DNMT1 has also been shown to accumulate at DNA damage sites via the PCNA binding domain to restore epigenetic information during DNA repair [Mortusewicz et al., 2005]. Mouse studies have demonstrated DNMT1 plays an important role in inhibiting proliferation and metastasis in esophageal squamous cell carcinoma (ESCC) suggesting DNMT1 is an important target in ESCC therapy [Bai et al., 2016]. Moreover, deletion of DNMT1 in human embryonic stem cells has previously been demonstrated to prompt rapid cell death indicating a role in its requirement for viability in human embryonic stem cells (hESC's) specifically [Liao et al., 2015]. Despite this, there is evidence implicating DNMT1's role in de-novo methylation of a sub telomeric repeat [Egger et al., 2006] indicating a more complex relationship between DNMT1 and its effects on cell development.

On the other hand, DNMT3A and DNMT3B are essential in establishing DNA methylation patterns and therefore are recognized as the de novo methyltransferases due to their ability to target unmethylated CpG sites [Probst et al., 2009]. These enzymes are most active during the blastocyst stage of embryonic development and are highly expressed in undifferentiated cells, however poorly expressed in somatic tissues [Okano et al., 1999]. DNMT3A has been shown to be expressed in most adult tissues and expressed ubiquitously, whereas DNMT3B is found to be expressed at low levels within adult tissues (excluding the testes, bone marrow and thyroid). DNMT3A and DNMT3B both contain conserved cysteine rich regions [Xie et al., 1999] and variable N terminals which are followed by a PWWP domain, important for non-specific binding to DNA. Research involving silencing of DNMT3a addressed the idea that DNMT3A acts as a tumour suppressor gene in lung cancer and can be a determinant of malignancy highlighting the harmful consequences upon absence of the enzyme warranting further research into the activation of silenced DNMT3A in lung cancer therapy [Gao et al., 2011]. The downregulation of DNMT3B in PC3 cells has also been demonstrated to increase apoptosis and restrict cell growth and migration [Yaqinuddin et al., 2008].

DNMT3L is an accessory molecule to the de novo methyltransferases DNMT3A and DNMT3B which function by increasing their binding to S-adenosyl-L-methionine (SAM), the methyl donor [Jin et al., 2011]. DNMT3L lacks a catalytic domain therefore interacts with DNMT3A and DNMT3B to facilitate their enzymatic activity [Chedin et al., 2002],[Chen et al., 2005]. DNMT3L expression with DNMT3A is more significant in stimulating de novo methylation than with DNMTB at various endogenous and non imprinted sequences within the genome, suggesting DNMT3L is essential for the establishment of methylation imprints [Chedin et al., 2002]. However, much like the complications seen in the function of DNMT1, theories have been supported in which DNMT3B plays a role in the methylation of germ line genes in somatic cells [Walton et al., 2011]. DNMT3B is also responsible for de-novo methylation of bodies of transcribed genes ultimately leading to their preferential methylation [Baubec et al., 2015].

1.4 DNA methylation throughout life

It is fairly well established that most methylation patterns observed in embryos remain fixed throughout the life span, but prior to this, two waves of genome wide demethylation and re-methylation occur [Greenberg and Bourc'his, 2019]. The first wave of demethylation occurs after fertilization and again, in the germline. This is necessary for dealing with the asymmetry of DNA methylation exhibited by the paternal and maternal genomes [Greenberg and Bourc'his, 2019]. At the blastocyst stage, roughly only 20% of CpGs remain methylated and these are restricted to transposable elements and imprints. Re-methylation occurs through activation of the aforementioned DNMT's which increase CpG methylation to around 70-80%. This process is then repeated in primordial germ cells where additionally, imprinted regions are erased and reset in a sex specific manner [Greenberg and Bourc'his, 2019]. Both waves affect the maternal and paternal genomes.

1.5 DNA Demethylation

Recent work in the field is revealing that DNAm is not in fact as stable as previously thought despite the stability of the bond between the methyl group and the 5-position of the cytosine base. Now it is widely recognized that demethylation of DNA or loss of DNAm can occur through mechanisms which are passive or active. DNA demethylation is important for imprints and development in mammals and is thought to play a role in the development of tumorigenesis [Popp et al., 2010],[Szyf et al., 2004],[Jin and Liu, 2018] . Therefore unveiling the mechanisms by which this process occurs offers insightful implications for regenerative medicine . Passive DNA demethylation involves the loss of a methyl group from 5-methylcytosine due to DNMT1 deficiency during cell division. Active DNA demethylation results in the conversion of 5-mC to cytosine by ten-eleven translocation methylcytosine dioxygenase (TET) enzymes. The TET family proteins include TET1-3 which all possess two conserved domains (a Cys rich and dioxygenase domain) and Fe(II) and 2-OG cofactors [Rasmussen and Helin, 2016]. These enzymes have been shown to have similar enzymatic activities and are part of the ES cell self-renewal network.

Furthermore, high levels of these enzymes have also been reported to block DNMT access and facilitate the conversion of 5mC to 5hmC [Ito et al., 2010]. TET enzymes work by oxidising 5-methylcytosine to 5-hydroxymethylcytosine (5-hmC) while utilising oxygen, Fe(II) and a-ketoglutarate as substrates [Moen et al., 2015]. Oxidation can continue further, consequently converting 5-mC to (5-fC) 5-formylcytosine and then 5-carboxylcytosine (5-caC). Previous work involving a double knockout of TET1 and TET2 dioxygenases observed changes in nucleosome positioning, TF binding and more importantly DNA methylation

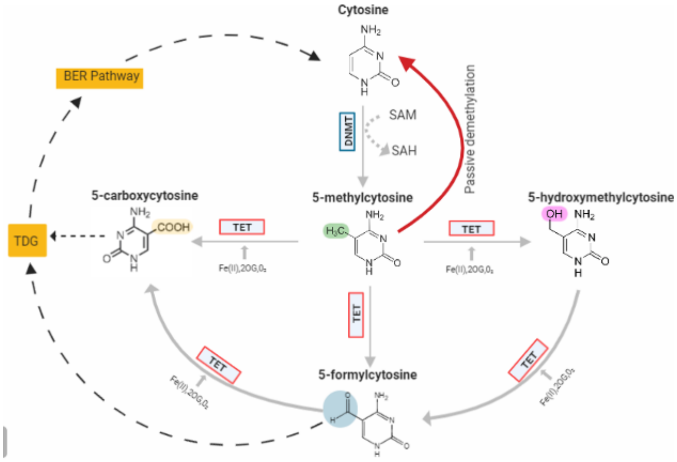


Figure 1.2: Overview of the mechanisms of DNA methylation and DNA demethylation. Cytosines contained in CpG rich areas may undergo the addition of a methyl group (from SAM) forming 5-methylcytosine catalysed by DNMTs. Passive dilution can occur by which modified cytosine is transformed to its original state-naked cytosine. Members of the ten eleven translocation enzyme family may oxidise 5-methylcytosine to form 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxycytosine using cofactors Fe(II), 2OG and O2. Thymine DNA glycosylase may recognize G/T mismatches and excise 5-formylcytosine or 5-carboxycytosine to produce naked cytosine through activation of the base excision repair (BER) pathway.

patterns, highlighting the functional role of TET's in this process [Wiehle et al., 2019].

Nucleotide excision repair has been considered in the context of DNA demethylation however a more widely accepted pathway is the base excision repair (BER pathway). Thymine DNA glycosylase capable of excising 5-fC prompts the BER pathway to convert modified cytosine to naked cytosine [Bochtler et al., 2017]. Interestingly, DNMTs have also previously been implicated in DNA demethylation upon absence of the methyl donor, SAM [Liutkeviačiūte et al., 2011] identifying many of the necessary molecules and key players in this process. We lack a complete mechanism of how DNA methylation is established, maintained and lost.

1.6 Assessing DNA methylation

Numerous methods are available for the identification of methylation states of DNA samples. Some methods allow for detection of epigenetic changes and others allow assessment of DNAm within particular genes of interest. Older techniques used for measuring DNAm aimed to measure DNAm quantity across specific regions of the genome however these methods did not provide the required specificity and genomic scope. Recent advances in methodological techniques by next generation sequencing (NGS) such as microarrays, whole genome bisulphite sequencing (WGBS) and bisulphite amplicon sequencing (BSAS) allow single base nucleotide resolution and CpG and non CpG sites across the genome [Masser et al., 2018]. Popular methods include the use of antibodies specific to 5-mC and 5-hmC, high-performance liquid chromatography (HPLC), ELISA based methods and pyrosequencing [Kurdyukov and Bullock, 2016]. However, many of these methods do not allow for distinction between cytosine and 5-mC unless bisulphite treatment is also applied. Sodium bisulphite treatment ensures deamination of cytosine into uracil allowing identification between naked and modified cytosine [Frommer et al., 1992]. However, it is worth noting that bisulphite treatment does not allow discrimination between different modified cytosine states such as 5-mC and 5-fC which could result in methylation patterns observed being confounded by signals from various forms of modified cytosine and different functions of these two states have been proposed previously [Song and He, 2013],[Gorrie-Stone, 2019]

Nanopore sequencing, a single molecule method for sequencing DNA does not however require bisulphite treatment due to its ability to distinguish 5-mC from the four standard DNA bases with high accuracy [Clarke et al., 2009]. However, whole genome bisulphite sequencing is considered the standard method to use as it covers 28 million CpG sites in the genome at a single base level providing required specificity and genomic scope, it is however a costly method [Pidsley et al., 2016].[Bibikova et al., 2009] describes a genotyping method adapted from the frequently adopted SNP arrays, the Infinium BeadChip microarrays. The first iteration of this method was the Illumina HumanMethylation27 (27k array) which was released in 2009 [Campagna et al., 2021]. This methylation array assayed methylation at more than 27,000 CpG sites across the genome (which represent less than 0.1% of the total CpG sites in the genome) of which covered around 14,000 genes. At the time, this array paved the way for the EWAS field as we know it today. With early studies using the 27k array to link DNA methylation changes to diseases and phenotypes such as ovarian cancer, breast cancer and smoking [Teschendorff et al., 2009],[Xu et al., 2013],[Breitling et al., 2011]. Following this, there was the release of the IlluminaHumanMethylation450k array, which covered an increased 450,000 sites across the human genome which includes coverage of 96% of CpG islands in the human genome and 99% of RefSeq genes [Wang et al., 2018]. This iteration of the methylation array is the most frequently used array in EWAS thus far. However, this is likely to change with the growing popularity of the Illumina HumanMethylation850 (EPIC) array, which measures nearly double the number of CpG sites as its predecessor (850,000 CpG sites). Its growing popularity can be attributed to its increased coverage in general, and also that of CpG sites that are not located in CpGI's, but located in open sea regions and enhancer regions. For example, it covers 58% of FANTOM enhancers, 7% of distal regulatory elements and 27% of proximal regulatory elements [Moran et al., 2016a]. Meaning the EPIC array allows for exploration of regions thought to be previously 'non functional' regions, thus allowing for a further understanding of the methylome in various tissues through utilisation within epigenome wide association studies. [Moran et al., 2016b],[Pidsley et al., 2016].

1.7 Epigenome wide association studies

Similar to genome wide association studies (GWAS), epigenome wide association studies (EWAS) arose from the field of epidemiology. The aim of an EWAS is to perform an assessment of epigenetic variants, most often DNA methylation genome wide, in order to identify epigenetic signatures that are associated with a particular phenotype or disease of interest. Methylation is measured using the aforementioned methods at CpG sites and then later compared in relation to the phenotype or disease of interest. This results in the identification of differentially methylated positions (DMPs) or differentially methylated regions (DMRs). A DMP is a single CpG site which shows differential methylation in relation to your exposure of interest, whereas a differentially methylated region can involve several CpG sites spanning a short or large region of the genome. The exact definition of what constitutes a DMR differs between research studies and detection methods but most often they are associated with either a specific gene or functional region such as a CpG island in a promoter region of a gene. DMRs are thought to be more biologically and functionally relevant for gene expression due to the strong correlation amongst neighbouring CpGs, although this is a debated area of the field.

Epigenome wide association studies can be conducted in a number of ways, such as through observational designs such as cohort and case-control designs. The latter are the most widely adopted study design due to the cost effectiveness and accessibility of these types of research studies. A case control study essentially involves arranging unrelated samples into two groups with relation to your phenotype of interest. For example, we can consider the case of an EWAS of smoking status. Here, we group our samples into those who smoke, and those who are non smokers. Put simply, we then would compare CpG methylation levels between these two groups in order to identify differentially methylated positions or regions in relation to smoking behaviours. It is important to note that at this step, we would also need to consider covariates, which I will discuss later in the chapter in more detail. One caveat of this type of design, is the inability to discriminate whether the DNA methylation changes are causative or just indirect effects of disease. In this case, longitudinal designs may be more appropriate, although still inconclusive.

Cohort designs effectively measure CpG methylation and phenotype. Longitudinal studies capture change over time These are less common due to the increased cost and difficulty of recruitment for these types of studies. Despite this, they are highly informative studies, as the epigenome is highly susceptible to change over time, in response to the environment and ageing [Snir et al., 2019],[Kulakova et al., 2016],[Martino et al., 2011],[Bjornsson et al., 2008],[Madrigano et al., 2012],[Wikenius et al., 2019]. Moreover, the longitudinal nature allows for a clearer picture of how the epigenome can be remodelled over time in response to phenotypes and disease.

Lastly, there has been an expansion of EWAS in replication and validation EWAS. Essentially, these types of EWAS involve two independent cohorts to confirm preliminary results and effect size and direction. These are proving necessary, as it is well known that the signals from EWAS can contain large amounts of noise, originating from confounding factors such as environmental exposures. Obtaining two independent cohorts can prove quite difficult for some researchers, especially in the case of investigation of rare diseases or complex exposure measures (such as air pollution or diet).

1.8 Considerations for epigenome wide association studies

Expanding on this, it is necessary to also highlight some important considerations and underlying assumptions for epigenome wide association studies which i will briefly summarise here.

1. **EWAS do not indicate causality:** EWAS are only sufficient to describe a relationship between DNA methylation and a phenotype or disease. Therefore, they cannot explain a direction of causation. In order to explore that idea further, techniques such as Mendelian randomisation may be employed [Burgess et al., 2020].

2. **Coverage may not be sufficient for research hypothesis:** This will depend on the method used to measure methylation across the genome, however, CpG sites measured may not be sufficient to explain disease state

or phenotype response. For example, CpG sites on the Illumina 450k array were 'cherry picked' as sites predicted to be functional or of of biological importance. Therefore, array based measurements may not be sufficient for complex diseases such as neurodevelopmental disorders or explaining phenotypes such as depression etc. If this is the case, it may lead to no relevant results being detected, potentially leading to an incorrect conclusion that the disease or phenotype has no significant relationship with DNA methylation.

3. **Consideration needs to be given to the choice of tissue** It is well established by now that the methylome varies largely between cell and tissue types. Therefore, it is of utmost importance in study design to consider whether the tissue one is studying will be of relevance to the disease or phenotype of interest. For example, if we consider the case where we are interested in studying the link between depression and DNA methylation signatures, we may ideally want to measure methylation in brain tissue, as we could assume this would be the most biologically relevant tissue for this research question. However, the limitations of this come from sample collection, so it is not always feasible to find a large sample willing to provide relevant tissues for research. Despite this, we should note that it is still possible to identify DNA methylation based biomarkers in less relevant tissues such as peripheral blood.

4. **DNAm may not be sufficient to explain or predict disease or phenotype** Often, it is assumed that DNA methylation changes are always functional. Moreover, frequently in EWAS, effect sizes although statistically significant, are often small, suggesting that disentangling the importance of DNAm may be difficult to distinguish. This makes it difficult to determine the biological function, if any, that these epigenetic changes identified in EWAS have.

1.9 Aims of this thesis

1. Provide an overview of the fundamental concepts that underlie this thesis.

2. Provide an overview of the materials and methods used throughout this thesis.

3. Describe an epigenome wide association study between exposure to ambient and traffic related air pollution and DNA methylation from a cohort of 1168 participants from Understanding Society: the UK household longitudinal study. This analysis aimed to delineate the different ways in which environmental exposures may influence the human epigenome by using different measures of environmental exposures.

4. Describe an epigenome wide association study between biological sex and DNA methylation from 3642 participants from Understanding Society: the UK household longitudinal study. This analysis aimed to identify a robust catalogue of CpG sites displaying sex differences in DNA methylation and elucidate characteristics of these sites.

5. Identify CpG sites across the genome showing highly variable or highly stable DNA methylation patterns and identify how these sites may play a role in biological pathways and or chromatin organisation. I also aimed to identify epiallele like sites in human whole blood using a bioinformatics approach.

Chapter 2

Materials and methodology

2.1 Materials

The work presented in this thesis is based on several different data sources that utilise publicly available, and mostly population based samples. These datasets are summarised in table 2.1 and detailed below.

2.1.1 Understanding Society

The main data source forming this thesis was obtained from the UK Household longitudinal study: Understanding Society [Bao et al., 2022] (Understanding Society). Understanding society is a longitudinal study based in the UK consisting of 40,000 households. The study aim is to collect information for researchers and policymakers on the dynamic nature and stability of individuals lives. In wave 3 of the study (2011-12) whole blood samples were collected from a subset of the study

Table 2.1: Datasets used throughout this thesis

Cohorts	**Study**	**Description**
Understanding Society (Discovery)	Chapter 3 : How does the environment affect our epigenome? Chapter 4: The role of DNA methylation in autosomal sex differences Chapter 5: Identifying interindividual variation and stability of DNA methylation	Tissue: Whole blood Platform : EPIC array (n=1171)
Understanding Society (Validation)	Chapter 4: The role of DNA methylation in autosomal sex differences Chapter 5: Identifying interindividual variation and stability of DNA methylation	Tissue: Whole blood Platform : EPIC array (n=2471)
GSE120312	Chapter 4: The role of DNA methylation in autosomal sex differences	Tissue: Whole blood Platform: RNA sequencing (n=20)
SHIP-TREND cohort	Chapter 5: Identifying interindividual variation and stability of DNA methylation	Tissue: Whole blood Platform: HumanHT-12 v3 Expression BeadChip (n-=991)
GSE197305	Chapter 5: Identifying interindividual variation and stability of DNA methylation	Tissue: Cortex Platform: EPIC array (n=1221)
GSE171140	Chapter 5: Identifying interindividual variation and stability of DNA methylation	Tissue: Skeletal muscle Platform: EPIC array (n=160)

participants. Individuals were considered eligible to give a blood sample if they were over the age of 16, consented to blood sampling and genetic analysis, and participated in all annual interviews between 1999 and 2011. Our study population was restricted to participants of white ethnicity. A full description of the data set and data processing has been described by [Hughes et al., 2018] but Illumina Infinium MethylationEPIC BeadChip DNAm data was collected from this data, along with genetic information. Due to two independent rounds of funding, I obtained a discovery and validation cohort from this study. Following quality checks of the data, our final discovery data set consisted of 1171 participants (males =489, females = 686) and 2471 (males = 1135, females = 1345) participants for validation. The age ranges for each data set were 28-98 years old and 16-99 years old respectively.

2.1.2 Publicly available data sets used

Additional DNA methylation data sets

The study presented in chapter 5 used data from two additional studies. Firstly, I obtained data collected as part of the Brains for Dementia cohort [Francis et al., 2018]. This study is a planned brain donation programme spanning five brain banks in the UK with one donation point. An overview of the data set which I accessed through GEO under the accession number *GSE197305* (see Table 2.1) has been previously described by [Shireby et al., 2022] but briefly contained 1221 samples. Secondly, I utilised publicly available data measuring skeletal muscle methylation using the EPIC array in 160 samples. This data set was published as an extension of the following research study [Voisin et al., 2020] and is available under the accession number *GSE171140* (see Table 2.1).

RNA-sequencing data

Chapter 4 of this thesis also utilised publicly available data from GEO under the accession number *GSE120312* (see Table 2.1). This data provided gene expression profiling by RNA-sequencing in human whole blood for 20 healthy donors (10 males and 10 females) using the SOLid 5500xl system. Specifically, we used the pre-processed count matrices.

Microarray data

The study in chapter 5 also used data collected as part of The study of health in Pomerania (SHIP-Trend). This study is also a longitudinal population based study based in Germany, aiming to assess common diseases and their relevant risk factors. Examinations took place from 2008-2012 and originally involved 4420 participants. Gene expression levels for a subset of these participants were measured using the Illumina HumanHT-12 v3 BeadChip array from whole blood cells (n=991). Details of the pre-processing methods have been previously reported here [Schurmann et al., 2012].

Hi-C data

Lastly, chapters 3-5 also used Hi-C data obtained using the GEO accession number *GSE124974* obtained from 2 healthy individuals. Hi-C library preparation was performed on white blood cells and neutrophils and the details have been previously described here [Liu et al., 2019].

2.2 Epigenome wide association studies

2.2.1 DNA methylation measurements and quality control

DNA methylation measurements were obtained from DNA extracted from human whole blood samples. Genome-wide DNA methylation data was measured using the Infinium HumanMethylationEPIC BeadChip array. Raw signal intensities were processed using the R package bigmelon [Gorrie-Stone, 2019] and wateRmelon [Pidsley et al., 2013] from idat files. Prior to normalisation of the data, outlier samples were identified using principal component analysis and subsequently removed from the data set. The reported age of the participant from which each sample was obtained was compared to predicted age using the epigenetic age method implemented by agep in the R package bigmelon [Gorrie-Stone, 2019]. Further, the participant's reported sex was checked in the sample using a DNA methylation-based sex classifier [Wang et al., 2021] which predicts sex based on the methylation difference of X and Y chromosomes. 4 samples were subsequently removed from our discovery data set and 9 samples were removed from our validation data set, as reported and predicted sex did not match. The data were then normalised via the interpolatedXY adjusted dasen method implemented in the R package, wateRmelon [Pidsley et al., 2013].

2.2.2 Identifying sex associated autosomal differential methylation

Linear regression models were used to assess the relationship between autosomal DNA methylation and biological sex using methylation as an outcome. First, following normalisation of the data, SNP probes, cross hybridising probes [Che,] and X or Y linked probes were removed from the data set. The final discovery and validation data set consisted of 1171 and 2471 samples, respectively, and 747,302 DNA methylation sites.

Further, as whole blood is a heterogeneous tissue and contains different cell types, individual samples will have different cell type proportions which may confound analyses. To ensure that whole blood cell composition did not differ significantly by sex and would not introduce bias to our results, the relative proportions of Granulocytes, mononuclear, natural killer, CD4T, CD8T and B cells were estimated for all samples using the estimateCellCounts function implemented in bigmelon [Gorrie-Stone, 2019] which utilises an implementation of the method developed by [Houseman et al., 2012] for cell type deconvolution. The estimation is based on epigenetic data and expected DNA methylation signatures at specific loci in each cell types are used to estimate cell type composition. To assess whether the sex differences I observed were age independent, I performed a Mann-Whitney U test between the age distribution of males and females. Our results confirmed that there is no statistical difference in age between our male and female samples for our discovery data set (p value 0.07; median values of 60 and 58, respectively) and also for our validation data set (p value 0.26; median values of 52 and 51, respectively).

Sex associated autosomal differentially methylated positions (saDMPs) were identified by performing linear modelling using the limma package in R [Ritchie et al., 2015] using sex annotation and Beta values while adjusting for age, cell type proportions and batch effects. Correction for multiple testing was performed with the Benjamini-Hochberg false discovery rate method (FDR values). I further used the Bayesian method for controlling p-value inflation using the R package bacon for both our discovery and validation data sets [Morris et al., 2014]. A probe was considered significantly differentially methylated if the difference in Beta values between males and females was greater than 0.05 in either direction and the FDR value was smaller than 0.05 in both the discovery and validation cohort. I applied a very stringent cut off by only considering a saDMP to be validated if it met these two criteria in both the discovery and validation data set. I further characterized differentially methylated regions (DMRs) by applying the DMRcate function from the R package ChAMP to detect DMRs between males and females on the autosomes. A DMR was considered to be significantly associated with sex (saDMR) if it consisted of at least 5 CpG sites with a maximum difference in beta values between males and females greater than 0.05.

2.2.3 Identifying pollution associated differential methylation

Linear regression models were also used to assess the relationship between DNA methylation and exposure to air pollutants. I searched epigenome wide for pollution and distance to road associated DMPs by constructing linear models using the limma package in R using age, sex, cell type proportions and smoking status as covariates [Ritchie et al., 2015]. I corrected for multiple testing by using the Benjamini-Hochberg false discovery rate method. In the case of the pollution associated DMPs, I fitted a separate model each of the four pollutants; NO_2,$PM_{2.5}$, PM_{10} and O_3 using an average exposure estimate from a 5 year period prior to DNA methylation measurement as estimated by our models described in section 3.4.3. Lastly, I used an individuals minimum distance to a busy road as a measure of TRAP to calculate distance to road associated DMPs. A CpG was considered to be a significant road or pollution associated DMP if it had a corrected p value lower than 0.05.

Road associated differentially methylated regions (raDMRs) were identified by performing linear modelling using the limma package in R [Ritchie et al., 2015] using pollution exposure measures and Beta values while adjusting for age, smoking status, cell type proportions and batch effects. Correction for multiple testing was performed with the Benjamini-Hochberg false discovery rate method (FDR values).

2.3 Identifying variable and stable sites on the Illumina EPIC array

To identify variably methylated and stably methylated probes on the Illumina EPIC array across individuals, I applied a down sampling approach. To measure variability and stability, I calculated standard deviation (SD) of methylation values on each probe across all individuals in our discovery data set and those with the top 10% highest SD values were labelled as the highly variable probes, and those with the top 10% lowest SD values were labelled as the highly stable probes. To ensure I detect robust variable and stable sites, I then down sampled the data set by 10% and kept only the probes which appeared in all down sampled data sets. This analysis was then repeated in our validation data set and only the replicated probes were carried forward for further analysis. This provided us with a highly robust catalogue of CpG sites which had either highly variable or highly stable DNA methylation patterns.

2.4 Genomic annotation of CpG sites

I annotated CpGs using the manufacturer supplied annotation data (MethylationEPIC v-1-0 B2 manifest file). Annotation was completed in the R package Minfi [Aryee et al., 2014]. Several categories were used as annotations in relation to CpG islands and divided into the following categories: CGIs, CGI shores (S and N), CGI shelfs (S and N) and open sea regions. Further, I also annotated the autosomal CpGs to several genomic features, including exons, introns, 5' UTR, 3'UTR, enhancers, promoters and transposable elements (TEs) using data from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables).

2.5 Enrichment of CpGs in transcription factor binding motifs

The enrichment analysis of known motifs at CpG sites was performed using the R package PWMEnrich [R, 2020] using the MotifDb collection of TF motifs [P and M, 2021]. Specifically, for the analysis presented in the chapter titled *The role of DNA methylation in autosomal sex differences* the DNA sequences within a 100 bp range from the saDMP which were female biased CpGs were extracted from the genome and compared to the saDMPs which were male biased CpGs as the background to reveal unique enriched motifs (adjusted p-value $<$ 0.05, FDR). Similarly, in the chapter titled *Identifying interindividual variation and stability of DNA methylation*, the DNA sequences within a 50 bp range from the VMPs were extracted from the genome and compared to the SMPs as the background to reveal unique enriched motifs (adjusted p-value $<$ 0.05, FDR).

2.6 Gene ontology analyses

GO analyses were conducted using the gometh function in the missMethyl package [B et al., 2016] which tests gene ontology enrichment for significant CpGs while accounting for the differing number of probes per gene present on the EPIC array. The background list used in these annotations was specific to the EPIC array also. For GO ontology analyses of enriched transcription factor motifs I used enrichGO from the clusterProfiler package in R [Yu et al., 2012], which performs FDR adjustment.

2.7 Protein-protein network visualisation and hub gene identification

I searched all of the genes annotated to our CpGs using the Search Tool for the Retrieval of Interacting Genes (STRING) (https:://string-db.org) database to generate our networks. I extracted protein-protein interactions with the default value of a combined score of >0.4. Following this, I utilised the cytoscape plugin tool Cytohubba (Chin et al., 2014) in order to identify hub genes within the networks. This was done by employing the local based method called maximum clique centrality (MCC). The same analysis was applied to the enriched transcription factor motifs found at CpGs.

2.8 Integration with gene expression

RNA-seq data for 20 healthy donors (10 males and 10 females) from publicly available data from GEO discussed in section 2.1.2 (GSE120312) was used in chapter 4 of this thesis. Firstly, in chapter 4, I explored the saDMPs in association with the expression levels of their annotated genes in whole blood in order to investigate how DNAm changes might lead to changes in gene expression. Specifically, I used the pre-processed count matrices with DESeq2 [Love et al., 2014] to calculate differentially expressed genes between males and females with an FDR adjusted p-value of 0.05 and log2 fold change of 1. DESeq2 does apply an automatic filtering step to remove genes with low counts but I did also apply our own independent filtering to this data by removing genes that do not have counts of at least 10 in all samples. Next, in chapter 5 I investigated the correlation between methylation variation and target gene expression variation at epialleles in microarray data. Details of the datasets used can be found in section 2.1.2. Again, I used the pre-processed count matrices [Love et al., 2014] and applied normalisation using the '*normalizeBetweenArrays*' function from the limma package in R [Ritchie et al., 2015]. Then, I calculated the coefficent of variation for each gene as a measure of variance. The coefficient of variation was calculated by dividing the standard deviation of its expression value by its average expression value across the sample population. Following this, I then calculated pearsons correlation between methylation variation at each CpG and gene expression variation of its associated target gene.

2.9 Overlap of CpGs with chromatin loops

I examined whether any of our CpGs of interest made 3D contacts with distal genes using Hi-C data available from the GEO under accession number (GSE124974) for white blood cells and neutrophils)see section 2.1.2). Hi-C library preparation was performed using the Arima-HiC kit and pre-processing of the data was performed using Juicer command line tools [Durand et al., 2016]. Reads were aligned to the human (hg38) genome using BWA-mem[Li and Durbin, 2010] and then pre-processed using the Juicer pre-processing pipeline. I called chromatin loops using the HICCUPS tool from Juicer using a 10 Kb resolution. I then constructed GenomicInteractions objects to annotate CpGs to loop anchors using the findOverlaps function from the GenomicRanges package using a maxgap of 10000. Following this, I then annotated the corresponding anchor to the relevant gene ID.

2.10 Annotation of VMPs and SMPs to methylation quantitative trait loci and their occupancy in topologically associated domains

To investigate the proportion of VMPs and SMPs described in Chapter 5 under genetic control, I annotated them to mQTLs using data from a previously reported study [Hannon et al., 2018]. Briefly, Understanding Society samples were genotyped using the Illumina Infinium HumanCoreExome array comprised of over 250,000 genome-wide tagging single-nucleotide polymorphisms (SNPs). This dataset was then subsetted according to those samples for whom there was also matched methylation data, details of the pre-processing of this data have previously been reported here [Hannon et al., 2018]. This data was then used in order to calculate DNA methylation quantitative trait loci (mQTL). In total, 766,714 DNAm sites were tested against 5,210,475 genetic variants while adjusting for age, sex and cell type proportions using the R package MatrixEQTL [Shabalin, 2012]. This resulted in the identification of 12,689,548 significant mQTL associations. I further characterised these into cis, trans and cis and trans mQTLs. I defined cis mQTLs as cases where the SNP and CpG site were located within 500bp of each other and trans as cases where the SNP and the CpG site were greater 500bp apart. Several CpG sites were annotated to more than one SNP, where one was a cis mQTL and another was a variable mQTL, thereby I classified these as cis and variable mQTLs. To annotate the mQTLs SNPs to biological processes I performed KEGG and GO enrichment analyses as previously described (see Sections 2.4 and 2.6).

Next, I examined whether the mQTL pairs occupied the same TAD or were connected by loops using Hi-C data available from the GEO under accession number (GSE124974) for white blood cells and neutrophils. Chromatin loops were calculated as described in Section 2.9. Following this, I then investigated whether the corresponding chromatin loop anchor was annotated to the SNP in the relevant mQTL pair. I also analysed whether our SNP and CpG pairs annotated by the mQTL analysis occupied the same topologically associated domain by using the *findOverlaps* function too. To do this I generated contact domains at a 10kb resolution using the hic-FindTADs tool from HiCExplorer [Ramírez et al., 2018]. These steps then allowed us to calculate the percentage of mQTL pairs which were connected via loops or occupying the same TAD compared to the EPIC array background. Fishers exact test was used to determine whether the differences were statistically significant. Lastly, Hi-C maps were generated using the hicPlotTADs function from the command line tool,HiCExplorer [Wolff et al., 2022].

2.11 Characterising epialleles

Chapter 5 of this thesis describes a characterisation of epialleles in human whole blood. First, I identified those VMPs which had an average intermediate methylation across our sample as characterised by an arbitary average beta value between 0.40 and 0.60. This step was applied in line with previous studies, which use intermediate interindividual variability as an indication of epiallele hallmarks. Following this, I employed Hartigans dip test [Hartigan and Hartigan, 1985] using the diptest package in R to test for unimodality across these sites. I characterised epialleles as sites which had variable intermediate methylation and bimodal distribution (hartigan dip test, p val $<$ 0.05). We used a measure of bimodal distribution to represent the idea that the variable nature of the CpG was actually due to an 'on or off' epigenetic state amongst our sample population. This is an idea which has previously been described in similar studies [Marttila et al., 2021, Nikolaienko et al., 2022].

Chapter 3

How does the environment influence the human epigenome?

3.1 Introduction

The association of the environment with health has been widely recognised and there is well documented and evolving epidemiological evidence inferring a link between air pollution exposure and an increase in mortality and morbidity. Approximately 7 million deaths each year have been attributed to exposure to air pollution due to the sequential onset of diseases such as asthma, cardiovascular disease, stroke and the development of malignancies following exposure. Air pollution is the largest environmental risk factor and affects all regions, socioeconomic groups and age groups [WHO, 2016]. Certain geographical regions are still adversely affected, with South East Asia and the western Pacific region observing over 2 million deaths per year, compared to 500,000 deaths per year in the European region [WHO, 2016] Nevertheless, it is important to note that 91 percent of the world's population reside in areas in which the WHO air quality guidelines are not met. Exposure to air pollution has been demonstrated to impact our epigenome, and more specifically to alter our DNA methylation patterns [Rider and Carlsten, 2019]. However, the molecular mechanisms by which this association occurs is yet to be delineated. Furthermore, majority of the current studies focus only on acute effects and fail to consider socioeconomic factors which ultimately affect our exposure and response to air pollution [Hughes et al., 2018]. This was further demonstrated in a meta-analysis focusing on how socioeconomic stress is linked to general health in which DNA methylation is thought to play a role, especially in the aspect of mental health [Wood et al., 2020]. These aspects are crucial to furthering our understanding of how exposure to air pollution modulates DNA methylation changes, health and well-being and consequently help to build preventative policies.

3.1.1 Air pollution

Ambient air pollution is a combination of gaseous components such as nitrogen oxides, benzene, and particulate matter. Air pollution is typically divided into two categories: outdoor air pollution and indoor air pollution. Outdoor air pollution arises from sources such as the burning of fossil fuels, ground level ozone, tobacco smoke and gases such as nitrogen oxides, sulphur dioxide and carbon monoxide (CO). Fossil fuel combustion, primarily from TRAP (traffic related air pollution) in urban environments is the major source of ambient air pollution [Kelly and Fussell, 2015]. Indoor air pollution involves the exposure to particulates, asbestos, mould, and gases such as CO and radon. Due to the increase in population size and urbanisation, air quality is beginning to be considered as a vital determinant to public health, with increased focus on those who are more disadvantaged and less able to control or change the environment in which they live [Holgate, 2017]. Rural areas are also subject to dangerous levels of air pollution (specifically indoor air pollution) through the burning of biomass fuels and these areas also experience increased ozone levels (due to less depletion of ozone concentrations through the action of nitrous oxides found in TRAP. Ozone forms at ground level through interaction of NO2 with polyaromatic hydrocarbons (PAHs) and has previously been demonstrated to induce DNA methylation of apelin in rat lungs leading to reduced protective signalling of the apelinergic system [Miller et al., 2018].

Previous work suggests that exposure to air pollution increases biological ageing, oxidative stress and inflammation causing the onset of lung cancer and cardiorespiratory diseases. Short term exposure to ambient air pollution was related to an increase in levels of interleukin-6, TNF receptor 2 and C-reactive protein, suggesting a possible link to development of cardiovascular disease[Steinvil et al., 2008]. Air pollution exposure has also previously been implicated in the reduction of telomere length, ultimately resulting in development of cardiovascular disease [Zhao et al., 2018] . A proposed mechanism is that exposure to air pollution results in increased replication rate of cells and an increased telomere loss during these replications due to the G-rich structure of telomeres being highly sensitive to oxidative stress [Collins et al., 2011]. A second mechanism suggested is that exposure to air pollution causes an increase in the number of leukocytes caused by increased inflammation and oxidative DNA damage affected particularly by diesel exhaust [Chen and Schwartz, 2008, Miri et al., 2019]. These are proposed mechanisms and future work is required to disentangle this process.

The majority of morbidity associated with pollution is thought to be due to PM2.5 which is composed of nitrates, ammonia, sodium chloride, black carbon, mineral dust and water. Particulate matter can be of varying sizes (including PM10) and due to their incredible small aerodynamic diameters they are able to lodge deep into the alveoli of the lungs, and at their smallest size even penetrate the lung epithelium where they can then enter the blood stream [Plusquin et al., 2017]. However, investigating the effects of PM is a difficult task due to its complex composition and considering other environmental factors along with PM exposure is necessary to improve our understanding of how pollution affects the epigenome. Furthermore, investigating the effects of air pollution exposure can itself be challenging because of the difficulty to determine the effect of a single element such as Ozone (O3) due to the complex mixture of elements in air pollution.

The exact mechanisms by which exposure to pollution affects our health remains obscure and further work is necessary to clarify pathways involved in health defects which are observed with both short- and long-term exposure to air pollution. Nevertheless, there are many other lifestyle factors (smoking status, diet, physical activity) which should be considered when analysing exposure following their implications on health [Alegría-Torres et al., 2011]. Studying specific cohorts' exposure to pollution is expected to help shed light on the relationship between the environment and our health and help to understand individuals' responses to such factors and onset of disease. The biggest gap in this area of research is the lack of consideration for several environmental factors and furthermore, the interplay of these factors on altering DNA methylation signatures.

3.1.2 Assessing air pollution

Measuring air pollution is necessary for exposure assessments and epidemiological studies which are carried out to benefit our understanding about pollution and its importance in health. There are several types of study designs used within this research, ranging from land use regression (LUR) models to studies carrying out controlled exposures. Land use regression models require variables such as road type, traffic count and land cover [Ryan et al., 2005] and are gaining interest as they are able to provide data for locations which are otherwise unmonitored, for example a participants home address or place of work. [Plusquin et al., 2017] demonstrated the use of LUR models to determine levels of PM10, PM2.5, NOx and NO2 for varying residential locations, yearly mean concentrations of those pollutants and to measure exposure of each cohort member at their residential address. LUR models are however often temporally limited and do not consider wind direction or emission data and fail to represent the variation in exposure levels of addresses found within metres of each other due to being located near heavy traffic. This could be considered a large disadvantage to consider when using this particular type of study design. [Freijer and Bloemen, 2000] reported that interestingly, exposure levels fall dramatically behind a row of uninterrupted buildings. [Arain et al., 2007] made efforts to include wind directions in LUR models and found slight improvement in predictions made by the model (R2 0.65 to 0.69).

Other studies use geocoding to estimate pollution exposure at a given latitude-longitude location. These estimates are often imputed at unmonitored locations and this method can be highly computationally intensive. There have been developments to facilitate this type of analysis, such as the R package PARGASITE which allows researchers to determine pollution exposures for specific geographic location and specific time points often utilizing air quality monitors [Greenblatt and Himes, 2019]. [Gondalia et al., 2019] used geocoding to facilitate estimation of PM exposures at daily and monthly intervals for 8,397 participants to explore the effect of PM exposure at DNAm and gene expression . However, current literature reports a bias and misclassification often observed in geocoding of streets due to errors in the process of geographically positioning areas such as homes and schools with a median error of 41 metres ultimately producing many false positive and negatives in results [Zandbergen, 2007].

Other options for studying air pollution utilise experimental designs. For example-controlled exposure studies are often longitudinal studies which involve participants experiencing different types of exposures. [Clifford et al., 2017] performed a randomized controlled exposure study in which participants were exposed to diesel exhaust particles for 2 hours and following a 4-week washout period they were then exposed to filtered air. This study aimed to investigate the effect of diesel exhaust exposure on bronchial epithelial DNAm. Results from both exposures were compared in order to determine the effects of each in relation to a particular disease, in this case allergic disease. These types of study design often offer accurate results since factors such as age which may often lead to misinterpretation of results are eliminated or reduced. Despite this, this design still fails to eliminate misleading results between participants due to socioeconomic factors such as wealth and access to health care.

Other research may evaluate air pollution exposure through comparing individuals who live in areas of high air pollution, to individuals who live in areas of low air pollution using monitoring stations. This design is not very accurate unless there is consideration for socioeconomic factors such as income, occupation and deprivation however this method does allow real life exposures to be observed rendering the results potentially more reliable. [Maghbooli et al., 2018] investigated placental adaptation to PM exposure by comparing individuals living in the most polluted region or Iran to those living in the least polluted region in Iran using monitoring stations of the Tehran air quality control company [Maghbooli et al., 2018].

Alternatively, some research studies depend on survey data collected from participants eliciting environmental and health histories in order to build histories on the participants' exposure levels. This method allows researchers to consider commuting habits, workplace exposures and exposures which may be determined by one's occupation which may greatly affect exposure levels. [Perera et al., 2009] used this approach combined with air quality monitoring in order to investigate prenatal exposure to polycyclic aromatic hydrocarbons (PAH's) and child intelligence. Furthermore, if an individual's location is known, kriging, inverse distance weighting or nearest distance methods may be applied in order to estimate exposures [Brauer et al., 2008]. Kriging can be applied in several forms; simple kriging, universal kriging or block kriging and has been previously used to estimate air pollution exposure.

3.1.3 The implications of air pollution exposure preconception and in utero

Exposure to air pollution has been associated with a wide range of health defects across all life stages. With a growing interest in whether early life conditions may influence health and Ill-being during adulthood, there is ever-growing evidence suggesting that exposure to air pollutants during embryonic development can result in epigenetic modifications such as DNAm changes in the placenta [Maghbooli et al., 2018, Gruzieva et al., 2019, Perera et al., 2009]. Studies investigating the effects of air pollution preconception, during pregnancy and early childhood show relatively little effect on the methylome [Forman and Finch, 2018].

Several studies have investigated the effect of air pollutants on male and female fertility [Conforti et al., 2018]. Recent evidence to suggest that NO2 and O3 are associated with impaired live-birth rates in IVF and high rates of miscarriage were also observed when experiencing high levels of PM10 [Legro et al., 2009]. A prospective study identified statistically significant associations between PM size and distance to road to incidence. infertility with PM2.5 and PM10 being most detrimental to fertility [Mahalingaiah et al., 2016] suggesting exposure to air pollution can affect fertility by causing defects during gametogenesis [Bains, 2021]. Despite this, no studies have been conducted focusing on the relationship between air pollution and DNA methylation in human sperm particularly as of yet; although it has been speculated that exposure to air pollution could result in DNA damage to sperm gametes therefore resulting in male infertility [Jafarabadi, 2007]. A review by Rider and Carlton, 2019 suggested that there is a global hypomethylation following air pollution exposure [Rider and Carlsten, 2019]. However, research from [Maghbooli et al., 2018] suggested that living in an area with more air pollution was associated with a higher placental DNAm and increased systolic blood pressure. Interestingly, they also found that exposure to air pollution can also reduce expression of DNMT1-a and increase expression of the methyl donor, SAMe contributing to greater passive dilution due to the decreased levels of DNMT1-a.

A great deal of research also focuses on studying DNA methylation in cord blood and placenta in order to decipher the relationship between exposure to air pollution and health in utero. Throughout pregnancy, the foetus experiences a time of plasticity in which DNA methylation plays a specific role. This is due to the occurrence of epigenetic ‘reprogramming’ during foetal development. Defects in this extremely fragile process may potentially have short- and long-term effects on the individual's health. Air pollution exposure during pregnancy has been consistently found to be associated with lower birth weight and impaired foetal programming [Burris and Baccarelli, 2014, Stieb et al., 2012]. A study by Breton et al, demonstrated that prenatal exposure to multiple pollutants in the first trimester of pregnancy was associated with lower LINE1 methylation levels in new-born blood spots whereas exposure to ozone later in pregnancy was associated with elevated levels of LINE1 methylation. This research also investigated single nucleotide polymorphisms (SNP) located in or near DNMTs, TETs and thymine DNA glycosylases (TDG) and subsequently found (after FDR adjustment for multiple testing) that 11 unique SNPs in four different genes showed significant association with a first-trimester pollutant and LINE1 methylation. Two of these SNPs found in DNMT genes also interestingly affected the susceptibility of children aged 11 years old to prenatal ozone induced changes in cardiovascular phenotypes [Breton et al., 2016]. It is important to consider that the changes in DNA methylation over time were not monitored and so it is possible that critical changes occurred at earlier points of foetal development, not captured by this study. Nevertheless, further research validates the results found from [Breton et al., 2016] in which exposure to air pollution resulted in DNA methylation of LINE1 and HSD11B2, this was especially pronounced in new-borns with foetal growth restriction [Cai et al., 2017]. Kingsley et al also reported that residing in locations in close proximity to major roads resulted in lower placental LINE-1 methylation levels and lower foetal growth [Kingsley et al., 2016].

Gruzieva and colleagues showed that exposure to NO2 during pregnancy is also associated with DNA methylation at 3 CpG sites in mitochondrial related genes (LONP1, HIBADH and SLC25A28). Analysis of several genes also revealed differentially methylated CpGs in catalase and thyroid peroxidase which have previously been recognized to be important in the maintenance of redox homeostasis [Gruzieva et al., 2019, Faria et al., 2019]. Previous studies have validated the idea that NO2 is capable to indirectly form ROS, cause inflammation and cell death

[Lodovici and Bigagli, 2011, Kelly, 2003]. A study focused on 120 participants from China living in either high polluted areas or lower polluted areas identified 371 DMR's using whole genome bisulfite sequencing. Most of these DMRs were located primarily in gene regulatory elements such as promoters and enhancers. Interestingly, gene enrichment analysis revealed that these DMR-related genes were also enriched significantly in diseases related to cancer and pulmonary disease. The DMR-related genes were also found to be involved in processes relating to cytokine production and mitochondrial assembly. Participants residing in areas experiencing higher air pollution were also found to show a higher mitochondrial DNA copy number. Following these results, a cytokine assay was performed which supported a link between increased levels of IL-5 and higher air pollution levels [Wang et al., 2020a]. Zhou and colleagues were also interested in the relationship between air pollution and oxidative stress as a potential mechanism for the development of malignancies and cardiovascular disease due to SOD2's role in protecting against damage from superoxide. This led to investigations into the effects of PM exposure on methylation of superoxide dismutase 2 (SOD2) in which they detected a significant association between SOD2 promoter methylation levels and PM exposure concentrations during the entire pregnancy. This association was particularly pronounced in the second trimester and seemed to be affected by SOD2 promoter methylation levels observed in the maternal blood [Zhou et al., 2019]. Since a link between air pollution exposure and risk of cardiovascular disease has been suggested, an effort to focus on prenatal exposure and placental mutation rate and DNA methylation was made by [Neven et al., 2018] Their results demonstrated an increase in overall placental mutation rate with greater PM2.5 exposure but contrary to [Gruzieva et al., 2018] they did not see any change with increasing NO2 exposure. Mutations were found to be present in key DNA repair and tumour suppressor genes suggesting air pollution exposure can affect DNA repair mechanisms in foetus' [Neven et al., 2018].

Earlier studies focused on polyaromatic hydrocarbons (PAH) in polluted air and their effect on respiratory diseases in which they discuss six genes which they identified to be hyper-methylated upon PAH exposure which were all aligned to a known gene (ACSL3). The known gene had functions linked to inflammatory and immune responses [Perera et al., 2009]. Previous studies also support claims that exposure to PAH induce DNA methylation changes [Alegría-Torres et al., 2011].

3.1.4 Air pollution exposure in children

Prenatal exposures are likely to persist into childhood and adolescence due to children's distinctive response to exposures. Children are more vulnerable to exposure of air pollution as their lungs are still developing. Lung development begins with the proliferations of alveoli and capillaries, until alveolar expansion begins around the age of 5-8 years old. Due to children's short stature, they also breathe air closer to the ground and have an increased air intake [Selevan et al., 2000]. [Breton et al., 2016] illustrated this through their findings in which they attempted to find associations between maternal multipollutant exposures and childhood outcomes such as supine heart rate, systolic/diastolic blood pressure and b-mode carotid artery ultrasound results. Most specific air pollutant exposures were not found to be significantly associated with any of these outcomes. However, they did observe that a 21-ppb increase in maternal NO2 exposure (during the third trimester of pregnancy) was associated with a higher systolic blood pressure in children aged 11 years old. Interestingly, previous research had also formed a link between maternal air pollution exposure and childhood systolic blood pressure (BP). The link in this case, was however observed to be due to high $PM_{2.5}$ and black carbon exposure rather than NO_2 [van Rossem et al., 2015] and this study also analysed a larger cohort of individuals. Nevertheless, it still shows support for this link between prenatal exposure and subsequent childhood health. In parallel with Breton and colleagues' findings on NO_2 and childhood BP outcomes, they also recognized that first trimester exposure was also associated with lower DNA methylation in LINE1. The latter association did however depend on the individuals genetic sequence variation in the DNMT's, TET2 and TDG genes. These findings suggest a plausible explanation for the differences seen in associations between prenatal exposure and childhood outcomes observed across studies. Since many mQTL associations have recently been identified [Hannon et al., 2018] which offers support for the idea that genetic variants are responsible for varying DNA methylation patterns within individuals. It is reasonable to consider that alterations in the expression of significant enzymes which are responsible for the harmonization of DNA methylation could be an underlying cause for the development of certain disease phenotypes.

Large scale epigenome wide meta-analysis performed by [Gruzieva et al., 2019] also demonstrated how methylation differences in several genes which occurred due to PM2.5 exposure during pregnancy can be conveyed to childhood. 20 CpG's were recognized to associated with various PM diameters (PM_{10} and $PM_{2.5}$), two of which mapped to cg00905156 and cg068499931. Pathway analysis revealed that these CpG sites mapped to FAM13A and NOTCH4 respectively, which are implicated in maintenance of respiratory health. Both CpG's were found to later be significant in 7-9-year olds. NOTCH4 expression was also detected to be much higher in those who experienced higher PM10 exposure, although the direction of the association was found to be rather inconsistent throughout their research [Gruzieva et al., 2019].

Due to established knowledge regarding the importance of TET enzymes in the DNA methylation process, it is no surprise that recent work has focused on associations between TET methylation, TRAP and childhood asthma. [Somineni et al., 2016] collected DNA derived from the nasal airway epithelial cells of 12 children with diagnosed asthma and their non asthmatic siblings. They found an increase in the global 5-hmC levels which were significantly associated with asthma, and furthermore a loss of methylation at cg23602092, a CpG in the promoter region of the TET1 promoter. The asthmatic children also displayed lower mean methylation. In addition, at the same single CpG site they also showed an increase in methylation due to TRAP exposure which were tissue consistent [Somineni et al., 2016]. This study was only focused on African American children and therefore further work should be considered in order to determine if results can be generalised to all ethnicity's. For that reason, using the Illumina 450K array and Pyrosequencing, several authors intended to replicate this work in two other cohorts. This research reported variation in DNA methylation in nasal DNA from sibling pairs and identified six CpG sites of which their methylation is associated with asthma. One of which was a CpG site (cg23602092) in the TET1 promoter identified in their previous study, and two others were CpG sites located within the promoter region of OR2B11 located upstream of a protein required for asthma phenotypes as previously demonstrated in mouse models [Zhang et al., 2018].

[Lovinsky-Desir et al., 2018] suggests that children with greater risk of impaired lung function due to high levels of air pollution may benefit from immunological benefits following physical activity. In their cross-sectional study in which they analysed children aged 9-14 in New York city they measured black carbon exposure and performed lung function assessments to attempt to draw significant associations between the two. Their research focused on 6 CpG sites located in the FOXP3 promoter as previous literature suggests that activation of FOXP3 maintains the differentiation and functionality of T-cells. Hypermethylation of CpG islands in the promoter region of FOXP3 leads to transcriptional silencing of FOXP3 which has been identified to contribute to development of several human diseases, but most importantly asthma [Marques et al., 2015]. Stratified analysis revealed that children exposed to higher black carbon levels who partook in physical activity had lower DNA methylation levels at FOXP3 promoter region. Hew and colleagues obtained similar results in their research investigating the effect of the exposure of polyaromatic hydrocarbons on epigenetic modifications leading to asthma. Increased methylation in FOXP3 locus was observed following high levels of PAH exposure which was also shown to impair Treg function. Increased IgE and IFN-y levels were also observed in concordance with impaired Treg function suggesting that PAH exposure is associated with diminished immunity via epigenetic modifications to FOXP3 [Hew et al., 2015].

Evidence supporting a link between maternal polyaromatic hydrocarbon exposure and childhood outcomes is further supported from a New York city study which focused on black or Dominican-American women who were categorized as non-smokers. Levels of exposure to PAHs were monitored from in utero until children were 5 years old. Several factors were accounted for such as maternal intelligence, tobacco smoke exposure and details of the home environment to avoid false positives within the results. Linear models were used in order to reveal an association between high exposure to PAHs and lower IQ scores suggesting that air pollution exposure may affect children's development [Perera et al., 2009]. Research based in the United States comparing student achievement in schools which installed air filters following a large gas leak to those that did not receive installation of air filters. A 0.20 standard deviation increase in test scores for English and Mathematics was found for students exposed to air filters suggesting air filters may be a simple, cost effective way to counteract neurological effects of air pollution exposure [Gilraine, 2020]. This study was not a randomised and it is possible that those schools with filters may have had more resources or money which allowed them to provide a better learning environment or higher quality / more experienced teachers.

An extensive list of literature offers support for a link between air pollution exposure and impaired lung function in children [Liu et al., 2012] but nevertheless, there is still ongoing research within this particular area in order to determine the mechanisms underpinning this association. Following the introduction of air pollution interventions such as Low emission zones (LEZ), [Mudway et al., 2019] undertook a study to examine the impact of these zones on childhood health. Annual pollutant exposures were measured alongside various lung function assessments, including FEV, FVC and assessment of respiratory or allergic symptoms. After adjusting for possible tobacco smoke exposures, a smaller lung volume was observed in children experiencing higher annual air pollutant exposures. Changes in lung growth were inferred from the association of FVC with long term exposure to NO2 and PM10 [Mudway et al., 2019]

3.1.5 Air pollution induced DNAm changes in adults

The vast majority of studies within this field have focused on DNA methylation in blood samples and on adults. This is due to fact that adult populations are easier to recruit and manage for such studies. A large number of associations have been made in adult studies, however most of these studies are only based on small cohorts and individuals who reside in areas of relatively low pollution compared to high polluted areas. Evidence seems to suggest a reduction in LINE1 methylation following high PM exposure. A study by Plusquin and colleagues studied the effects of long-term exposure of air pollutants on DNA methylation in two large cohorts; the ESCAPE project and the EPIC cohort [Plusquin et al., 2017]. Their aim was to focus specifically on average DNA methylation at functional regions and also individually methylated CpG sites. Their study focused on 454 Italian and 159 Dutch participants from the European Prospective Investigation into Cancer and Nutrition. In response to high exposure of NOx and NO_2, levels of DNA methylation in functional regions across the genome are lower, including CpG islands, shores and shelves and gene bodies. Furthermore, in Italy, exposure to higher annual averages of NOx and NO2 was associated with global hypomethylation, however in the Netherlands an association between NOx was not significantly present which could be due to higher levels of air pollution in Italy in general or due to population differences between the two cohorts. Meta-analysis did not result in observations of single CpG site level associations. Despite this, when analysed by country, they were able to identify numerous significant CpG sites. When they analysed methylation levels against gene expression levels of these particular sites, they were able to identify 5 pathways related to the immune system and its regulation for NO2 and NOx. This research did not consider confounding factors such as socioeconomic position which could have offered explanations for the absence of differential DNA methylation observed in single probes when analysing the two cohorts together.

Investigating differentially methylated regions may offer promising routes for exploring the effect of environmental exposures on DNA-methylation as one study demonstrates that there are associations between short term exposure to air pollution and DNA methylation at regional clusters and single sites of CpGs [Mostafavi et al., 2018]. This study included a panel of 157 healthy non-smoking adults living in four European countries. 24 hour personal and ambient exposures of PM2.5 are taken along with measurements of PM2.5 absorbance, ultrafine particles and peripheral blood samples taken. Personal exposure to high levels of PM2.5 was associated with methylation differences in 13 CpG sites and 69 differentially methylated regions. Out of these 13 sites, two mapped to the following genes: KNDC1 and FAM50B located in DMRs. DMRs were also identified to be associated with UFP and absorbance exposure. This study demonstrates an association but no attempt to identify the underlying mechanism was made, as is often seen in studies such as this. Moreover, it is worth considering that this study only accounts for short term effects and does not consider early-life exposures which may have an impact on associations found. A more recent study looked at the effects of daily and monthly mean PM concentrations over 2,7,28, 365 days and 1 and 12 months before blood samples were taken. Here they identified significant DNAm-PM associations at three CpG sites; cg19004594, cg24102420 and cg12124767. These sites are linked to the following genes, MATN4, ARPP21 and CFTR which have all been mapped to endocrine, neurological and pulmonary pathways, and cardiovascular disease related genes [Gondalia et al., 2019].

Nevertheless, some studies do investigate association between the effects of long-term air pollution exposure and DNA-methylation. Sayols-Baixeras and colleagues identified 81 CpGs associated with air pollution in REGICOR study. This research consisted of two large cohorts; 630 individuals from the REGICOR and 454 participants from EPIC Italy for validation. They screened for previously identified CpGs associated with air pollution in their study. 81 unique CpGs were identified in the EWAS which they could not validate, they also did not validate any previously identified CpGs in relation to PM and NO2. This study did control for many factors such as diabetes, hypertension, smoking status and cancer [Sayols-Baixeras et al., 2019].

Gao and colleagues explored the effects of environmental exposures on immune function by estimating leukocyte distribution using DNA methylation data from 774 White/European males from the Normative Aging Study. DNA methylation data was collected from blood samples by using the Illumina 450K array and leukocyte distribution was estimated using Horvath's algorithms [Horvath, 2013]. A 28-day period was used in which measurements of air pollution was taken and a higher mean PM2.5 exposure was associated with lower proportions of Natural killer (NK) cells, naïve CD8+ and plasma cells, but higher proportions of CD8+ T cells and naïve CD4+ T cells. This research suggests that methylation patterns could be useful for assessing the influences of air pollution on human health, more specifically morbidity could be achieved through detrimental effects on the immune system [Gao et al., 2019]. These findings can also offer support for the association between high PM exposure and the risks of cardiovascular disease and hypertension [Miller et al., 2007].

Although most studies focus mainly on particulate matter, [Samoli et al., 2008] offers support for a relationship between carbon monoxide exposure and cardiovascular disease and resulting mortality. This study focused on a very large cohort, over 40 million people across 19 cities in Europe as part of the APHEA-2 project. This research found significant associations of carbon monoxide with total and cardiovascular mortality. A 1.2 percent increase in total deaths occurred following a 2 day increase of mean carbon monoxide levels. This study also considered socioeconomic factors, and so coupled with the large cohort size these results are reliable [Samoli et al., 2008].

Exposure to particulate matter has also been reported to increase susceptibility to respiratory syncytial virus (RSV) infection by modulating the lung host defence to respiratory viral infections. Ultimately, this may lead to a greater chance of developing a lung disease [Harrod et al., 2003]. Several other studies have also formed a link between exposure to particulate matter and decreased lung function [Carlsten et al., 2011, Rice et al., 2016, Schultz et al., 2017] but most of these studies are focused on effects of exposure in children. However, [Clifford et al., 2017] were motivated to investigate the effects in adults, they performed a randomized cross over controlled exposure study in 17 humans and measured CpG methylation at single site resolution. During this study, they exposed individals to diesel exhaust. They observed significant methylation changes in 7 CpG sites at 48 hours, and 500 CpG sites in 4 weeks. Clifford and colleagues concluded that specific exposures could prime the lung for methylation changes as different methylation patters were seen according to different exposures [Clifford et al., 2017]. This research measured particulate matter in diesel exhaust, an approach also adopted by [Carlsten et al., 2016]. However, a limitation of this study is the small sample size. Despite this, later research from Carlsten and colleagues also detected a decrease in DNA methylation on promoter regions of genes following diesel exhaust exposure. This work suggests that air pollution exposure may induce cytokine changes as interleukin-5 and interleukin-6 levels specifically increased. These observations hoIver may only be extrapolated to allergic individuals, and those especially with the GSTT1 genotype [Carlsten et al., 2016].

Bind and colleagues assessed the effects of exposure to ozone and particulate matter on methylation of genes coding for tissue factor (F3), interleukin-6 (IL 6), intracellular adhesion molecules 1 (ICAM 1), toll-like receptor 2 (TLR 2) and interferon gamma (IFN y). The investigation was carried out on a cohort of 777 elderly men from the Normative Ageing Study. Methylation was investigated in the promoter regions of these genes and hypermethylation of these regions was observed in all genes generally. Negative associations were seen upon PM and black carbon exposure with F3 methylation, specifically an interquartile range increase in black carbon was linked to a 12 percent reduction in F3 methylation. Whereas sulphate and ozone exposure were negatively associated with ICAM 1 methylation [Bind et al., 2014]. The observation that air pollution exposure is associated with inflammatory processes is an emerging concept, claiming that obesity may increase susceptibility to the effects of particulate matter exposure via low grade inflammation. A study on 186 overweight individuals measured the mean methylation of CD14, TNF alpha and TLR4 and percent 5mC values for log transformed genes. All of which are important genes in inflammatory pathways. CD14 and TLR4 are necessary for recognition of LPS in the inflammation pathway, and also for activation of the immune system. TNF alpha is a cytokine produced by monocyte following inflammation which has been linked to age-related inflammatory diseases [Cantone et al., 2017]. A negative association between the methylation of the genes involved in inflammatory processes and PM exposure was found. Specifically, CD14 and TLR4 were associated with PM10 4 to 6 days and 6 to 8 days before, respectively. In contrast to other research, no association was found between TNF alpha and PM10 suggesting TNF alpha may be affected by other epigenetic marks such as histone modifications or non-coding RNA's [Cantone et al., 2017].

A controlled human exposure study also demonstrated that 15 healthy adults who were exposed to concentrated ambient particles or filtered medical air in a 2-week washout showed lowered Alu and TLR4 methylation in parallel with altered systolic/diastolic blood pressure in concordance with previous research [Bellavia et al., 2013, Breton et al., 2016, Wang et al., 2016]. Futhermore, exposure to woodsmoke has also been linked to cognitive dysfunction including in particular Alzheimers disease. Changes in DNA methylation patterns were Ill documented following exposure and have been shown to influence disease favouring inflammatory cascades, induce oxidative stress and modulate the response in vivo and in vitro following exposure. Research suggests that inflammation, oxidative stress and epigenetic modifications such as DNA methylation may link airborne pollution exposure and subsequent disease pathogenesis [Schuller and Montrose, 2020]. This research is further supported by longitudinal studies which studied the effect of long-term exposure to particulate matter PM2.5 on neurological disease in the American Medicare population which included 63,000,000 individuals. Throughout this research, they identified 1 million cases of Parkinson's disease and 3.4 million cases of Alzheimer disease and related dementia. There was strong evidence of linearity at PM2.5 concentrations less than 16 ug/m-3 followed by a plateaued association with increasingly larger confidence bands with hospital admissions. However, this research doesn't indicate disease onset as admissions for diseases such as these are normally found later on in the disease. Secondly, this does not uncover an underlying mechanism for this association or determine cause and effect [Shi et al., 2020].

3.1.6 Additional factors affecting DNA methylation and air pollution

It is well established that other factors such as physical activity, stress, diet, smoking and alcohol consumption can also influence epigenetic mechanisms such as DNA methylation, histone modifications and mRNAs [Quach et al., 2017, Alegría-Torres et al., 2011]. Several studies have been conducted to provide robust evidence for this idea, for example [Joehanes et al., 2016] and colleagues identified 2623 CpGs associated with smoking between current versus never smokers which were annotated to 1405 genes. Enrichment analysis with the genes annotated to these smoking associated CpGs revealed enrichment for pulmonary function, inflammatory diseases and heart diseases. With exposure to airborne carcinogens being more likely to occur in the presence of cigarette smoke, this suggests it is necessary to consider smoking status when estimating exposure to air pollution [Cho et al., 2014] as smoking status may influence exposure to airborne pollutants. A limited number of studies investigating air pollution exposure on DNA methylation consider smoking as confounding variable, but increasing evidence suggests that this is vital to correctly estimating pollution exposure. For example, particulate matter levels have previously been estimated to be higher in ‘street canyons’ when compared to high traffic areas [Ruprecht et al., 2016]. Street canyons (where the street is flanked by buildings on both sides) are areas which are limited to pedestrians only and are therefore often accompanied by a higher presence of smokers which potentially generate larger quantities of second-hand smoking). A study on 996 individuals in Salvador, Brazil who were exposed to cigarette smoke alone, household air pollution alone and dual exposure were investigated. It was identified that those individuals who experienced dual exposure presented poorer lung function, more severe asthma symptoms and poorer asthma control when compared to those who experienced exposure to cigarette smoke alone, or household air pollution alone. This study demonstrated the need to consider additional factors when investigating associations between air pollution exposure and health, as ‘double exposed’ individuals presented worse adverse health symptoms [Fernandes et al., 2018].

Recent evidence also suggests that exposure to high ozone and particulate matter levels are capable of activating the hypothalamic pituitary adrenal (HPA) axis triggering the release of the hormone cortisol which ultimately increases physiological stress and allostatic load [Kodavanti, 2016]. Disruption to the normal regulation of the HPA axis is commonly identified in several diseases such as depression, Alzheimer’s disease and type 2 diabetes [Du and Pang, 2015] making this an attractive mechanism to research. Social declivity in health that affect exposure to air pollution can be enlightened through study of allostatic load which suggests that socioeconomic position should also be considered when pollution exposure estimates are being drawn. Gonzalez et al suggested explanations as to why those of a disadvantaged socioeconomic position may be more susceptible to adverse effects of air pollution exposure. These are that they experience limited access to health care, more likely to be living in poor housing meaning they are exposed to air pollution more frequently and that they also adopt health behaviours such as smoking and drinking, which negatively impact their susceptibility [Rivera-González et al., 2015a].

Increased alcohol consumption as previously mentioned, may also be detrimental to mechanisms necessary for managing chronic exposure to air pollutants as found recently in a study on elderly Korean adults. Air pollution exposure was found to contribute to an increase in liver enzyme levels including alanine aminotransferase (ALT) and aspartate aminotransferase (AST). Interestingly, they found that the effect decreased during abstinence from alcohol (Kim et al., 2015) a finding which was also recently replicated in a cohort of 36,151 adults [Kim et al., 2019].

Some efforts have been made to investigate lifestyle factors which may mitigate the adverse health effects from chronic air pollution exposure. Studies such as these may assist investigations into the underlying mechanisms for air pollution's effects on epigenetic markers. The inhalation of PM2.5 was proposed to modify methylation patterns within CD4+ Th cells through inflammation and oxidative stress. Zhong et al conducted a cross over trial in which ten healthy adults were exposed to 2 hours of; sham under placebo, PM2.5 under placebo and PM2.5 with supplementation of vitamin-B. Methylation profiles were made before and after exposure using Illumina 450K array in CD4+ Th cells. They observed changes in methylation induced by PM2.5, with the top two loci being cg06194186 and cg17157498 which were located in the promoter region of the carboxypeptidase O gene and the promoter region of the NADH dehydrogenase ubiquinone Fe-S protein gene, respectively. Both of these genes are involved in mitochondrial oxidative energy metabolism. Four-week supplement of vitamin B reduced the detrimental effects of PM2.5 by 28-76% in the top ten associated loci [Zhong et al., 2017]. Despite this, a preceding study mentions the limitations of this research by highlighting the inappropriate dosage of B-vitamins administered to participants. Lucock et al, mention the potential health risks from pteroylmonoglutamic acid (PteGlu) when given in such high doses, and also notes previous research linking

B vitamins to an increased risk of myocardial infarction [Lucock et al., 2017].

In addition to vitamin supplement, folic acid supplementation is also thought to play a role in DNA methylation. Folate is required for the generation of methionine from homocysteine through reactions involving vitamin B12 and folate. A diet deficient in folic acid may therefore impede this reaction. Previous work has presented a link between long term air pollution exposure and increased likelihood for developing autistic spectrum disorder (ASD), but interestingly identified a role for folic acid in alleviating this risk. Mothers self -reporting consumption of folic acid levels above 800µg in the first month of pregnancy attenuated the risk of their child developing ASD when they were exposed to high levels of air pollution. Though these results were not found to be statistically significant following adjustment for confounding factors (such as financial hardship and zinc intake) ,they do offer support for future research to study diet supplementation as a method to alleviate harmful effects of chronic air pollution exposure [Goodrich et al., 2018].

3.1.7 Air pollution and accelerated ageing

Previous research suggests a link between air pollution and age acceleration, although currently limited [Ward-Caviness et al., 2016, Nwanaji-Enwerem et al., 2016b]. A study of non-Hispanic white women in the United States investigated the effects of particulate matter and NO2 on accelerated ageing by using Hannum and Horvath epigenetic clocks [Hannum et al., 2013, Horvath, 2013]. Exposure to NO2 was recognized to be inversely associated with age acceleration and associated with methylation of 2 CpG sites where as PM2.5 association with accelerated ageing depended on levels of exposure of each individual. For instance, in geographical regions with lower nitrate concentrations, PM2.5 was also inversely associated with age acceleration but there was no overall association between the two.

Results like these suggest a relationship between chronic air pollution exposure and accelerated ageing, an important indicator of morbidity. Epigenetic age clocks may be measures for investigating the effects of pollution on our health. Following research demonstrating that shorter telomere lengths are identified in parallel with lower levels of global DNA methylation [Zhao et al., 2018] speculations have been made that genome stability decreases following air pollution exposure, ultimately mediating the susceptibility to disease.

3.1.8 Limitations of the current research

While there have been many studies inferring a link between air pollution exposure and changes in the epigenome, there is no duplication of particular methylation sites which have been identified. To add to the complexity of research in this area, there are a number of factors which may be responsible for these differences. Firstly, DNA methylation signatures are altered throughout the life course and as explained, can be affected by several different lifestyle and socioeconomic factors. The interplay of these is also significantly important yet complex, meaning many researchers to date have not controlled for any or all confounding factors. This interplay between several factors also makes it hard for studies to determine causal effect from a single environmental factor. This highlights the important of intervention studies such as those from Zhong and colleagues which aim to determine mechanisms to mitigate harmful consequences of environmental factors. In doing so, pathways and mechanisms underlying the pathway to disease can be highlighted, thereby leading to more focused research in this field. However, these types of study designs are often only based on small cohorts which may lead to false negatives due to decreased statistical power and sample variability. Many study designs also focus on particular ethnic groups, age groups or sex. This poses difficulties in generalisation; for example, white elderly men may be more socially patterned to be at a higher risk for cardiovascular disease which could be the reason small associations are observed. Therefore, findings may not reflect cause and effect. Therefore, larger, and more diverse cohorts may be necessary to increase accuracy and reliability of results from such association studies and to warrant further research into this area.

Furthermore, the nature of air pollution makes this research very tricky and interpretation of research can often be even more complex. Particulate matter varies from region to region and is a mixture of different particles making the results from such studies extremely difficult to generalize to populations residing in different geographical regions. Moreover, most of the current research does not account for an individual's commuting habits or workplace exposures. While this is a necessary step, this method also means that exposures from commutes and the workplace are often ignored. Considering a vast majority of people spend lots of time away from their residential address, many studies miss an important aspect of data when computing pollution exposure here. This is especially important considering the vast amount of literature focused on workplace pollution and its detrimental effects on health [Edwards et al., 2001]. These models also often produce data for several pollutants such as NO2, PM2.5 and NOX for instance, any contributions of effects to health made from other pollutants are often missed in studies attempting to establish a link betIen pollution and health. To add to this, LUR models have previously been identified as imprecise in rural areas and islands, meteorological data should be improved in LUR models to increase precision and accuracy of exposure estimates drawn from them [Habermann et al., 2015]. There is also previous work demonstrating the ability of green spaces, tall buildings and street canyons to diminish levels of pollutants in the air, an aspect of 'real-life' which LUR models do not account for.

The nature of time between exposure estimates and sample analysis may offer reasons for the failure of replication of findings in this field [Rider and Carlsten, 2019]. Time dependent changes in methylation patterns of DNA in response to environmental pollutants have previously been discussed in this section and are not accounted for in several of these studies. Research focusing on long term exposure and long-term effects may be more important for this field.

Lastly, all studies mentioned in this review except one from Zhong and colleagues quantified methylation from blood samples and peripheral blood is a mixture of different cell types[Zhong et al., 2017] . Statistical adjustment for such sample types must be considered to account for the different cell types and proportions found. Further problems with methylation analysis are present in studies in which techniques are used which are not able to identify 5-hmC modification versions. Different modifications of cytosine pose different effects to genome stability and transcription factor binding, thereby having a different effect on gene expression and disease states. The vast majority of studies also mentioned in this review have not demonstrate intention to control for cell type correction.

3.1.9 Conclusions

Exposure to air pollution is linked to changes in methylation patterns of DNA and is important in morbidity from in utero to elderly age. Current research suggests that chronic exposure to air pollution has a larger effect on our health, especially in the case of lung function and development of cardiovascular disease when compared to short term exposures. Most of the research in this field has been non replicative, highlighting the complexity of this particular field of research. Despite this, efforts have been made to expose interventional treatments which guide research to important pathways for study. Future work should aim to determine whether air pollutant exposure induced DNA methylation changes are indicators of future disease risk or are merely a demonstration of the epigenomes temporary response to environmental stressors. Efforts are needed to define the underlying mechanisms by which exposure to air pollution increases morbidity and mortality around the globe as they are currently unclear. Following this, vulnerable populations can be targeted for intervention and policy discussions will be better informed. Most importantly, significant advances will be made when consideration is given to the interplay and impact of several lifestyle factors and exposures outside of the immediate home address to allow for a more complex and thorough assessment of the association between air pollution and health.

3.2 Methods

3.3 Development of air pollution exposure measures

3.3.1 Introduction of current methods and limitations

Air pollution continues to be a growing concern for public health across the world. Accurate estimates of individual exposure to ambient air pollution are crucial to delineating the relationship between long term exposure to air pollution and poor human health. There are several types of study designs used within this research, ranging from land use regression (LUR) models to studies carrying out controlled exposures [Ryan et al., 2005] and are gaining interest as they are able to provide data for locations which are otherwise un monitored, for example a participants home address or place of work. Plusquin and colleagues [Plusquin et al., 2017] demonstrated the use of LUR models to determine levels of some types of pollutants such as PM10, PM2.5, NOx and NO2 for varying residential locations, yearly mean concentrations of those pollutants and lastly, to measure exposure of each cohort member at their residential address [Plusquin et al., 2018]. LUR models are however often temporally limited and do not consider wind direction or emission data and fail to represent the variation in exposure levels of addresses found within metres of each other due to being located near heavy traffic. This could be considered a limitation when using this particular type of study design as [Freijer and Bloemen, 2000] reported that interestingly, exposure levels fall dramatically behind a row of uninterrupted buildings. Some studies [Arain et al., 2007] have included wind directions in LUR models and report slight improvement in predictions made by the model (R2 0.65 to 0.69).

Other studies use geocoding to estimate pollution exposure at a given latitude-longitude location. These estimates are often imputed at un monitored locations and this method can be highly computationally intensive. There have been developments to facilitate this type of analysis, such as the R package PARGASITE which allows researchers to estimate pollution exposures for specific geographic location and specific time points often utilizing air quality monitors [Greenblatt and Himes, 2019]. [Gondalia et al., 2019] used geocoding to facilitate estimation of PM exposures at daily and monthly intervals for 8,397 participants to explore the effect of PM exposure at DNAm and gene expression. However, current literature reports a bias and misclassification often observed in geocoding of streets due to errors in the process of geographically positioning areas such as homes and schools with a median error of 41 metres ultimately producing many false positive and negatives in results [Zandbergen, 2007].

Other options for studying air pollution utilise experimental designs, for example-controlled exposure studies are often longitudinal studies which involve participants experiencing different types of exposures. [Clifford et al., 2017] performed a randomized controlled exposure study in which participants were exposed to diesel exhaust particles for 2 hours and following a 4-week washout period they were then exposed to filtered air. This study aimed to investigate the effect of diesel exhaust exposure on bronchial epithelial DNAm. Results from both exposures were compared in order to determine the effects of each in relation to a particular disease, in this case allergic disease. These types of study design often offer accurate results since factors such as weight and age, which may often lead to misinterpretation of results are eliminated or reduced. However, it is important to note that these studies are not entirely reflective of an individuals real life exposure and so results should be considered with this in mind.

Alternatively, some research studies depend on survey data collected from participants eliciting environmental and health histories in order to build histories on the participants exposure levels. This method allows researchers to consider commuting habits, workplace exposures and exposures which may be determined by one's occupation which may greatly affect exposure levels. [Perera et al., 2009] used this approach combined with air quality monitoring in order to investigate prenatal exposure to polycyclic aromatic hydrocarbons (PAH's) and child intelligence.

Further, if an individual's location is known, kriging, inverse distance weighting or nearest neighbour methods may be applied in order to estimate exposures [Brauer et al., 2008] and these methods have previously been used in epigenome wide association studies [Son et al., 2010],[Lu and Mar, 2020],[Neven et al., 2018],[Li et al., 2018]. All of these methods are interpolation techniques, and work by using pollution values at known locations in order to estimate pollution values at unknown locations.

Nevertheless, modelling of air pollution data in the UK is a challenging task due to a number of reasons, first, the small number of monitoring sites located across this large region makes interpolation a difficult task. The second difficulty with pollution modelling within the UK is that there is a large amount of missing data for several pollutants due to faulty equipment, discontinuation of monitoring sites and in the case of particulate matter, no monitoring for several years. Table 3.1 summarises the percentage of missing monitored data yearly from the period of 2007 to 2011, for the period of interest [Mukhopadhyay and Sahu, 2018]. These difficulties could be reason for the sparse evidence supporting the idea that DNA methylation may play a role in underlying the relationship between air pollution and poor health.

3.3.2 Air pollution data sources

Air pollution data for the UK mainland is collected from the UK's automatic monitoring network which consists of 144 AURN (Automatic Urban and Rural network) stations. Monitoring is done for compliance with the European directive on Ambient air quality. Figure 3.1 shows the location of the 144

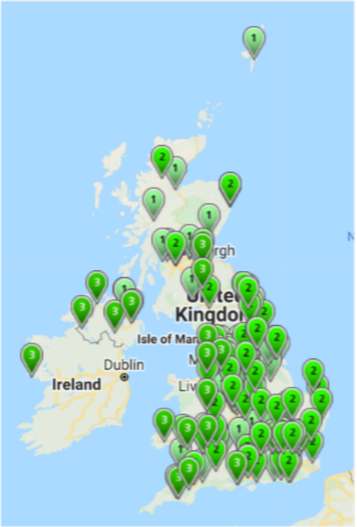


Table 3.1: Missing data (percent) for the four pollutants of interest from the period of 2007-2011.

Pollutant	**2007**	**2008**	2009	2010	2011	Overall missing data for study period
Nitrogen dioxide	39.16	39.23	38.39	38.09	35.47	38.07
Ozone	56.56	62.87	64.05	63.05	61.81	61.67
Particulate matter 10	64.99	66.71	70.24	72.41	69.78	68.83
Particulate matter 2.5	96.66	92.21	67.57	65.18	64.75	77.26

AURN monitoring sites within the UK (Interactive monitoring networks map - Defra, UK, 2020). Data collected at these stations is publicly available for download online. Table 3.2 shows an exhaustive list of parameters collected at these stations at a height between 1.5 and 4 metres from ground level. These sites provide the public with access to high resolution information which is taken on an hourly basis. There are currently 150 active sites under the AURN, with data available from the period of 1973-2020. However, due to the large percentage of missing data and small amount of sites across the study region, it would not be feasible to use this data alone to interpolate from. We therefore utilised models which included this data, as described in the next section.

3.3.3 Bayesian model-based estimates

To address the large amounts of missing data, Bayesian model-based estimates have recently been formed and used in order to study the long-term effects of the environment on health [Mukhopadhyay and Sahu, 2018]. There have previously been efforts to perform air pollution modelling using non-Bayesian methods which fail to account for temporal variation of pollutants and are mainly based in smaller areas such as cities [Gulliver and Briggs, 2011], [Wang et al., 2011]. Recently, a Bayesian spatiotemporal model incorporating data from both the AURN stations and also an atmospheric air quality dispersion model was formed in which concentrations of each pollutant was estimated for various latitude and

Table 3.2: Parameters collected from the AURN monitoring stations across the UK

Parameters collected
Ambient Temperature
Barometric pressure
Carbon monoxide
Daily measured PM10 (uncorrected)
Daily measured PM2.5 (uncorrected)
Modelled Temperature
Modelled Wind Direction
Modelled Wind Speed
Nitric oxide
Nitrogen dioxide
Nitrogen oxides as nitrogen dioxide
Non-volatile PM10 (Hourly measured)
Non-volatile PM2.5 (Hourly measured)
Ozone
PM10 Ambient pressure measured
PM10 Ambient Temperature
PM2.5 Ambient Preasure
PM2.5 Ambient Temperature
PM10 particulate matter (Daily measured)
PM10 particulate matter (Hourly measured)
PM1 particulate matter (Hourly measured)
PM2.5 particulate matter (Daily measured)
PM2.5 particulate matter (Hourly measured)
Rainfall
Relative Humidity
Sulphur dioxide
Total Particulates
Volatile PM10 (Hourly measured)
Volatile PM2.5 (Hourly measured)
Wind Direction
Wind Speed

longitude points within the UK [Mukhopadhyay and Sahu, 2018]. This particular model predicted concentrations of the 4 pollutants at the corners of 1-km grid squares and used this data to obtain estimated daily concentrations of NO2., PM10, PM2.5 and O3 from the periods of 2007-2011. I have used this data for the next steps of my analysis in order to benefit from accurate air pollution estimates at any spatial and temporal resolution within the study region.

3.3.4 Exposure estimates at postcode level

In this thesis, use of the methods mentioned above have been combined with further interpolation methods. This is because although the bayesian model estimates have greatly improved the exposure estimate resolution, it was still necessary to obtain postcode level measurements for each individual in our study. Methods such as Inverse distance weighting (IDW) and Ordinary Kriging (OK) can be extremely useful for this as it allows for concentrations of air pollutants to be interpolated to estimate individual exposure. In such cases, interpolation methods allow estimation of values of a particular variable at unsampled sites based on point observation data from other sites within the same region [Li et al., 2014a]. These methods are based on weighted averages and follow a general mathematical equation 3.1 and are therefore a mathematical best guess determined by the known values. The aim of this section is to discuss the development and use of IDW and Kriging models to estimate annual values for PM10, PM2.5, O3 and NO2 from the study period of 2007-2011 at given postcodes within the UK. Further, I evaluate the accuracy of both implemented models using cross validation techniques to decipher which technique should be used in the downstream analysis. It is important to note that the models used were capable of maintaining individual confidentiality as postcodes were converted to latitude and longitudinal data before use in the model. Inverse distance weighting (IDW) is a deterministic interpolation method which enables estimation of unknown values at particular locations based on known values at other locations in the same region using weighted values (Figure 3.2). Further, in

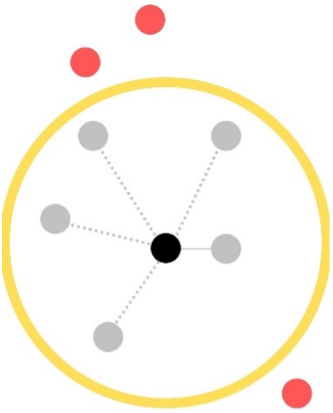


Figure 3.2: Inverse distance weighting interpolation illustration.

both models, we utilised data from the Bayesian model-based estimates, consiting of 500,000 known data points over a 5 year period prior to blood collection (2007-2011). The known data points were further aggregated into a 5 year average mean for use in our interpolation models.

$$x* = \frac{w^1x^1 + w^2x^2 + w^3x^3 + \ldots + w^nx^n}{w^1 + w^2 + w^3 + \ldots + w^n} \tag{3.1}$$

This method assumes that things that are closer to one another are more similar compared to those that are further apart, in this case a particular location is likely to have similar air pollution concentrations to those that are a larger distance away. This assumption is based on Toblers first law of geography [Waters, 2017]. The IDW equation (Equation 3.1) where x* represents the unknown value, and $\mathrm{w^n x^n}$ represents the distance and weights given at known values assumes that each known point has a local influence that decreases with distance and weights points closer to the unknown location greater than those that are further away [Li et al., 2014a]. Several studies have previously used inverse distance weighting-based methods to study the effect of pollution on health [Brauer et al., 2008],[Kim et al., 2009].Kriging is another interpolation method which consists of geostatistical methods and can be applied in several forms: simple kriging, universal kriging, and ordinary kriging. Ordinary kriging (OK) in the most applied form and works by estimating a value at a point in the study area for which a variogram is used, which uses neighbourhood data of the point to be estimated [Swan, 1996]. The main difference between IDW and OK is that kriging uses spatial autocorrellation among pollution values to determine the weights to assign to unknown points rather than simply assuming a function of inverse distance. The model assumes that the distance between sample points reflects a spatial correlation which is able to explain variation in pollutant levels. It works by fitting a mathematical equation to a user defined number of points to determine predicted values for unknown locations. The model is a multistep process including statistical analysis of the data and generation of a variogram model. The function of the variogram is to demonstrate the how dissimilar two observations separated by a given distance are [Son et al., 2010],[Greenblatt and Himes, 2019].

Suggest I want to find the pollution values at an unknown value (Figure 3.3), our prediction will be a linear combination of pollution values at neighbouring locations. In this model, I selected 3 nearest neighbours to be included in the estimations. The equation I will follow is as follows:

$$Y^{new} = w^1y^1 + w^2y^2 + w^3y^3 + E^{new} \tag{3.2}$$

Where $\mathrm{Y}^{\mathrm{new}}$ is the unknown location I want to estimate pollution levels for. Wn indicates the weight should be given to Y^{n} when Y^{n} is the pollution values at one of the 3 neighbouring locations. In order to complete equation 3.2, the weighting given to the known locations must first be determined. These weights are determined by equation 3.3. Given that I know the distance between the known point and the unknown points.

$$Y = X^iX^j = \frac{1}{2}(Y^iY^j)^2 \tag{3.3}$$

Nevertheless, in order to practice equation 3.3, I must first generate variogram models used to generate variogram functions between the known values and the

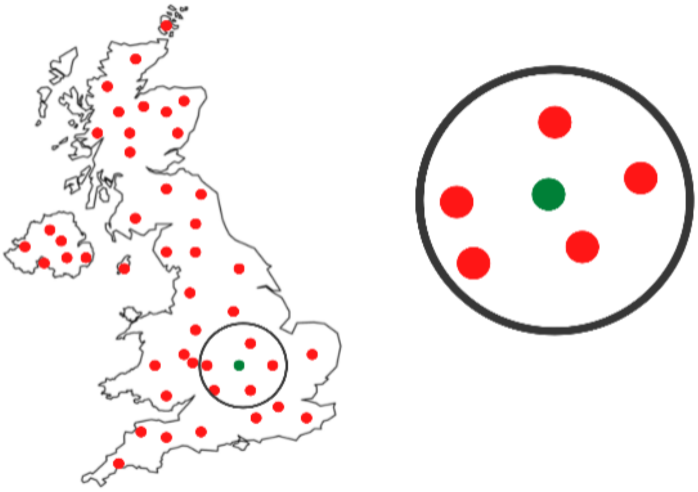


Figure 3.3: Study area of interest demonstrating known locations and unknown locations for pollution interpolation

unknown values, and further, the known values alone. The variogram models help to provide information on the spatial autocorrelation of the pollution values for all possible directions and distances. Essentially this is similar to regression analysis, where I fit a ‘line of best fit’ to the pollution data points to explain the nature of pollutants across the study region.

In equation 3.3, X represents the distance between the points, and Y represents the relationship between the pollution values. Simply, if I input 3 points into equation 3, the function is essentially one half of the difference in pollution between those 3 points. I therefore expect that the closer two points are in space, the smaller gamma will ultimately be.

$$w = A - 1b \tag{3.4}$$

A is comprised of variogram functions of the 3 known points: = (XiXj) and b is also comprised of variogram functions but between three known points and the unknown point which is given by gamma = ($X^{new}X^{i}$). Equation 3.3 generates the weighted values for each neighbouring point to be used in equation 3.1, expressed as W^{n}. To generate the weights of each point, I can take the inverse of equation 3.3 to produce equation 3.4, which ultimately provides the weights of each known point.

$$w = A^{-/}b \tag{3.5}$$

Implementation of inverse distance weighting and Ordinary Kriging (OK) models were completed in R. All plots and predictions were generated using the R packages; ggplot2,raster, gstat, Sp and rgdal [Pebesma, 2004],[Pebesma and Bivand, 2005a],[Wickham, 2011].

3.3.5 Distance to road measures

In the previous sections, I described methods to estimate air pollution from regional or secondary pollutants. However, these methods are not a sufficient measure for freshly emitted or primary pollutants such as traffic related air pollution (TRAP). Considering it is well known that TRAP shows very strong spatial gradients, meaning that concentrations of TRAP rapidly decrease with distance to busy roads, I decided to incorporate an additional measure of pollution exposure into this research described in Chapter 3 of this thesis. This is of importance due to the distinction made in the literature that secondary and primary air pollution sources may have different impacts on human health [Oudin et al., 2016],[Salam et al., 2008],[Künzli et al., 2009],[Henschel et al., 2012],[on the Health Effects of Traffic-Related Air Pollution, 2010]. Therefore, I calculated for each individual in our study, the distance from their residential address (measured by their postcode) to a busy road across the UK. I classified a busy road as either a motorway link, motorway, trunk, trunk link, primary, primary link or secondary road. I obtained UK road network data from the UK Open Roads ordnance survey and converted this data to spatial points using the R package sp [Pebesma and Bivand, 2005b]. I was then able to construct distance matrices between UK roads and an individuals postcode using the *distm* function from the geosphere package in R [Hijmans et al., 2016]. I then extracted the minimum distance to a busy road and this was used as a measure of TRAP in our EWAS described in Chapter 3.

3.4 Results

The results presented in this section are presented in two individual sections. First, I discuss the estimates of air pollution in my cohort sample, and secondly, I present the results from an epigenome wide association study for exposure to air pollution. This epigenome wide association study was performed using linear regression for the Illumina EPIC BeadChip for 1171 individuals in Understanding Society, aged 16 and above.

3.4.1 Statistical analysis for models to estimate air pollution exposure

I performed cross validation for each individual pollutant (NO2, PM10 , PM$^{2.5}$ and O^{3}) by separating the dataset into a test and training set. I sampled random locations across the UK which I had known pollution values for and reserved these locations to later be our test set for the models. The remaining data was used to train the models and in the case of the Kriging model, fit the variogram to. Once the models were trained in R, I then used the models to interpolate pollution levels on a random subset of 10% of the test data set which allowed us to test the accuracy of the models at estimating pollution levels. In order to assess the accuracy of the models, I then plotted actual vs predicted values in a scatter plot and calculated the Pearson correlation coefficients between the actual and observed values for each model, for each pollutant.

The observed value is the pollution values obtained from the Bayesian model estimates [Mukhopadhyay and Sahu, 2018] resulting in 500,000 known data points distributed across the study region. Specifically, these data points were an average mean obtained from annual measurements across a 5 year period prior to blood collection (2007-2011). Moreover, the predicted values are the values obtained from implementation of IDW and OK method in R. Once predictions had been made, predicted values vs observed values were plotted and Spearman correlation tests were performed in order to assess the accuracy of the predictions made by both methods. This assisted in validating the parameters that could have affected the interpolation accuracy of the pollution values.

3.4.2 Inverse distance weighting model

In order to compare the interpolation models, I computed correlation coefficients for observed values at known locations and predicted values at known locations. The coefficients between observed and predicted values were generally high ranging from 0.90 to 0.97. The lowest correlation was seen in Nitrogen dioxide values predicted by the IDW model, and the highest value observed was Particulate matter^{10} predictions by the Ordinary kriging model.

Several parameters were tested when applying the inverse distance weighting model to the pollution data for the study period of 2007-2011. The most significant parameter when applying the model was ‘nmax’ which involves defining the number of neighbour points to include in the model. The number of neighbours which produced the most accurate results was 3. In order to explore the model’s accuracy, I computed Pearson’s correlation coefficients for the actual and predicted values for each pollutant (Figure 3.4). The IDW model performed best for particulate matter 2.5 and particulate matter 10 with correlation coefficients of 0.95 and 0.94, respectively. The model was able to accurately predict values for these pollutants, even at the higher end of the range of yearly values. However, for nitrogen dioxide and ozone despite the correlations coefficients still being greater than 0.90, I observe large amounts of skewness at ‘extreme’ values where the pollution values fall far from the mean observed values. For example, for NO_2 , 87% of the actual

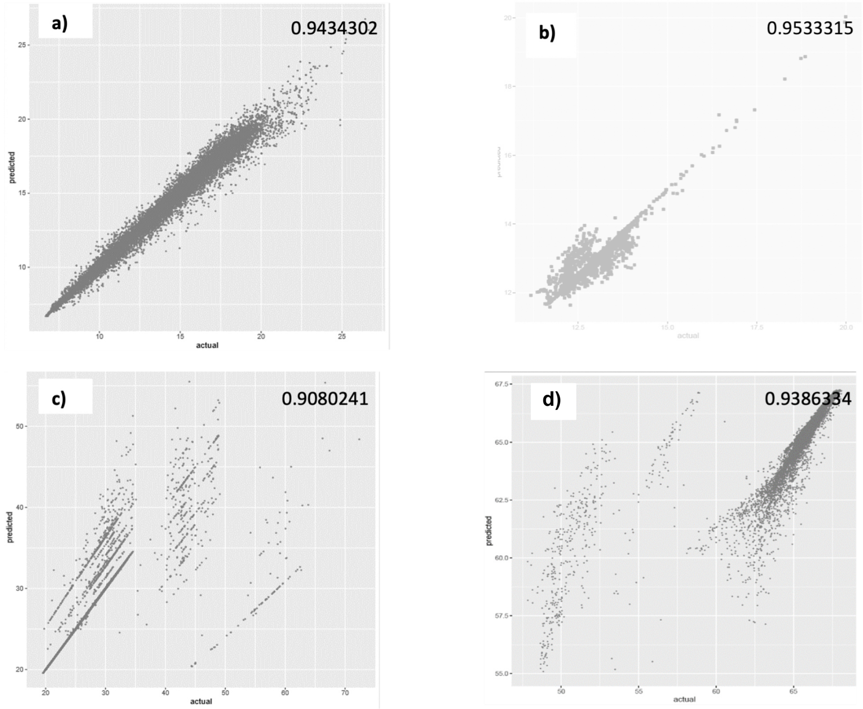


Figure 3.4: Scatterplots showing actual values of pollutant concentrations at known locations vs values predicted by the IDW model a) PM10 b) PM2.5 c) NO2 d) O3. Correlation coefficients are displayed at the top of each figure (confidence level=0.95, p-value of 2.2e-16).

pollution values fell in the range of 38-46 $\mathrm{ug}^{\text{m-3}}$. , but the model predicted 94% of the values to be within the range of 38-46 $\mathrm{ug}^{\text{m-3}}$.

3.4.3 Ordinary Kriging model

The ordinary kriging model provided the most accurate estimates of pollution exposure for the four pollutants NO$_2$, PM$_{10}$, PM$_{2.5}$ and O$_3$ for the study period of 2007-2011 (Figure 3.4.4). The two most accurate pollutants, similar to the IDW model were again, PM10 and PM2.5 with Pearson correlation coefficients of 0.97 and 0.96 respectively and even at the minimum and maximum values the model performed well. The ordinary Kriging model also performed better than the inverse distance weighting model for the two pollutants Ozone and Nitrogen dioxide with Pearson correlation coefficients of 0.9414516 and 0.9102311 respectively. Similar to the IDW model, the accuracy of the OK model decreases for the minimum and maximum values of these two pollutants, meaning it is less accurate at interpolating ‘extreme’ values which may be observed near extremely high air pollution sources. Therefore, I decided to use the OK models to estimate individual air pollution exposure, later used in my EWAS of air pollution exposure as discussed in Section 3.5.

3.4.4 Mapping in R

Mapping of the interpolated data from the Ordinary Kriging model was then carried out in order to visually validate the predicted data. Figure 3.4.4 shows a map of interpolated values determined by OK projected across the UK. A coloured scale ranging from white to green represents minimum and maximum values of nitrogen dioxide (a) and particulate matter 10 (b) in ug$^{\text{m-3}}$. Figure 3.4.4 shows maximum values of nitrogen dioxide and annual average exceedances of 40 ug$^{\text{m-3}}$. can be observed in the major cities such as London, Birmingham, Liverpool and Manchester. Cities such as Glasgow, Aberdeen, Edinburgh and the majority of southern England had values in the range of 10-20 ug$^{\text{m-3}}$, deemed safe by the WHO. Interestingly, the northern region of the UK has lower levels of nitrogen dioxide that fall below the annual average guideline of 40 ug$^{\text{m-3}}$. suggesting that nitrogen dioxide values are higher at busy roadside locations and areas located near large motorways more densely observed in the south. Figure 3.4.4B suggests that similar to NO_2, higher values of PM_{10} can be observed in the southern region of the UK and lower values are observed in the northern region of the UK. Figure 3.4.4 shows that in all major UK cities and in fact the majority of the UK are living in areas exceeding the air quality guidelines of 10 ug$^{\text{m-3}}$.

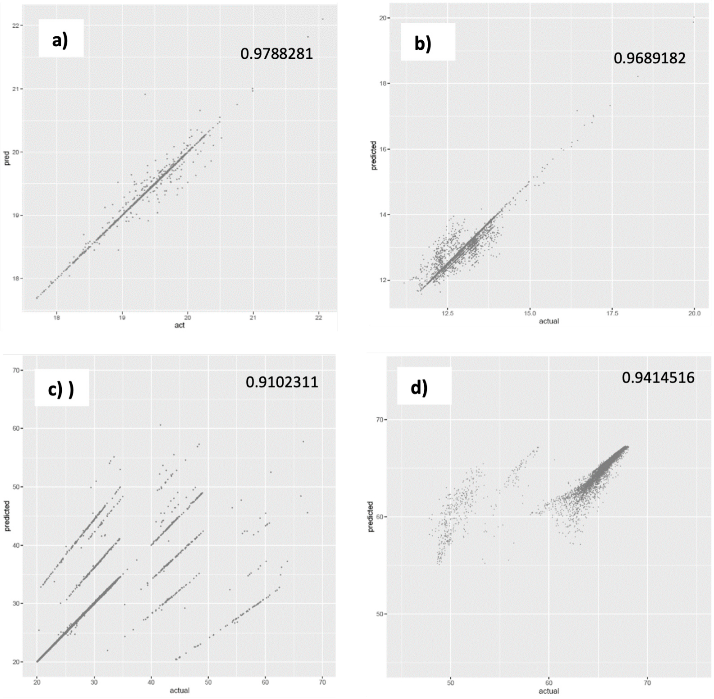


Figure 3.5: Scatterplots showing actual values of pollutant concentrations at known locations vs values predicted by the Ordinary Kriging model a) PM10 b) PM2.5 c) NO2 d) O3. Correlation coefficients are displayed at the top of each figure (confidence level=0.95, p-value of 2.2e-16).

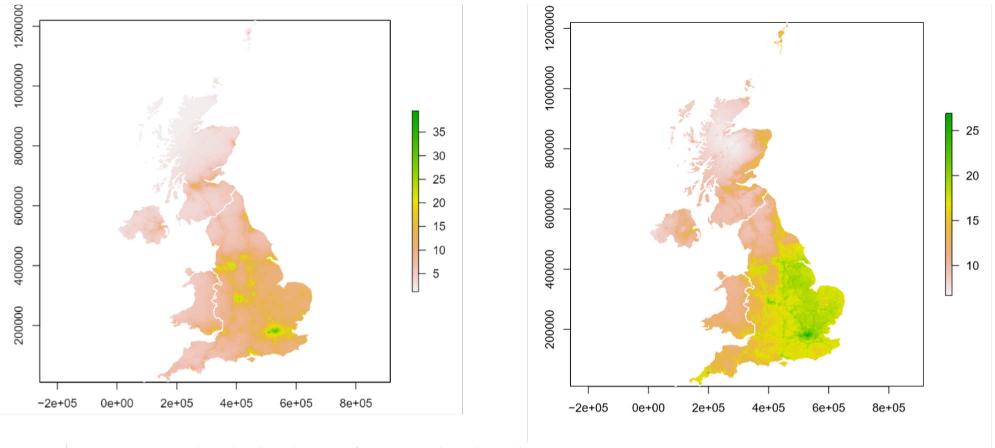


Figure 4: Interpolated values by OK of nitrogen dioxide and PM$_{210}$ (ug m-3) across the UK.

Figure 3.6: Interpolated values by Ordinary Kriging of nitrogen dioxide and PM$_{2.5}$

3.5 Epigenome wide association study of air pollution

Analysis of DNA methylation signatures related to air pollution were explored by performing linear regression for the IlluminaEPIC BeadChip array. Due to data access restrictions, I performed this analysis in one individual cohort, comprised of 1161 participants. Details of this cohort are described in 2. In order to explore this relationship, I performed several epigenome wide association studies using different measures of air pollution. Briefly, I ran a separate EWAS for four pollutants (Nitrogen dioxide, ozone, particulate matter 10 and particulate matter 2.5) and an additional EWAS using distance to road as a measure of TRAP. After data processing and cleaning, a total of 747,302 CpGs were analysed (see Chapter 2). Since whole blood is a bulk tissue, each EWAS adjusted for cell type proportions, and a number of other potential confounding factors as described in equation 3.6.

$$DNAm \sim Pollution\ measure + Batch\ effects + CD8T + CD4T + NK + BCell + Mono + Gran + Age + Sex + Smoking\ score \tag{3.6}$$

3.5.1 There are no significant detectable DNA methylation changes in response to exposure to background levels of ambient air pollution

To address my first goal of investigating how background levels of air pollution may modulate the human epigenome, I performed EWAS for the four main pollutants previously described. This involved using a measure of average air pollution exposure over a five year period prior to blood collection and DNA methylation measurement (2007-2011). In the first instance, after adjusting for multiple testing using the Benjamini Hochberg FDR method (FDR < 0.05), I found no significant detectable DNA methylation changes related to air pollution exposure for PM_{10},

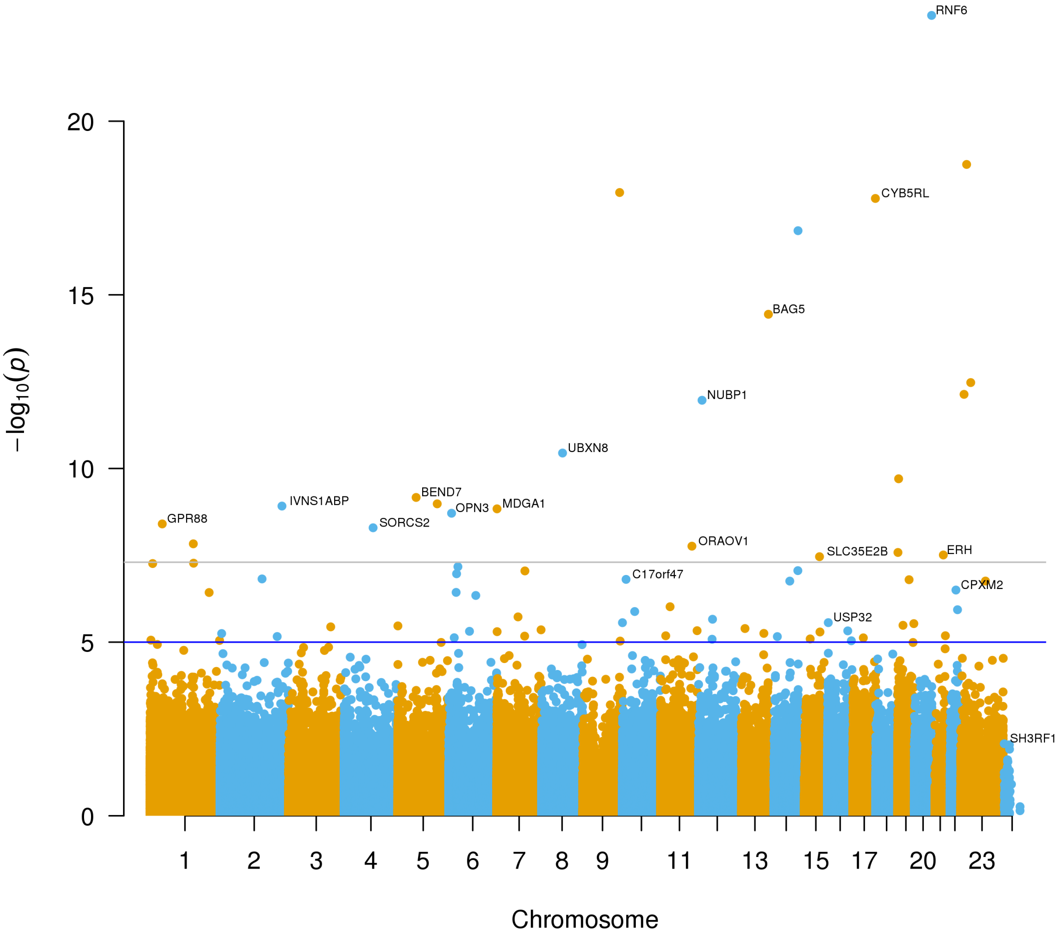


Figure 3.7: Manhattan plot for EWAS analysis of exposure to $PM_{2.5}$. The x axis represents the genomic position, and the y axis represents the -$\log_{10}$ pvalue significance. Each individual point represents an individual CpG site with a number of these being labelled with their annotated gene name. Darker coloured points represent CpG sites deemed to be significantly associated with $PM_{2.5}$ exposure by EWAS analysis.

NO_2 or O_3.

However, I did initially identify 43 CpGs significantly associated with $PM_{2.5}$ following FDR correction (see Figure 3.7), which I found to be positioned in or near the following genes: RNF6,CYB5RL,FAM18A, UBXN8, BEND7, AC062028.1, IVNS1ABP, MDGA1,OPN3, ORAOV1, RNF220, SLC39A9,ERH,C17orf47, GDAP1, CPXM2, WISP1, Y RNA, TLK2, NT5C2, R3HDM1, ZRANB3 and GOLGB1. Several of these genes have been previously reported to be related to air pollution exposure and lung function [Sayols-Baixeras et al., 2019, Jackson et al., 2018].

Prior to downstream analysis, I carried out some quality checks of our EWAS results. Firstly, I plotted methylation values (Beta values) against $PM_{2.5}$ at the most significant CpG site, cg21553490 in gene RNF6 3.8. Upon this analysis, I identified two participants in our sample who had extremely high exposure to $PM_{2.5}$ compared to the rest of the sample 3.9. To ensure these were not just outliers of our PM2.5 prediction models, I checked the geographical location of these two participants and were able to confirm that they lived in close proximity, meaning our $PM_{2.5}$ exposure values were likely accurate. I hypothesised that these two samples could potentially be driving our EWAS results, due to the fact that I was seeing extremely small P values in genes which did not seem to be highly biologically relevant which although not completely unreliable, deemed further investigation.

Therefore, I decided to remove these samples from our analysis and re-run our EWAS for $PM_{2.5}$ exposure. Following this, I observed that all of our signals were no longer significant and I did not identify any other significant $PM_{2.5}$ exposure associated DMPs. These results suggest that it is necessary in these types of EWAS, to check distribution of phenotype measurements within the sample. Although it is entirely possible that these signals may just have been inflated and not insignificant, I note that the beta values at the top hit at CpG site cg21553490 in gene RNF6 do not show a consistent methylation pattern, one sample has very low methylation whereas the other sample has intermediate methylation despite the similar high

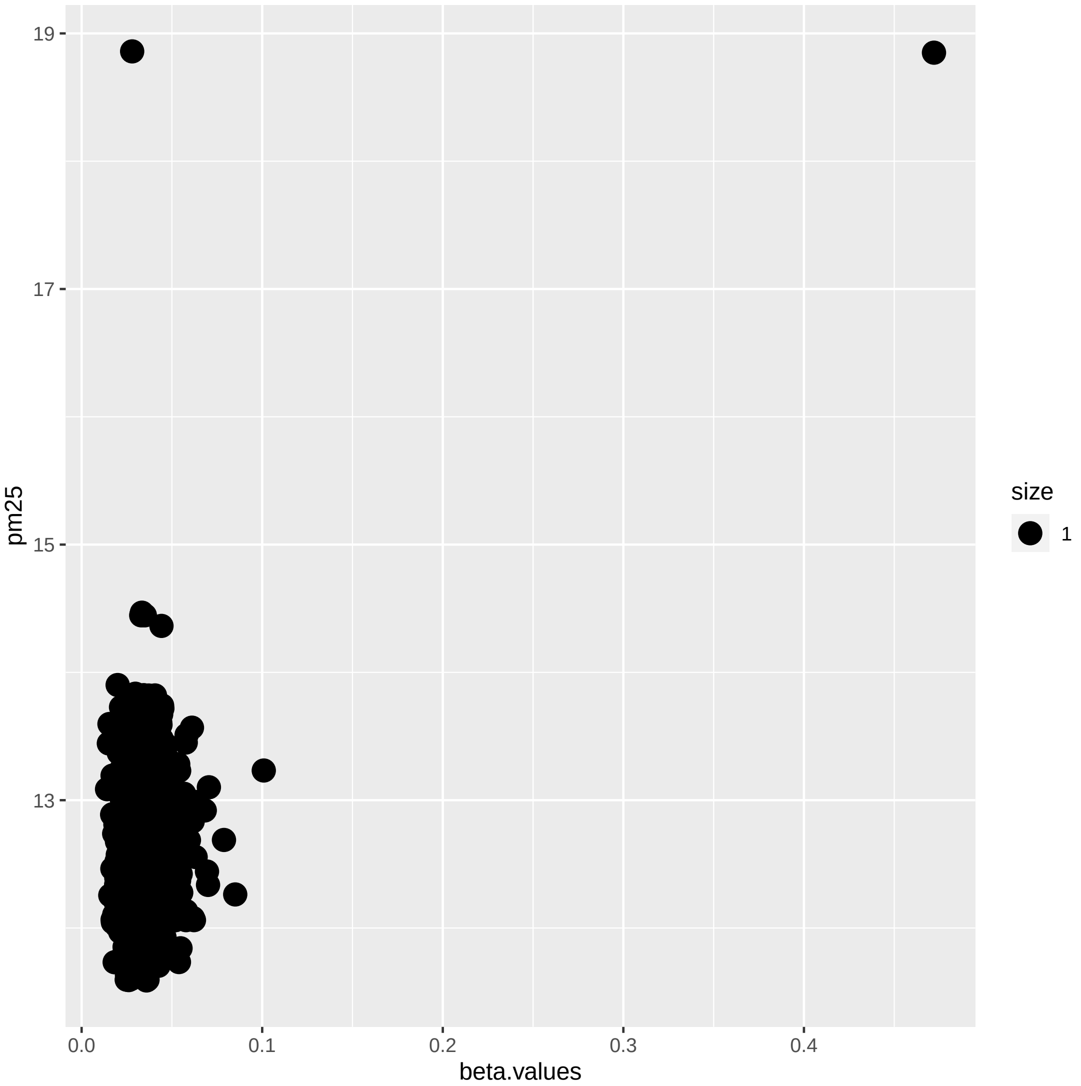


Figure 3.8: Scatterplot showing relationship between methylation value and $PM_{2.5}$ exposure at CpG site cg21553490 in gene RNF6.

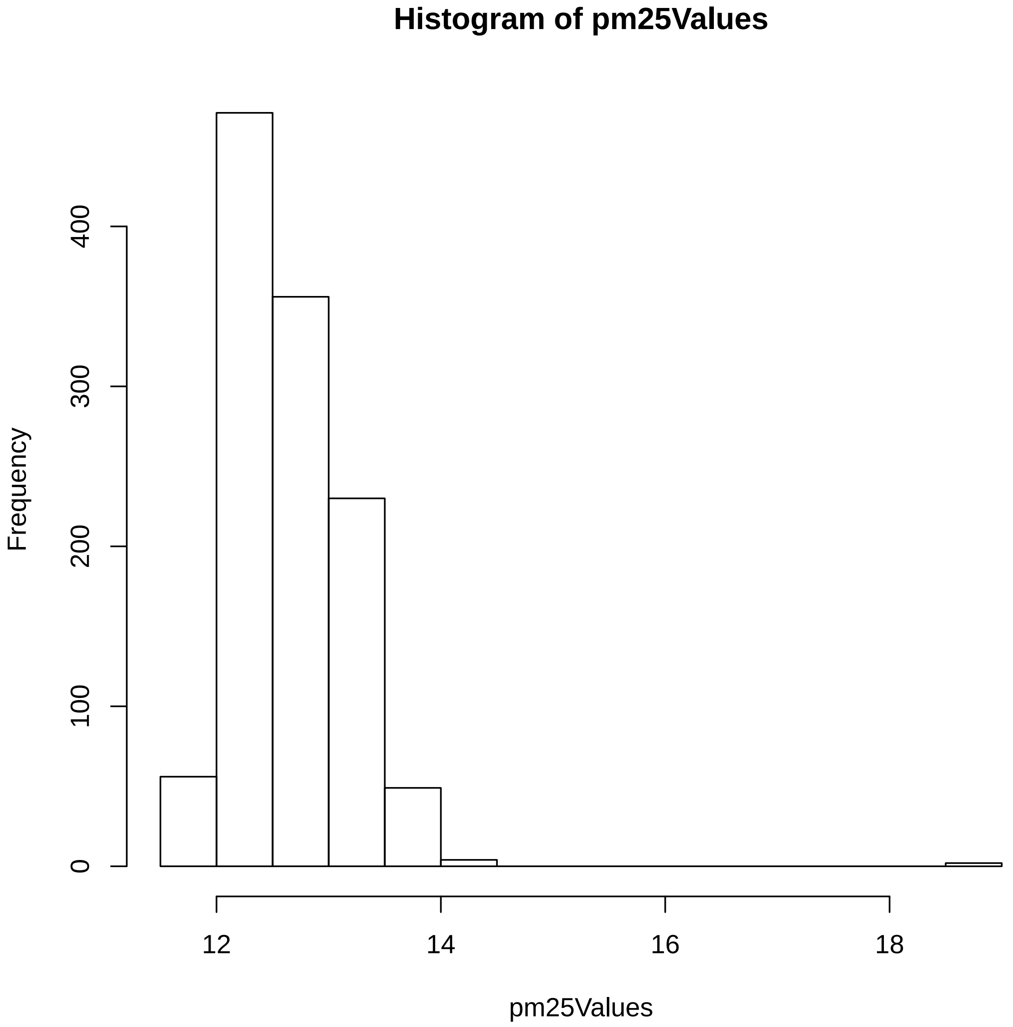


Figure 3.9: Histogram showing range of $PM_{2.5}$ exposure values across our study sample for the study period of 2007-2011.

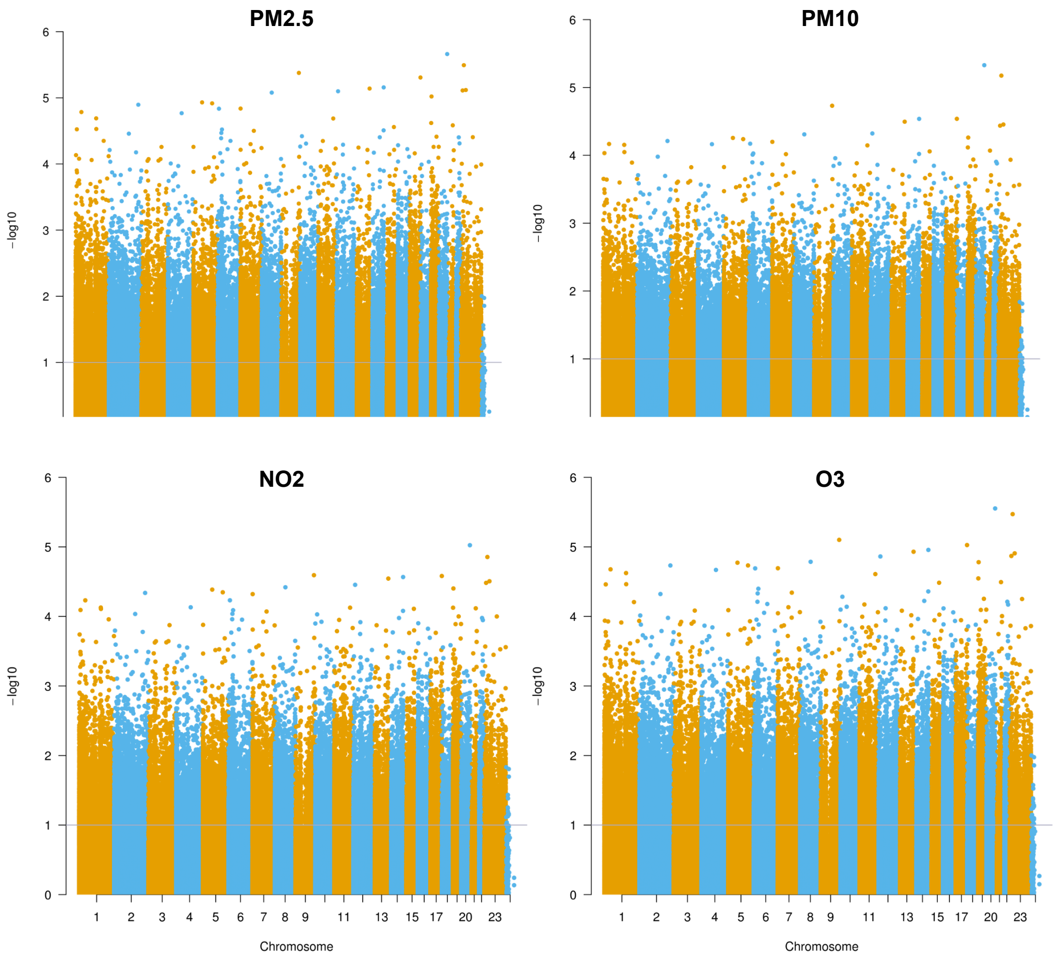


Figure 3.10: Manhattan plot for EWAS analysis of exposure to the four main air pollutants. The x axis represents the genomic position, and the y axis represents the -log10 p-value significance. Each individual point represents an individual CpG site.

$PM_{2.5}$ exposure values 3.8.Nevertheless, this suggests that this loci may warrant further investigation in response to air pollution exposure.

3.5.2 A distance to road measure gives insight into how exposure to traffic related air pollution may modify DNA methylation signatures across the genome

In line with my second aim of investigating how traffic related air pollution may influence DNA methylation patterns in human whole blood, I performed EWAS

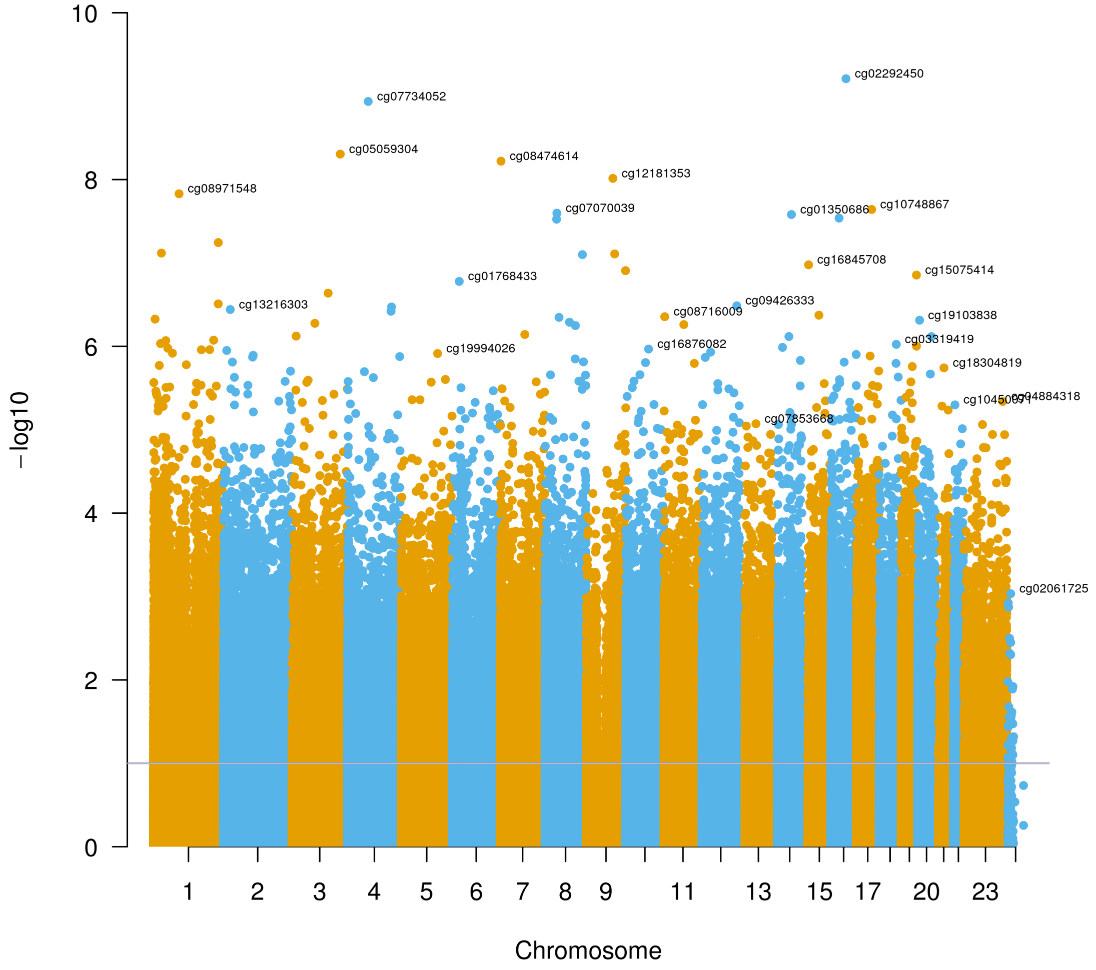


Figure 3.11: Manhattan plot for EWAS analysis of exposure to TRAP. The x axis represents the genomic position, and the y axis represents the -$\log_{10}$ p value significance. Each individual point represents an individual CpG site with a number of these being labelled with their respective CG identifier. Darker coloured points represent CpG sites deemed to be significantly associated with TRAP exposure by EWAS analysis.

using a distance to road measure. This consisted of a measure of how close a participant lives to a busy road as described in Section 2. Following adjustment for multiple testing using the Benjamini–Hochberg FDR method (FDR $p<0.05$), I identified 531 significant CpG sites associated with traffic related air pollution (Figure 3.11). Links to the full list of these sites is available in the appendix (see A) Of these, 266 (50,09%) of these CpGs were hypomethylated in response to TRAP exposure and 265 were hypermethylated in response to TRAP exposure. Gene ontology analyses showed no enriched terms for these 531 CpGs following FDR adjustment. Similarly, KEGG analyses also revealed no enriched terms either. From hereon, I refer to these 531 significant CpGs as road associated DMPs (raDMPs).

3.5.3 Characterisation and location of road associated DMPs

The raDMPs were found in 374 unique genes with 15 of these genes harbouring several raDMPs (Figure 3.11). The number of raDMPs harboured by individual genes ranged from 1 to 4. Two genes, C10orf4 and ZNF678 contained 4 raDMPs each. The 4 CpGs located in C10orf4 were all in the transcription start site or 5' UTR and were all hypermethylated in response to TRAP exposure. On the other hand, all 4 CpGs located in ZNF678 were all located in the 5' UTR and all but 1 were hypermethylated in response to TRAP exposure. Additionally, neither of these genes have previously been linked to air pollution. However, I did identify CpG sites overlapping genes which have previously been linked to exposure to air pollution. For example, the most significant road associated CpG, cg02292450 associated with the gene CSMD3 which is a gene which has previously characterised as a somatic mutation in air pollution related lung cancer, been been implicated in asthma–COPD overlap syndrome and is thought to lead to increased proliferation or airway epithelial cells [McGeachie et al., 2016, Yu et al., 2015, Liu et al., 2012]. Although, the exact function of this gene does remain unknown, previous reports suggest that it may be involved in proliferation of epithelial cells and maintaining smooth muscle tone, of which both can lead to impaired lung function [Steinke, 2016]. Moreover, the third most significant raDMP was associated with the gene SYCP2L, a gene which has previously been linked to a causal SNP related to car or vehicle pollution [D'Antona et al., 2022] and also exposure to herbicides [Rytel et al., 2021]. Furthermore, another top raDMP was annotated to a gene called GRID2IP, which contains several CpGs which have previously been identified to be differentially methylated to gases and fumes exposure. [Van Der Plaat et al., 2019].

Next, I characterised the genomic locations of the raDMPs in order to learn more about their functional role. First, I found that the raDMPs are slightly depleted at CpG shelves and slightly enriched at CpG shores and open sea regions of the genome compared to the EPIC background (see Figure 3.12). As I found an enrichment in open sea regions, I then annotated the raDMPs to functional regions in the genome and compared this with the EPIC background also. Road associated DMPs were found to be very slightly enriched at enhancer regions and 5'UTR regions of the genome compared to the EPIC background. This suggests that these raDMPs could potentially play a role in gene regulation. Interestingly, I did not identify significant depletion of the raDMPs in any of these regions compared to the EPIC background (Figure 3.13).

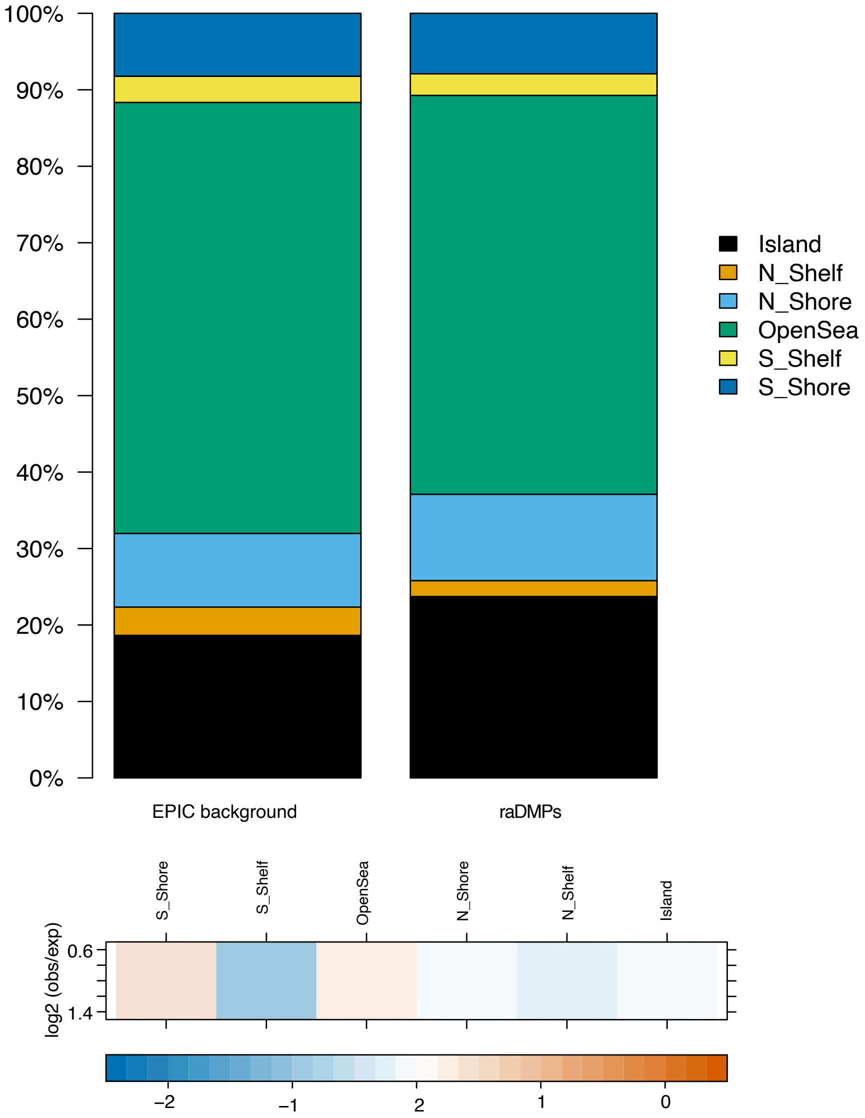


Figure 3.12: Annotation of raDMPs relative to CpG islands. Top panel shows the annotation of all raDMPs compared to the EPIC background. Bottom panel shows the log2 (obs/exp) annotations based on the EPIC background of the different annotations.

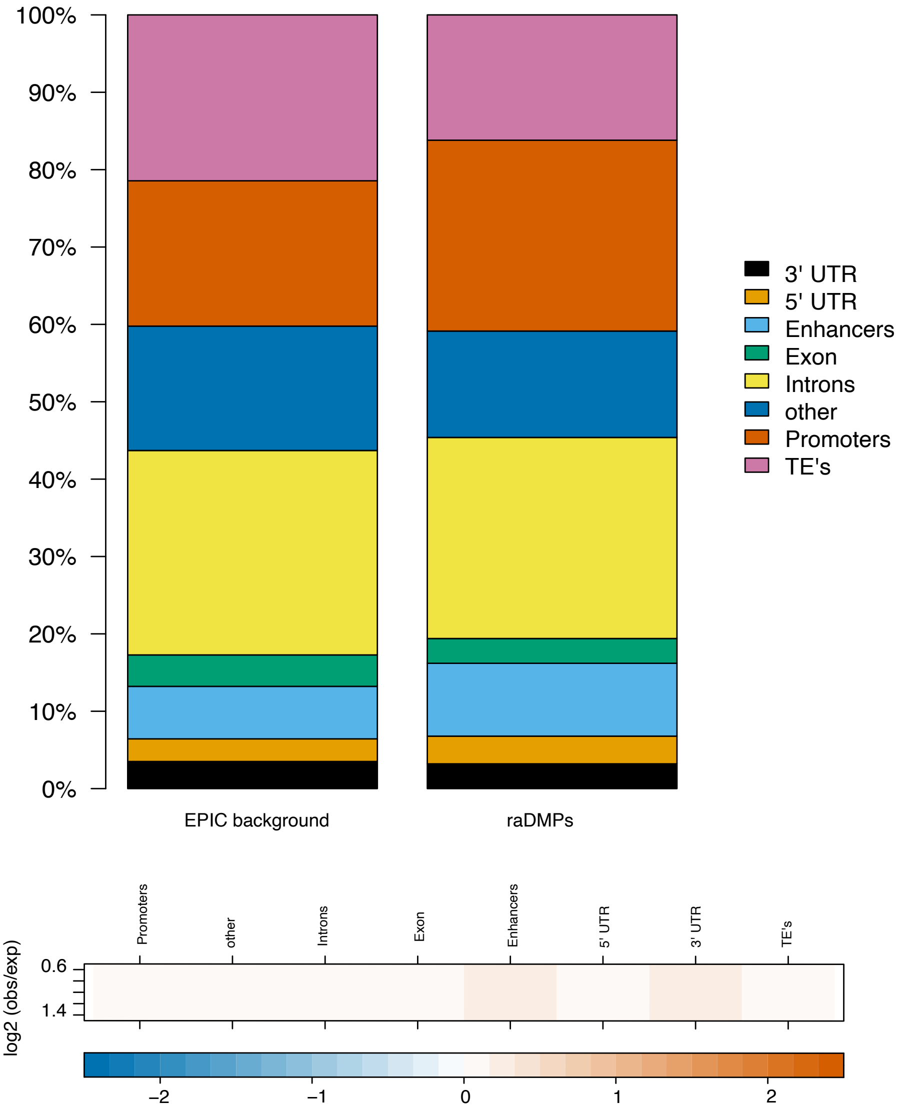


Figure 3.13: Functional annotations of raDMPs. Top panel shows the overlap of all raDMPs with functional regions compared to the EPIC background. Bottom panel shows the log2 (obs/exp) annotations based on the EPIC background of the different annotations.

Table 3.3: Top 20 significantly associated raDMPs identified by EWAS analysis.

	Chr	Position	Relation_to_Island	UCSC_RefGene_Name	UCSC_RefGene_Group	adj.P.Val
cg02292450	chr8	114389388	OpenSea	CSMD3;CSMD3;CSMD3	Body;Body;TSS200	0.00049606362702384
cg07734052	chr4	124820603	OpenSea	LOC285419	Body	0.00049606362702384
cg05059304	chr6	10887047	N_Shore	SYCP2L	TSS200	0.00128957181944048
cg08474614	chr16	66912266	N_Shore			0.00128957181944048
cg12181353	chr1	227751318	N_Shore	ZNF678;ZNF678;ZNF678	1stExon;Body;5'UTR	0.00165820907312231
cg08971548	chr1	161676199	OpenSea	FCRLA;FCRLA;FCRLA;FCRLA;FCRLA;FCRLA;FCRLA	TSS1500;TSS1500;TSS1500;TSS1500;TSS1500;TSS1500;TSS1500	0.00211649479819944
cg10748867	chr1	19577036	N_Shore	MRTO4;KIAA0090	TSS1500;Body	0.00232792561286792
cg07070039	chr7	6576529	S_Shore	GRID2IP	Body	0.00232792561286792
cg01350686	chr1	24439075	OpenSea	MYOM3	TSS1500	0.00232792561286792
cg14314533	chr10	3577056	OpenSea			0.00232792561286792
cg17333791	chr11	64811900	N_Shelf	SAC3D1	Body	0.00232792561286792
ch.6.573831F	chr6	22835108	OpenSea			0.00406993583019027
cg02851625	chr8	79064223	OpenSea			0.00454148652414284
cg19160858	chr16	57452988	OpenSea			0.00454148652414284
cg06601131	chr8	26492999	OpenSea	DPYSL2;DPYSL2;DPYSL2	Body;Body;Body	0.00454148652414284
cg16845708	chr1	22379018	Island	CDC42;CDC42;CDC42	TSS200;TSS200;TSS200	0.00564294241027037
cg13071326	chr7	75368332	Island	HIP1	TSS200	0.00623567841066201
cg15075414	chr17	31240005	OpenSea			0.00665467670388827
cg01768433	chr5	90185030	OpenSea	GPR98;GPR98	Body;Body	0.00750584805361633
cg02365742	chr5	174151204	N_Shore	MSX2	TSS1500	0.00986775699563627

As I found an enrichment of these raDMPs in open sea and enhancer regions in the genome, I hypothesised that perhaps these raDMPs could potentially be involved in regulating or contacting distal genes. I therefore further annotated our raDMPs to additional genes using chromatin loop data as described in section 2. Through this analysis, I was able to annotate our raDMPs to an additional 21 genes (see Table 3.4). Interestingly, some of the genes identified in this analysis have previously been reported to contain differentially methylated CpGs associated with NO_2 and $PM_{2.5}$ such as CCDC88C-DT and DCBLD2 [Sayols-Baixeras et al., 2019]. However, most of these are genes which have not previously been linked to air pollution, revealing novel loci related to TRAP exposure.

Table 3.4: Additional genes annotated to raDMPs via chromatin loops

ENTREZID	SYMBOL	ENSEMBL
101929093	LINC01350	ENSG00000228309
10402	ST3GAL6	ENSG00000064225
105370625	CCDC88C-DT	NA
105371814	LOC105371814	ENSG00000248278
131566	DCBLD2	ENSG00000057019
1666	DECR1	ENSG00000104325
2101	ESRRA	ENSG00000173153
25824	PRDX5	ENSG00000126432
285540	SEPSECS-AS1	NA
41	ASIC1	ENSG00000110881
4683	NBN	ENSG00000104320
51091	SEPSECS	ENSG00000109618
51504	TRMT112	ENSG00000173113
55300	PI4K2B	ENSG00000038210
56990	CDC42SE2	ENSG00000158985

643836	ZFP62	ENSG00000196670
676	BRDT	ENSG00000137948
7587	ZNF37A	ENSG00000075407
79726	WDR59	ENSG00000103091
80012	PHC3	ENSG00000173889
84937	ZNRF1	ENSG00000186187

3.5.4 TFs enriched at raDMPs

Next, I performed transcription factor (TF) binding site and gene ontology analyses for enriched TF motifs to try and identify if the raDMPs may be part of any important gene regulatory networks. First, I assessed whether the raDMPs were enriched in motifs for TFs (50 bp window). I identified a total of 65 enriched transcription factor motifs at an adjusted p-value lower than 0.01. (FDR) (see Table 3.5). The top 5 enriched motifs were TMSL3, NNT, CANX, TIMM8A and DNMT3A. Interestingly, all 5 of these motifs have previously been linked to air pollution exposure or lung function [Ward et al., 2020, Nwanaji-Enwerem et al., 2016a, Kobayashi et al., 2015, Yan et al., 2022].

To analyse whether the TF motifs were enriched for annotation to biological processes or pathways, I performed pathway analyses using the GO and KEGG databases. I identified 0 enriched KEGG or GO terms for the TF motifs enriched at raDMPs, likely due to the small number of enriched TFs.

Table 3.5: Enriched transcription factor motifs at raDMPs.

Rank	Target	ID	Raw score	P value
1	TMSB4XP8	TMSL3	0.827288096784864	1.21408623877925e-09
2	NNT	NNT	1.19025427326105	2.77645708375133e-09
3	CANX	CANX	1.10373551934907	4.64968091531798e-09

Table 3.5: Enriched transcription factor motifs at raDMPs.

Rank	Target	ID	Raw score	P value
4	TIMM8A	TIMM8A	1.23444645422692	4.96893080107032e-09
5	DNMT3A	DNMT3A	0.90129984407243	5.1118544220231e-09
6	UQCRB	UQCRB	0.628081088170516	1.8591517274927e-08
7	GLTPD1	MGC10334	1.39369620966956	4.89546033445032e-07
8	GOT1	GOT1	0.710031438064806	5.307896128857e-07
9	CELF6	BRUNOL6	1.05729167291615	7.6325603707753e-07
10	LINC00471	C2orf52	1.47158164725332	6.55989983994937e-06
11	PRDX5	PRDX5	0.753056343399507	1.37954383903783e-05
12	RBBP5	RBBP5	0.894185596946131	1.54112932858781e-05
13	P4HB	P4HB	1.35306404331275	1.5791808491221e-05
14	ZNF304	ZNF304	1.49183698809017	4.30427112197044e-05
15	NANOS1	NANOS1	1.52911227609363	5.07465193511333e-05
16	GTF2B	GTF2B	0.807539056443985	7.14923524773372e-05
17	TGIF2LX	TGIF2LX	1.3584976422527	7.4524640855131e-05
18	ESX1	ESX1	1.24864420270669	8.51169599222319e-05
19	SSX3	SSX3	1.58960117057919	0.000109518470485686
20	TAF9	TAF9	1.02805872401067	0.000114544271685169
21	THRA	THRA	1.19969956577158	0.000163849040887651
22	DUS3L	DUS3L	1.26089529595883	0.000216629340166059
23	TOB2	TOB2	2.03768716227727	0.000243422550342213
24	ODC1	ODC1	1.54509401301601	0.000269466244115514
25	DIABLO	DIABLO	1.64775740187529	0.000317442187719284
26	FEZF2	FEZF2	1.09103387052372	0.000335411487949143
27	NRL	NRL	0.789104546163279	0.0003641714268999
28	SOX13	SOX13	1.6270709116781	0.000397433228109449

Table 3.5: Enriched transcription factor motifs at raDMPs.

Rank	Target	ID	Raw score	P value
29	HSPA5	HSPA5	1.05249742055632	0.000400645761654624
30	NHLH1	M3376_1.02	1.874154065013	0.00044338032361957
31	RARA	RARA	0.674686172033714	0.000482444615334126
32	CEBPA	M3040_1.02	0.783521558133718	0.000520737416306532
33	USF1	M4514_1.02	1.14868483848455	0.000668757518554214
34	SEMA4A	SEMA4A	1.2745286133526	0.000768236225296607
35	BAX	BAX	0.836956572995134	0.000922947287432346
36	RBM17	RBM17	1.47525439420296	0.000960694407058086
37	PKNOX2	PKNOX2	1.28732670322194	0.00100442306221528
38	NFIC	NFIC	1.51916388336712	0.0013452939752786
39	ARFGAP1	ARFGAP1	1.08563669518683	0.00148826421223246
40	FAM127B	FAM127B	1.24625881364519	0.00172624237778789
41	PPP1R10	PPP1R10	1.1602170177035	0.00194280830183351
42	RAB7A	RAB7A	1.74582882803546	0.00207587955008505
43	C19orf40	C19orf40	1.12532742580417	0.00268168010028647
44	ZNF71	ZNF71	1.6032595636713	0.00274929396049037
45	DAZAP1	DAZAP1	1.58525795987432	0.003000956719951
46	ING3	ING3	0.808568050547915	0.00345708145584539
47	ID2	ID2	0.731856759058946	0.00349641644825198
48	CSNK2B	CSNK2B	1.29204593743006	0.00384662754667635
49	ZNF671	ZNF671	1.51814298384407	0.00397949598812966
50	RFX3	RFX3	1.74333508204794	0.00471609867724739
51	SNRPB2	SNRPB2	1.57618659806703	0.00513159036588882
52	PCK2	PCK2	0.720423117330663	0.00514800158428574
53	VPS4B	VPS4B	1.21510971117641	0.00515572963660667

Table 3.5: Enriched transcription factor motifs at raDMPs.

Rank	Target	ID	Raw score	P value
54	C9orf156	C9orf156	2.18318442483659	0.00525491121863917
55	H2AFY	H2AFY	1.24236227662047	0.00590983523255075
56	UBE2V1	UBE2V1	1.0952813923256	0.00629833012946529
57	ATF3	M4597_1.02	1.02714992696599	0.0072377886332191
58	TGIF1	TGIF1	1.07125578521327	0.0076086899473959
59	WISP2	WISP2	1.3395342023463	0.00772294578117267
60	MECP2	MECP2	1.06877876337402	0.00802649270508451
61	RAB2A	RAB2A	0.580058316966397	0.00810137020449613
62	ABCF2	ABCF2	1.22027636909557	0.00815952384451781
63	HTATIP2	HTATIP2	0.74604994580638	0.00846570525645831
64	ANXA1	ANXA1	1.29318087804033	0.00878088132549926
65	ESRP1	RBM35A	1.42805238965326	0.00941938966505146

3.5.5 Identifying road associated differentially methylated regions

As previously mentioned, some genes contained several raDMPs, I therefore attempted to identify larger differentially methylated regions associated with TRAP exposure. Thus, I searched for road associated differentially methylated regions (raDMRs). Following adjustment for multiple testing (FDR) and accounting for for cell type proportions, smoking and age, I identified 3 raDMRs. The number of CpGs within the DMRs ranged from 5 to 13 and were located on chromosome 1, 10 and 11. The first raDMR was located on chromosome 1 and had a width of 707 base pairs. This DMR contained 8 CpGs, 2 of which were also identified within our DMP analysis. This region showed hypermethylation in response to TRAP exposure and had an adjusted p value (FDR) of 1.87779e-12. Interestingly,

Table 3.6: 3 road associated differentially methylated regions identified

	seqnames	start	end	width	strand	no.cpgs	overlapping.genes
1	chr1	24438607	24439313	707	*	8	MYOM3
2	chr11	41481497	41481891	395	*	5	NA
3	chr10	95462161	95462662	502	*	13	FRA10AC1

this region overlapped a gene called MYOM3. This gene has not previously been linked to air pollution exposure, however it has previously been reported to be hypermethylated in relation to cardiovascular disease [Shi et al., 2022]. Secondly, a second raDMR located on chromosome 11, 395 base pairs long contained 5 CpGs. This contained one of the raDMPs identified in our earlier analysis, identified as cg08072402 found to be hypermethylated in open sea region. Lastly, the third raDMR was found on chromosome 10, spanning 502 base pairs. This region contained 13 CpGs, 2 of which were also identified as raDMPs, cg16661579 and cg10830758. This region did overlap the gene FRA10AC1, and the CpGs were found to be hypermethylated in a CpG island in the promoter region of this gene. FRA10AC1 is a known fragile site, although little else is known about this loci, its hypermethylation has previously been linked to neurodevelopmental disorders (NDs) and congenital anomalies [Barbosa et al., 2018].

3.6 Discussion

Here, I discussed research aiming to estimate pollution levels across the UK using Inverse Distance Weighting and Ordinary Kriging in R, two of the most widely used methods for spatial interpolation. Yearly exposure estimates simulated from our two models for 100,000 randomly sampled locations in the UK are compared in order to guide a choice of model for use in the epigenome wide association studies. The accuracy of the models was predicted by Pearson correlation coefficients and mapping of the data across the study region. High resolution models such as these, are necessary for use in EWAS as previous research suggests that levels of ultrafine particles can differ greatly even across distances as small as 100-300m [Zhou and Levy, 2007]. Further, several studies have shown an increase in susceptibility to adverse health effects for individuals located in close proximity to roads in a 100m range [Ryan et al., 2005],[Sekine et al., 2004]. Findings such as these strengthen the need for models which are able to distinguish individuals located next to busy motorways for example, from those who are living within a mile radius, as research suggests health effects may differ dramatically, even with a small difference in distance to an air pollution source, such as a busy motorway. Both of the models performed extremely well for pollutant values clustered around the mean, however, especially in the case of Nitrogen Dioxide and Ozone, the models performed less accurately for 'extreme' values. Although these values could be considered outliers, I know that it is probably more likely that these values are residential areas which are located next to large, busy motorways or other pollution sources such as coal or powerplants [WHO, 2016]. Therefore, the model which best predicts these extreme values, as well as the values clustered around the mean will be the best model for use in epigenome wide association studies. This is because, the individuals who live near these sources of high pollution, will potentially act as interesting comparisons within our epigenome wide association analysis allowing us to explore the difference between long-term high exposure, and extremely high exposure to ambient air pollution. This will ultimately allow us to assess health effects in comparison with the yearly values set by the world health organization [WHO, 2016].

Similar research has also investigated the use of interpolation models in estimating air pollution for use in epidemiological studies. Although interpolation models have previously been used in epigenome wide association studies, to our knowledge this is the first work which compares two interpolation methods for use in epigenome wide association studies. Previous work focusing on epidemiological studies, have found that OK models generally performed much better than other interpolation models [Rivera-González et al., 2015b], [Kim et al., 2009]. However, previous studies have found that error when interpolating PM2.5 was considerably higher than error for other pollutants such as NO2 and O3 although I found that our predictions for PM2.5 were the second most accurate. This could, however, be due to the fact that our data set consisted of a large amount of data compared to other studies. Our data set consisted of 500,000 observations, whereas previous studies have only had 9 observations to interpolate values from [Rivera-González et al., 2015b].

The strengths of this research come from the fact that I assessed methods which address both the temporal, and spatial nature of the air pollution data within the UK. Secondly, as our models provided an input of latitudinal and longitudinal coordinates, I are able to ensure I maintain privacy and confidentiality of participants involved in our research by not including their postcodes. This ensures that postcode data is not able to be linked to the epigenetic or genetic data which will later be used in genome wide association and/or epigenome wide association studies. Further, when values from monitoring stations are used within this type of research, there are a number of things which must be considered. Firstly, the results are obviously highly dependent on location and frequency of monitors available, distance of individuals from the monitors and also how representative the monitoring stations are of the study region. Our results are based on data from [Mukhopadhyay and Sahu, 2018] which utilised several data sources, including the air quality modelling unified model (AQUM) [Savage et al., 2013] which is a chemistry transport model utilising emission inventory and weather data. It also includes data from the UK Automatic Urban and Rural Network data which and lastly, map data from the moderate resolution imagine spectroradiometer (MODIS) which provide land use information at an accuracy of 93%. This data was obtained using the ordinance survey maps [Schneider and Garrett, 2009]. The vast use of data within this spatiotemporal model means our data is highly representative of the study region and increases our number of ‘known’ sites which were input into our model. The AURN in the UK itself, also covers a large unbiased region of the study area. Normally, monitoring sites are placed in areas where pollution is known to be high. Advantageously the AURN places site at roadside, urban and rural locations, removing that bias from this work. Lastly, the application of several interpolation methods builds a foundation for future work based on utilising this type of data in not only epigenome wide association studies, but also studies of a similar nature including genome wide association studies and epidemiological studies.

This research does, however, have some limitations. The first being that adoption of kriging models relies on a few assumptions, one being that the data that the variogram is built on is stationary [Olea, 1999]. This means that if I was to take a small chunk of the UK, and examine the attributes of that area, the mean of the pollution scores would be the very similar. Although most of our data is stationary, I acknowledge that there are certain areas of our study region in which I would observe higher values of pollution for. These include congested cities, industrialised areas, or areas within a certain radius of an airport for example. This could provide explanation as to why our model performs less accurately for 'extreme' values which I might consider outliers. Although, it is worth noting that the ordinary kriging model still performs better than the inverse distance weighting model despite this assumption.

The second limitation is that the model also assumes a constant variogram. That is, that if I again were to take a small section of the study region, the change in pollution based upon change in distance of two locations should be about the same as another section of the study region. Again, although this is generally true, I also acknowledge there are 'outliers' or 'extreme values' located near pollution hotspots which may defy these assumptions [Olea, 1999].

Lastly and most importantly, our research is only concerned with measuring outdoor air pollution however this does not accurately represent true exposure as individuals may be exposed to higher or lower levels of pollution in other settings, such as their work commute, or workplace, or inside their home as discussed in [Shaddick et al., 2008],[Shaddick et al., 2008]. However, modelling true personal exposure by incorporating workplace exposure estimates is a much more difficult, and computationally intensive task, especially when considering a 5- year study period.

The validity of epigenome wide association studies can be improved by utilising Inverse distance weighting and Ordinary Kriging models to estimate long term-exposure to the pollutants; nitrogen dioxide, ozone, particulate matter 10, and particulate matter 2.5. However, Ordinary Kriging provided the most accurate estimates for individual exposure estimates. As the IDW and OK models yielded similar results for particulate matter, the use of either of these models in epigenome wide association studies may also yield similar results of similar validity. However, when studying the effects of Nitrogen dioxide and Ozone, results suggest that the OK model is the most appropriate model to estimate exposure. Application of these models will help to delineate the relationship between long term exposure to ambient air pollution and health, and also help to understand what, if any, the role DNA methylation may play in mediating this relationship.

Here, I also sought to characterize pollution associated differences in DNAm in whole blood using the IlluminaEPIC BeadChip, which interrogates 850,000 sites across the genome. This study was not able to replicate or identify any novel epigenetic signatures related to background levels of air pollution, as measured by exposure to $PM_{2.5}$,PM_{10}, NO_2 or O_3. The lack of significant results with respect to this aim of our study should be taken with caution and several methodological issues must be considered. Previous work has reported contradicting results and whilst performing a detailed overlap of all CpGs previously linked to different measures of air pollution would be a task beyond the scope of this thesis, due to the vast amount of published work. I note that there is very limited overlap between any studies within the field and that studies of this nature are highly susceptible to yielding study specific results. This could be for a number of different reasons. Firstly, differences in the study region could result in large differences in exposure, studies focusing on cohorts in highly polluted areas such as India or China may result in largely different results compared to studies focusing on regions with lower air pollution levels like the UK or the Netherlands. Secondly, measuring air pollution is a very difficult task and with this, many studies take very different approaches to characterising an individuals exposure. For example, like ours, some studies may choose to focus on an individuals residential address, which will not capture accurately workplace exposures or outdoor levels of air pollution as accurately. Equally, studies which focus on solely workplace exposures will also miss pollution exposure from commuting behaviours and indoor air pollution in a participants home. Thirdly, as air pollution is a mixture of chemicals which may have different or combinatorial effects, this may be reflected as study specific results too.

Moreover, I demonstrate how an extremely small fraction of a study sample is capable of potentially driving genome wide significantly results. This highlights the importance for researchers within this field to consider the distribution of phenotype measurements and investigate further how this might be influencing EWAS results. Further, this also raises the question of how much of the current literature may have also suffered from similar outliers. Whilst exploring this was beyond the scope of this thesis, purely due to the fact that this raw data is never made available in such EWAS'. It still raises concerns to the robustness and reliability of such studies. Expanding on the importance of considering study sample characteristics, it is important to highlight that our results do not necessarily support the hypothesis that air pollution does not induce DNA methylation changes. Yet, it may suggest that background levels of air pollution in the UK are perhaps not high enough to produce genome wide significant results. Thus,focusing on study samples in other regions with higher levels of background air pollution levels may reveal more insights into the relationship between air pollution and DNA methylation changes.

This idea is reinforced further, as I was able to identify 531 significant DMPs (raDMPs) and 3 significant DMRs (raDMRs) related to TRAP exposure as measured using a distance to road measure. Further, I find that these regions are slightly enriched at enhancers and 5'UTRs, indicating a potential functional role.

I utilised the Illumina EPIC array technology, which has an increased coverage of distal regulatory elements and functional genomic regions. I observed differential methylation in CpG shores and open sea regions, but not in CpG islands or CpG shelves. Interestingly, CpG shores have previously been described as the 2kb sequence flanking a CpG island. Moreover, these regions have been reported to have a strong relationship with gene expression and to have a functional role [Bibikova et al., 2011, Irizarry et al., 2009]. Although, I was unable to reveal any enriched biological pathways for these CpGs, but was able to annotate these to additional genes via chromatin loop analysis and also to transcription factor motifs.

Additionally, I was able to identify novel and known loci previously related to air pollution. For example, the top raDMP, cg02292450 located on the CSMD3 gene was positively associated to a measure of traffic related air pollution (TRAP) exposure. CSMD3 is a protein coding gene, known as CUB And Sushi Multiple Domains and is thought to be involved in dendrite development. Interestingly, this gene has been reported to undergo expression changes in relation to hexanal exposure in rats, a known major component of indoor air pollution [Cho et al., 2017]. Similar findings have also been reported in humans, one study analysed the somatic mutations of 164 non-small cell lung cancers (NSCLCs) from XuanIi and control regions (CR) where CSD3 was reported to be mutated in air pollution related lung cancer [Yu et al., 2015]. This idea is reinforced further with additional research identifying CSMD3 as a gene showing mutations between patients at low vs high risk of lung adenocarcinoma [He et al., 2022]. Considering the well established link between exposure to air pollution and lung function decline or cancer, observation of differential methylation at this loci, offers a very interesting avenue for future research considering the genetic and epigenetic interplay of air pollution induced lung cancer. In contrast, the second most significant CpG is a novel raDMP found on chromosome 4 located at LOC285419, a non coding RNA with little to no literature reported on it. One study reports an association between a SNP at this loci with amyotrophic lateral sclerosis outcome in US veterans. Although just a hypothesis, given the link between air pollution exposure and neurological disorders, this is yet another loci that could warrant further investigation. On the other hand, the third most significant raDMP located on chromosome 6, was associated with the gene SYCP2L (synaptonemal complex protein 2). Interestingly, this gene has previously been linked to a causal SNP related to car or vehicle pollution [D'Antona et al., 2022] and exposure to herbicides [Rytel et al., 2021].

These results collectively support the hypothesis that DNA methylation differences in response to traffic related air pollution exposure may account for some of the links identified between air pollution and poor health. Future work should aim to replicate these findings in a separate cohort to obtain a robust catalogue of high confidence raDMPs.

Chapter 4

The role of DNA methylation in autosomal sex differences

4.1 Introduction

Sex is an important covariate in all epigenetic research due to its role in the incidence, progression and outcome of many phenotypic characteristics and human diseases [Beery and Zucker, 2011, Credendino et al., 2020].There is an increasing interest as to which role epigenetic modifications (such as DNA methylation) may play in the underpinnings for relationships between environmental exposures and disease onset. In addition, sex has previously been shown to have a strong influence on DNA methylation variation [Hartman et al., 2018],[Davegårdh et al., 2019],[Qin et al., 2019],[Xia et al., 2021],[Koo et al., 2020],[Xia et al., 2021]. However, the idea that DNA methylation variation between males and females may underlie the sex biases observed in diseases has not been well documented thus far.

Sex differences in disease prevalence are sometimes explained at the molecular level and rooted in genetic differences between males and females. Differences in sex chromosome complement have independently been shown to direct differences in gene expression and chromatin organization [Smith-Bouvier et al., 2008],[Wijchers and Festenstein, 2011],[Link et al., 2013],[Werner et al., 2017]. Furthermore, these differences in sex chromosome complement are sufficient to explain sex bias seen in some diseases. For example, X chromosome number has previously been shown to impact immune cell population and occasionally therefore the development of diseases such as autoimmunity [Fish, 2008],[Rubtsova et al., 2015].

Previous research has also revealed sex differences in gene expression of autosomal genes as well as sex chromosome linked genes [Andrews et al., 2022]. It is worth noting that most of the differences in gene expression on the autosomes are small differences [Lopes-Ramos et al., 2020a]. However small expression differences may still be associated with great effects on phenotypic characteristics and disease incidence and onset. Others also identified sex differences in chromatin accessibility and histone modifications, thus suggesting that different epigenetic factors contribute to gene expression sex biases seen in some diseases [Sugathan and Waxman, 2013].

Sex specific gene expression and levels of sex hormones may be mediated by epigenetic mechanisms, including DNA methylation. Several genome wide association methylome studies (or Epigenome Wide Association Studies, EWAS) have highlighted differences in DNA methylation patterns linked to sex differences in genes on the autosomes [Liu et al., 2010],[Numata et al., 2012],[Sugathan and Waxman, 2013],[Lopes-Ramos et al., 2020a]. Previous studies have reported sites and regions showing varying methylation due to sex differences in several tissues such as saliva, placenta, brain, pancreatic islets and whole blood [Liu et al., 2010],[Yousefi et al., 2011],[Price et al., 2013],[Hall et al., 2014],[Sun et al., 2014],[Inoshita et al., 2015],[Singmann et al., 2015],[Martin et al., 2017],[Suderman et al., 2017],[Lopes-Ramos et al., 2020a],[Gatev et al., 2021].These studies highlight the presence of autosomal loci displaying sex biased DNA methylation patterns across the genome for several tissues. In order to determine their role in disease and developmental processes, these loci warranted further exploration.

However, due to X chromosome inactivation in females, large differences in methylation levels of X chromosomes can be observed between males and females

[Wang et al., 2021]. Recent research suggests that normalising methylation data with the sex chromosomes introduces a large technical bias to many autosomal CpGs [Wang et al., 2022]. This technical bias has been reported to result in many autosomal CpG sites being falsely associated with sex even when male and female samples are normalised independently of each other, a method employed by some studies in the field. Moreover, it also leads to many autosomal CpGs being incorrectly identified to be more methylated in male samples compared to female samples. Therefore, the breadth of autosomal DNA methylation variation between males and females is still not well understood and requires further clarification. Extra steps were therefore employed in this study by applying a normalisation method which aims to reduce bias introduced to autosomal CpGs [Wang et al., 2022] to uncover true biological differences and determine patterns of global DNA methylation levels between males and females.

Additionally, it is worth noting that thousands of autosomal CpGs do show very small differences in DNA methylation patterns between males and females. However, a robust and well annotated catalogue of sites showing the largest differences still needs to be characterised.

Here, I use the EPIC BeadChip to assess autosomal sex differences in DNA methylation levels from whole blood at individual sites and genomic regions. All individuals involved in this study were part of Understanding Society: The UK Household longitudinal study [Bao et al., 2022]. Additionally, I adequately handle the technical bias introduced by sex chromosomes. To our knowledge, this is the largest study using the Illumina EPIC BeadChip (allowing for interrogation of 850,000 sites across the genome) to investigate autosomal sex differences in DNA methylation at CpG sites in whole blood.

4.2 Results

4.2.1 Females show higher methylation at a subset of autosomal loci

Analysis of DNA methylation (DNAm) differences between males and females on the autosomes was performed using linear regression for the IlluminaEPIC BeadChip for 1171 individuals (682 females and 489 males) for discovery and repeated in a validation data set of 2471 participants (1339 females and 1132 males). After data processing and cleaning, n=747,302 CpGs were analysed (see chapter 2). Sites which are known SNP probes, cross hybridizing or X/Y linked probes were excluded. Moreover, since whole blood is a bulk tissue, I calculated the estimated cell type proportions for whole blood between our male and female samples to assess whether any differences in cell type proportions would potentially be reflected in our results resulting in false positives. Using Wilcoxon test, I found no significant difference in the proportions of Granulocytes between males and females, but I did find statistically significant differences in proportions of CD4T, CD8T, Natural killer, B cells and monocytes (see Figure 4.2B and D). I therefore included cell type proportions in our models for identifying sex associated differentially methylated probes and regions. After adjusting for multiple testing using the Benjamini Hochberg FDR method (FDR $p < 0.05$) I identified 54,261 autosomal CpGs associated with sex in our discovery and validation data set (Figure 4.1). Of those CpGs, 60% (33,103 CpGs) were more highly methylated in females and the remaining 40% (21,788 CpGs) were more methylated in males. Gene ontology analyses showed several enriched terms for genes containing these 54,261 autosomal CpGs (see Table 4.1) which included terms related to mammalian sex determination and gonad development, specifically

Several signalling pathways such as Ras signalling, MAPK signalling, Wnt and Hippo signalling [C et al., 2007],[Nef and Vassalli, 2009],[SP and D, 2015],[Jiménez et al., 2021].Other terms included pathways related to cancer and cellular prolifera-

Table 4.1: Enriched GO terms among the genes containing the 54,261 CpGs identified to be significantly associated with sex

Path	Description	N	DE	P.DE	FDR
hsa04020	Calcium signalling pathway	240	192	4.10E–09	1.41E–06
hsa04015	Rap1 signalling pathway	210	169	1.15E–07	1.98E–05
hsa05200	Pathways in cancer	531	382.8333	1.91E–07	2.19E–05
hsa04014	Ras signalling pathway	232	180	2.97E–07	2.56E–05
hsa04010	MAPK signalling pathway	294	223.33333	2.24E–06	0.00015382
hsa04360	Axon guidance	182	148.5	3.46E–06	0.00019826
hsa04072	Phospholipase D signalling pathway	148	121	4.51E–06	0.00022149
hsa04310	Wnt signalling pathway	166	129.5	7.50E–05	0.00322397
hsa04371	Apelin signalling pathway	139	107.5	0.00011119	0.00363348
hsa04724	Glutamatergic synapse	114	93	0.00011675	0.00363348
hsa04390	Hippo signalling pathway	157	123.5	0.00012001	0.00363348
hsa01521	EGFR tyrosine kinase inhibitor resistance	79	67.5	0.00013811	0.00363348
hsa04071	Sphingolipid signalling pathway	119	94.5	0.00013944	0.00363348
hsa05226	Gastric cancer	149	115.5	0.00014787	0.00363348
hsa04550	Signalling pathways regulating pluripotency of stem cells	143	108.5	0.00059275	0.01359379
hsa04151	PI3K–Akt signalling pathway	354	245	0.00076003	0.01634056
hsa05224	Breast cancer	147	111.5	0.00092102	0.01863701
hsa04725	Cholinergic synapse	113	88	0.00148428	0.02836619
hsa04961	Endocrine and other factor regulated calcium reabsorption	53	44	0.00209794	0.03798375
hsa05225	Hepatocellular carcinoma	168	123.5	0.00284431	0.04892215

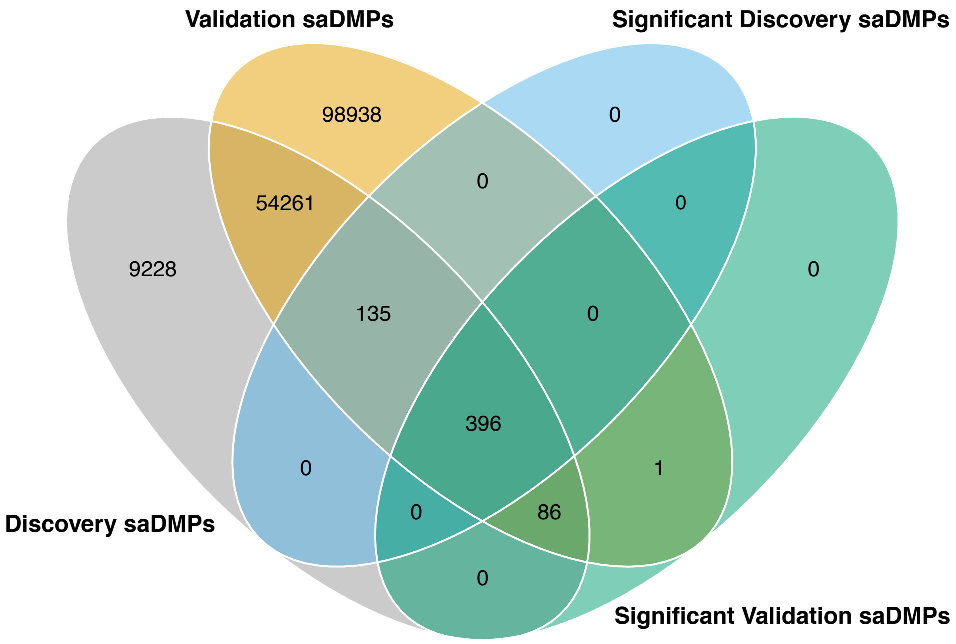


Figure 4.1: Venn diagram showing overlap of differentially methylated positions identified in our validation and discovery data set before and after filtering of the list of saDMPs

tion 4.1. This is not surprising though, as there is overwhelming evidence that sex influences cancer risk, progression, and treatment response [Lopes-Ramos et al., 2020a],[Lopes-Ramos et al., 2020b],[Zhu and Boutros, 2021],[Rubin, 2022]. It is also now well accepted that sex differences may significantly impact on the cell biology of cancer [Rubin et al., 2020]. Furthermore, epigenetic dysregulation is also now accepted widely as a mechanism for cancer initiation and progression. This may be through transcriptional activation or repression of specific autosomal loci through means of DNA methylation. Therefore, I hypothesise that sex specific patterns may influence the ability of cancer cells to adopt a stem cell like phenotype. This enables us to draw a link between epigenetic signatures and cancer pathways. It is likely that these sex differences in DNA methylation in part cause or are caused by differing levels of sex hormones such as androgen or oestrogen. This idea is supported by previous literature highlighting that DNA methylation transcriptionally represses masculinizing genes and that this depends on gonadal hormones during development [Rubin et al., 2020].

The Q-Q plots are slightly high (see Figure 4.2A and C) indicating small inflation of test statistics and, in order to ensure I detect true sex differences, I selected CpGs that displayed large differences in methylation. Thus, I further filtered our list of 54,261 CpGs by only considering those probes that displayed the largest sex differences, determined by a Beta value (absolute difference between average Beta values in male and female samples) greater than 0.05. A total of 396 CpGs met this criterion (called sex associated DMPs or saDMPs) in both our validation and discovery data sets and from here on, are the focus of this chapter (see Figure 4.3). A link to the full list of these sites are included in the Appendix A. CpG sites which I identified to have higher methylation in females are from here, referred to as 'female biased CpGs' and CpG sites which have higher methylation in males are here on referred to as 'male biased CpGs'. I found that these saDMPs were distributed across all autosomes (see Figure 4.3) with 74% of the saDMPs being female biased CpGs (293 CpGs) and 26% being male biased CpGs (103

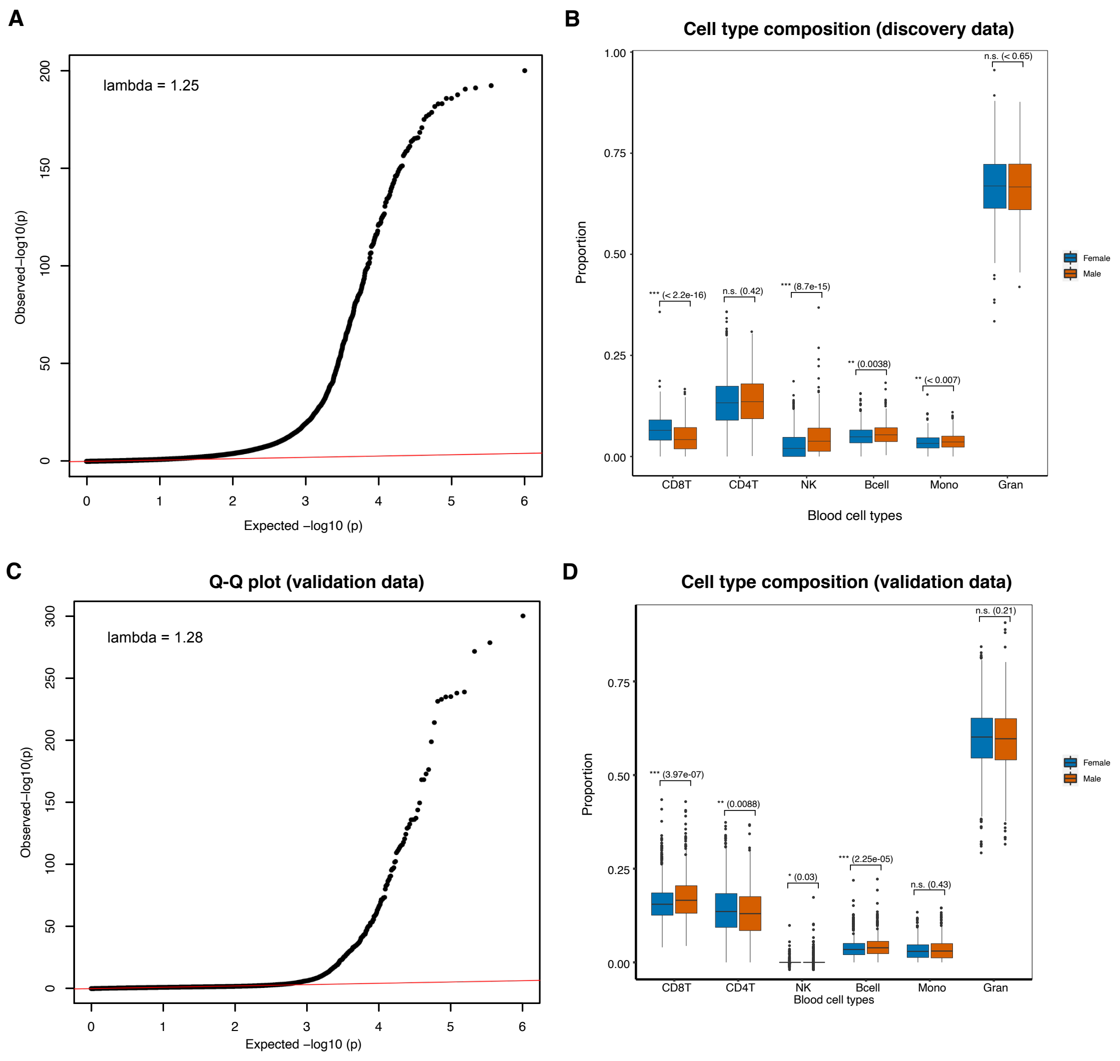


Figure 4.2: (A) QQ plot and lambda values (discovery data) distribution of the adjusted p values (FDR) against the null distribution for EWAS of sex in the understanding society cohort. Genomic inflation lambda score is indicated in the QQ plot to indicate statistical inflation of p values. (B) Boxplots of estimated whole blood cell type proportions for males (orange) and females (blue) in the discovery data set, estimated using the estimateCellCounts function from bigmelon. I performed a Mann-Whitney U test (p value: n.s. 0.05, *p value $<$ 0.05, **$<$ 0.01 and ***$<$ 0.001). (C) QQ plot and lambda values (validation data) distribution of the adjusted p values (FDR) against the null distribution for EWAS of sex in the understanding society cohort. Genomic inflation lambda score is indicated in the QQ plot to indicate statistical inflation of p values. (D) Boxplots of estimated whole blood cell type proportions for males (orange) and females (blue) in the validation data set, estimated using the estimateCellCounts function from bigmelon. I performed a Mann–Whitney U test (p value: n.s. 0.05, *p value $<$ 0.05, **$<$ 0.01 and ***$<$ 0.001)

CpGs) (see Figure 4.3B).

Since I had such stringent parameters to define a significantly associated saDMP for males and females, I performed principal component analysis (PCA) to see how male and female beta values clustered in PC space and to evaluate the effect of DNAm at the saDMPs. As shown in Figure 4.3C, male and female samples formed distinct clusters based on the beta values of the significant sex associated DMPs (396 CpGs). PC1 explained 16.1% of the variance and PC2 explained 4.2% of the variance. Based on Figure 4.3C I can conclude that these saDMPs are sufficient to contribute to the clear separation of male and female samples in PC space.

4.2.2 Characterisation of sex associated DMPs

The saDMPs were found in 174 unique genes with 48 of these genes harbouring several saDMPs (see Figure 4.3D). The number of saDMPs harboured by individual genes ranged from 1-8. CRISP2, a gene known to be involved in sperm function and male fertility [Lim et al., 2019], harboured the largest number of saDMPs, 8, which interestingly were all found to be female biased CpGs. I performed GO and KEGG analyses but did not identify any significantly enriched biological processes or pathways for these genes. Nevertheless, the genes which did harbour saDMPs are biologically relevant, as many are genes known to be involved in sexual development and processes, such as SOX18 [Prior and Walter, 1996]. Further, some genes are already known to exhibit sex specific methylation patterns, such as PRR4 and PTPRN2 [I et al., 2020],[Kochmanski et al., 2021]. Despite this, I was able to identify some novel genes which have not previously been reported to exhibit sex differences in DNA methylation such as GCK, HIP1R and KANK1.

To help me gain more insight into the functional role of these saDMPs, I characterised their genomic location and further compared this with the autosomal EPIC background. I found that saDMPs are preferentially located in CpG islands and CpG shores and depleted in open sea regions compared to the autosomal background (see Figure 4.4A and C). Moreover, female biased CpGs are enriched

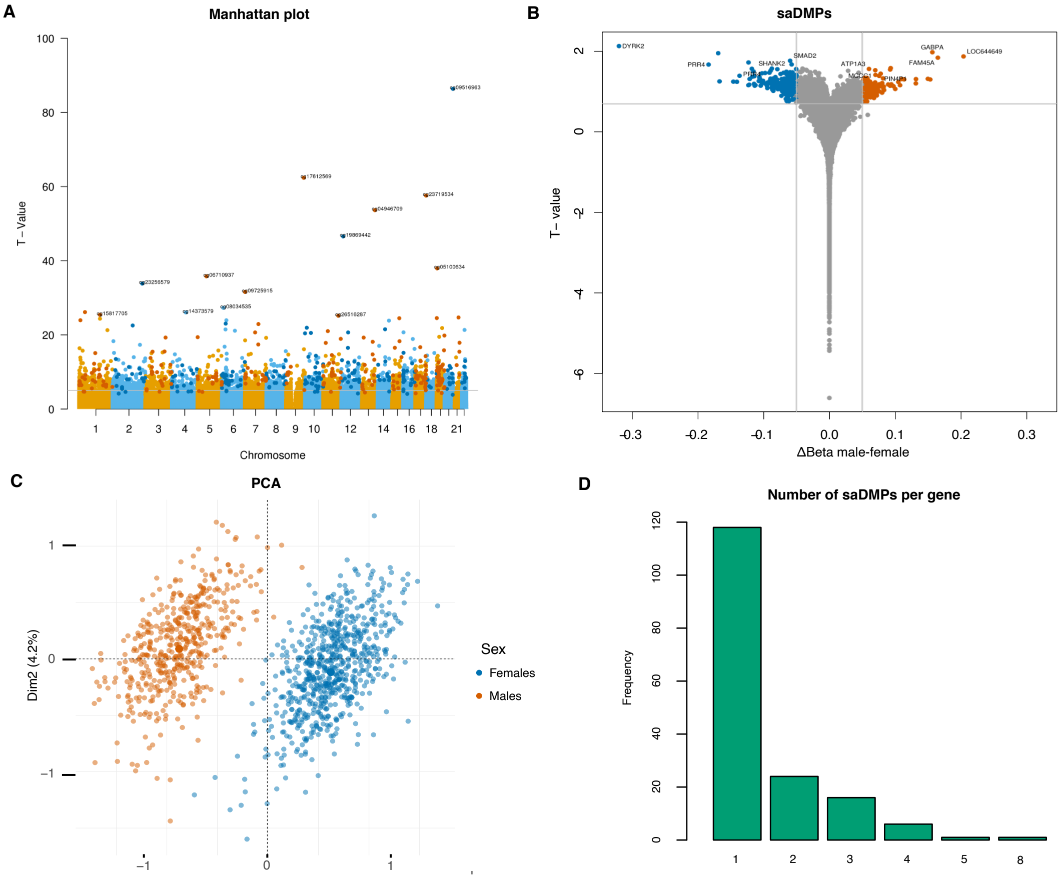


Figure 4.3: Location and characterisation of saDMPs. (A) Manhattan plot for EWAS analysis of sex. CpG sites which met a threshold of FDR < 0.05 and had an average beta change of > 0.05 and found in both discovery and validation data sets were considered significant and are represented by darker colours. B) Volcano plot for saDMPs. CpGs which are not significant in both the discovery and validation data sets are represented in grey, replicated saDMPs male-biased CpGs are in orange and replicated saDMPs female-biased CpGs in blue. Grey points displayed beyond the cut off points represent CpG sites which were met the criteria in the discovery data set (FDR < 0.05 and deltaBeta value $>$ than 0.05 in any direction) but were not replicated in the validation data set. C) Principal component analysis of beta values at the significant saDMPs. Male samples are indicated in orange while female samples are indicated in blue. D) Number of saDMPs harboured by individual genes

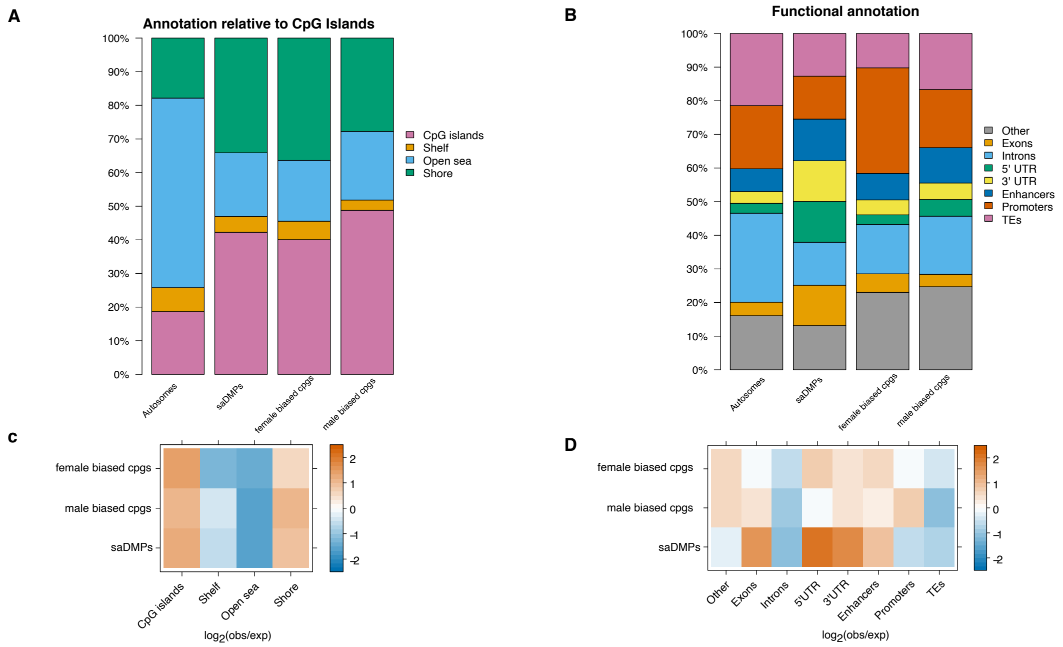


Figure 4.4: Location of saDMPs. A) Top panel shows the annotation of all saDMPs (n=396), female-biased CpGs (n=293) and male-biased CpGs (n=103) relative to CpG island regions compared to the autosomal background. Bottom panel shows the log2 (obs/exp) annotations based on the autosomal background of the different annotations. B) Top panel shows the overlap of all saDMPs (n=396), female-biased CpGs (n=293) and male-biased CpGs (n=103) with genomic features compared to the autosomal background. Bottom panel shows the log2 (obs/exp) annotations based on the autosomal background of the different annotations.

at 5' UTRs and enhancers, with male biased CpGs being enriched at promoters and exons (see Figure 4.4B and D).Interestingly, I observed that all saDMPs display enrichment at enhancers, which, together with their presence at promoters, indicates that they could play a role in gene regulation. Lastly, I also note that all saDMPs were depleted at transposable elements and introns compared to the autosomal EPIC background.

Enrichment of saDMPs at enhancers suggests that some of the saDMPs could potentially regulate distal genes [Chathoth and Zabet, 2019],[Nasser et al., 2021]. I further annotated the saDMPs to genes by identifying if their contacts with

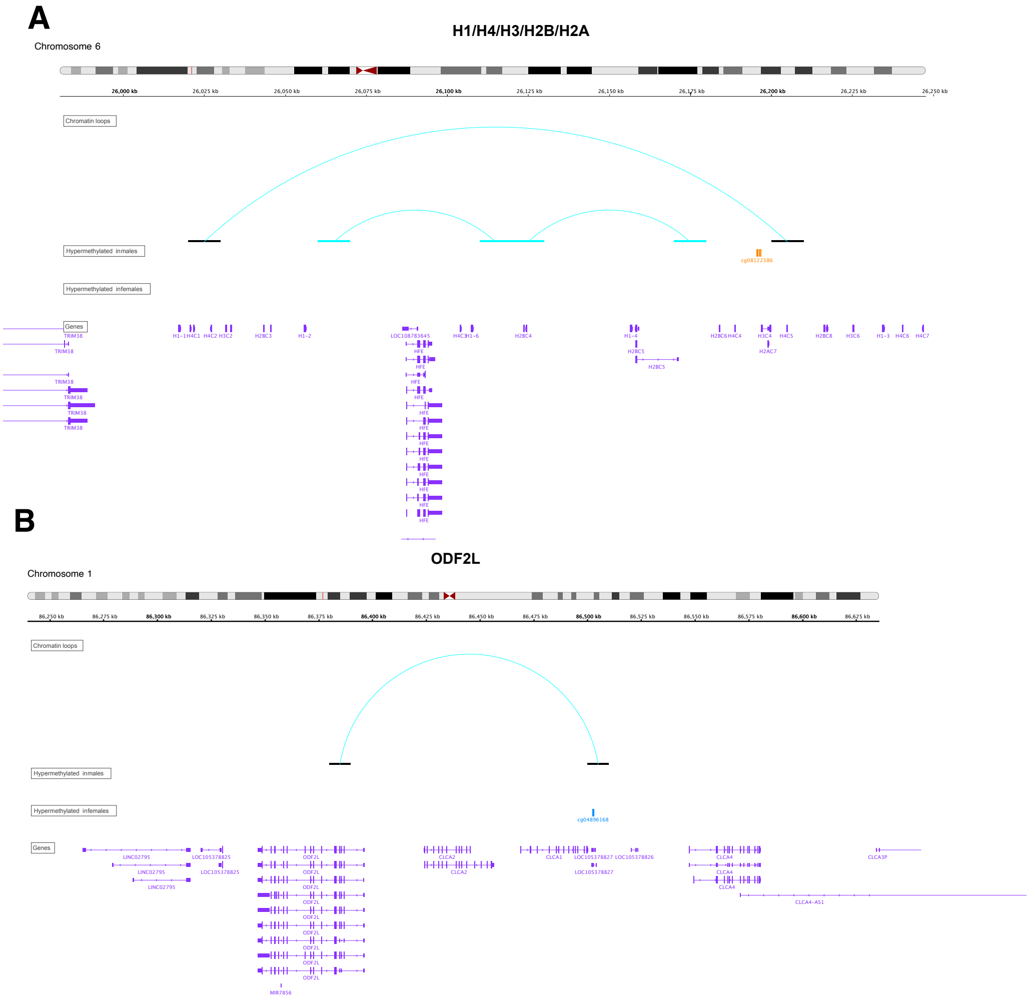


Figure 4.5: Chromatin loops connecting saDMPs to additional genes. A) Integrated genomics viewer track of chromatin loop on chromosome 6 showing two male-biased CpGs contacting H1/H4/H3/H2V/H2A. B) Integrated genomics viewer track of chromatin loop on chromosome 1 showing a female-biased CpG contacting the ODF2L gene. Blue lines represent the chromatin loops, with black lines showing the loop anchors. Orange vertical lines represent the male-biased CpGs and blue vertical lines represent the female-biased CpGs. Purple annotations represent genes.

promoters are mediated by 3D chromatin loops detected in Hi-C data. Following this, I further annotated the saDMPs to 37 additional genes, 28 of them being annotated to female biased CpGs and 8 to male biased CpGs (see Figure 4.5).

Of the 8 genes linked to male biased CpGs, I found three histones (HIST1H3A, HIST1H4A and HIST1H4B), which are known to interact with CDYL. Chromodomain Y-like protein (CDYL) is a chromatin reader binding to heterochromatin (H3K9me3, H3K27me2 and H3K27me3) that is crucial for spermatogenesis, male fertility and X chromosome inactivation [Xia et al., 2019]. In addition, ODF2L; outer dense fiber of sperm tails 2 like is linked to saDMPs female biased CpGs and has previously been shown to interact with PRSS23, which is involved in ovulation [Dimas et al., 2012].

Next, to evaluate whether the genes controlled by the saDMPs are part of the same regulatory network, I merged all proximal and distal genes and produced protein-protein interaction networks to visualize the networks of these genes. Following this, I was able to identify the top 30 hub genes by evaluating each gene by its network connectivity. The results for these analyses are produced in (see Figure 4.6).. The top hub gene (ranked by the maximum clique centrality method) for the male biased CpGs in males was HIST1H4B and the top hub gene for female biased CpGs was SLC17A7.

4.2.3 Enrichment of saDMPs in transcription factor binding sites

To identify common features among the sex associated DMPs, I performed transcription factor (TF) binding site and gene ontology analyses. First, I evaluated whether the saDMPs were enriched in motifs for TFs (100 bp window). For the 293 female biased CpGs I found 315 unique enriched TFs (p.value < 0.05) (see Figure 4.7) with strongest evidence for FOXB1, TIA1 and XRCC1. These are genes not previously reported to exhibit any sex differences or be enriched at areas exhibiting any sex differences. I did however find some TF motifs enriched which

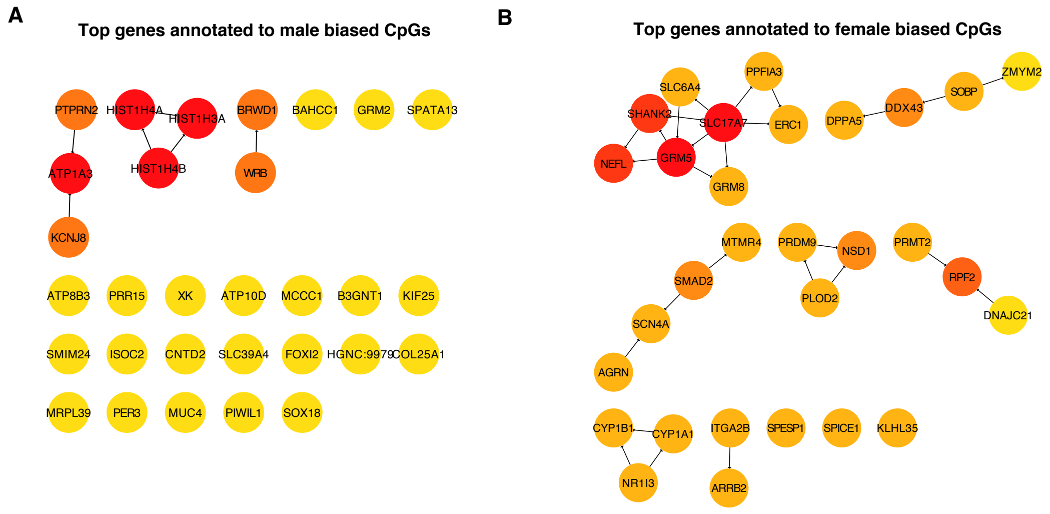


Figure 4.6: Subnetworks of the top 30 genes annotated to male-biased CpGs (C) and females (D). Node colour represents the degree of connectivity. The scale from red to yellow represents the top 30 enriched genes rank from 1-30, with red indicating highest degree and yellow indicating lowest degree.

have previously been shown to play a role in sexual development and hormone levels. For example, I found SOX13, SOX21 and SRY TF motif to be enriched in female biased CpGs, which are known to be involved in male sex determination [Harley et al., 2003a],[Li et al., 2014b].For the 103 male biased CpGs, I identified 64 enriched TFs, including ESR1 which encodes the oestrogen receptor, TCEAL6 HLCS and GPD1 (see Figure 4.7).

To analyse whether the TF motifs were enriched for annotation to biological processes or pathways, I performed pathway analyses using the GO and KEGG databases. I identified several enriched KEGG pathways for the TFBS enriched at female biased CpGs, spanning a wide range of processes such as transcriptional misregulation in cancer, several specific cancer pathways, PI3K-Akt signalling and more (see (see Figure 4.7). In addition, I also found 39 enriched GO terms ranging from transcription factor activity, E-box binding, transcription coactivator activity and interestingly, bHLH transcription factor binding (see Figure 4.8). Nevertheless, I found no enriched KEGG terms for the TFs enriched at male biased CpGs, likely

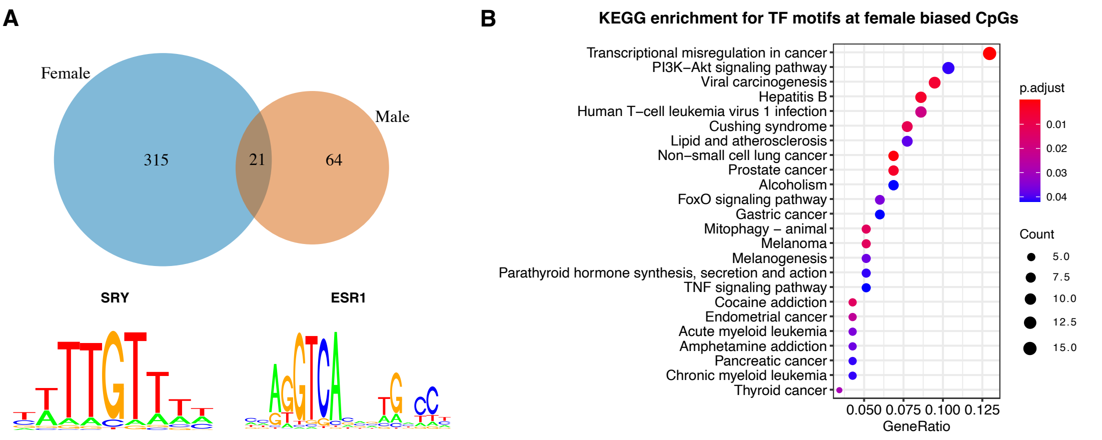


Figure 4.7: Transcription factor motif enrichment analysis. A) Overlap of enriched TF motifs for female-biased CpGs (blue) and male-biased CpGs (orange). saDMPs were enriched in TF binding motifs including SRY and ESR1. B) KEGG analyses for the significantly enriched TF motifs at female-biased CpGs

due to the small number of enriched TFs. However, I identified several enriched GO terms such as NAD, NADP binding and oxidoreductase activity (see Figure 4.8).

As I identified enrichment for some transcription factors encoded for on the sex chromosomes (e.g. SRY), I hypothesised that sex chromosome encoded transcription factors may influence CpG methylation at the saDMPs directly or indirectly by acting as hub genes in the enriched TF motif network. To assess this, I firstly produced protein-protein interaction networks to visualize the networks of these TFs (see Figure 4.9). Although I identified some enriched motifs for several TFs encoded on the X chromosomes in the male biased CpGs such as ELK1, TGIF2LX and TCEAL6 (see Figure 4.9A), I observed that they were not central nodes in the network. Nevertheless, I did identify several central TF motifs encoded on the sex chromosomes for the female biased CpGs (see Figure 4.9B). These included 15 TFs encoded on the X chromosome and 2 on the Y chromosome including SRY and KDM5D.

Secondly, I further utilised cytohubba a plug-in tool in cytoscape to robustly identify if these TFs were in fact hub genes in the network. This revealed that one TF encoded on the X chromosome (RPS4X) did in fact act as a hub gene

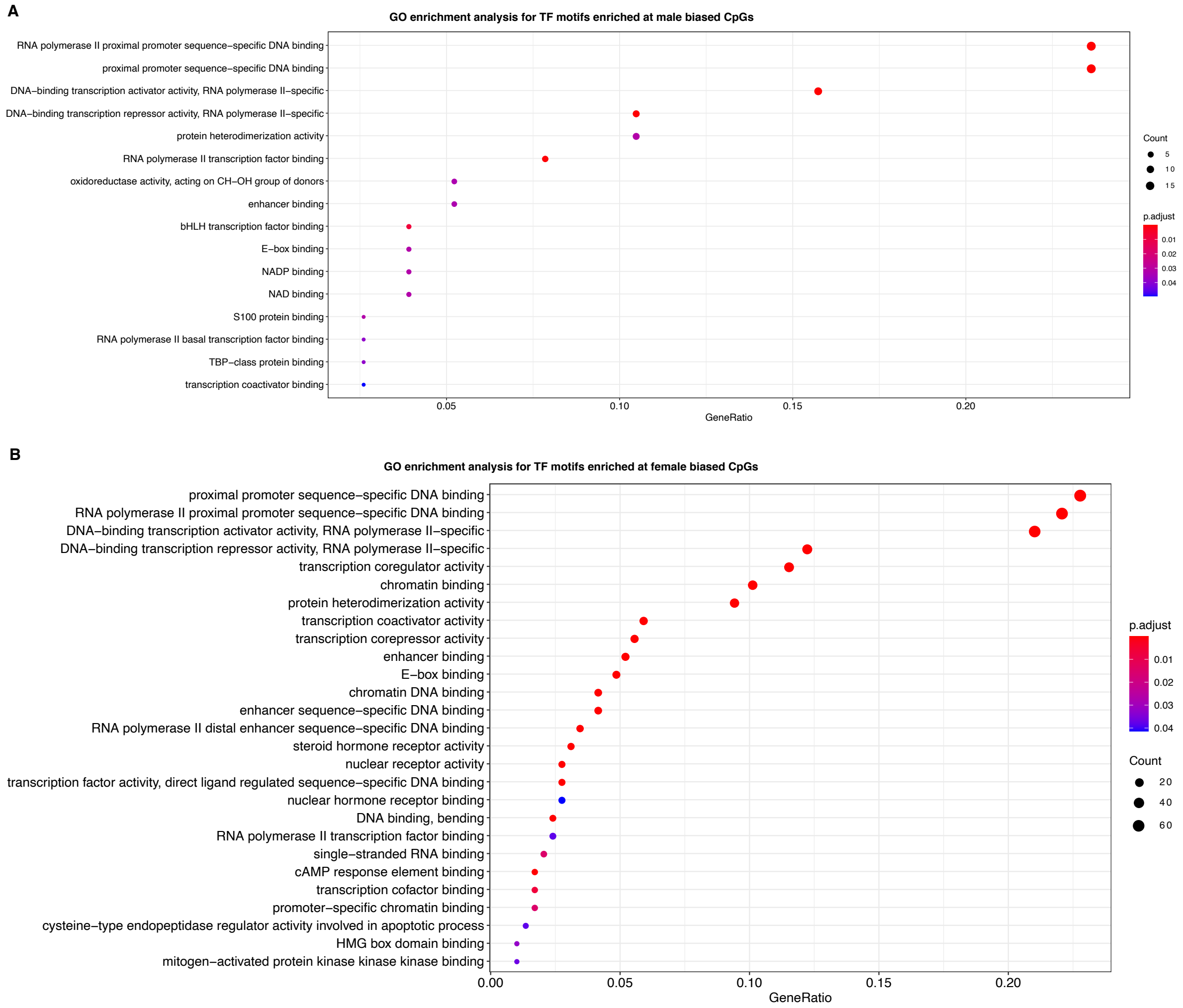


Figure 4.8: GO terms overrepresented for the significantly enriched TF motifs at male-biased CpGs (A) and female-biased CpGs (B).

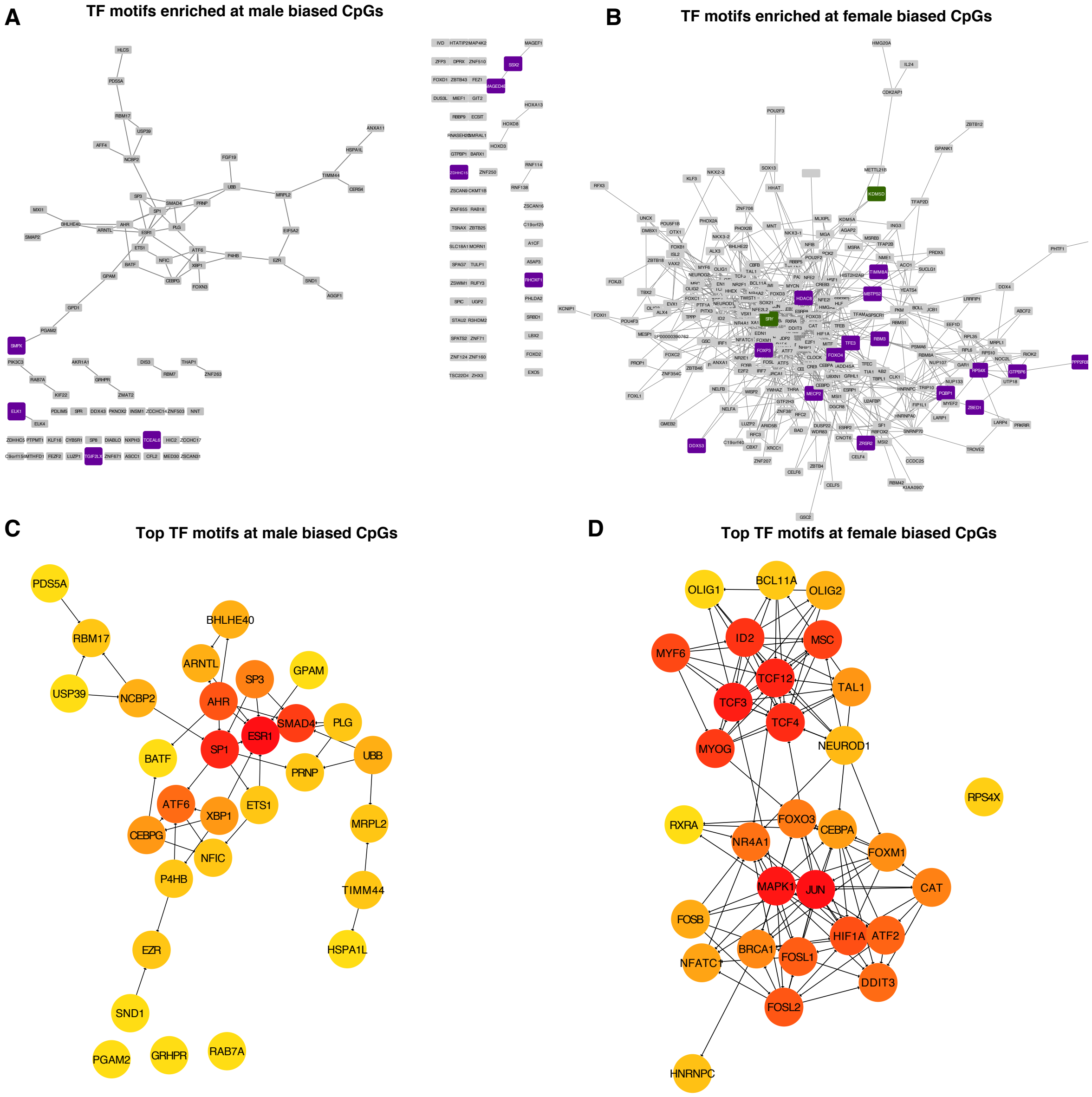


Figure 4.9: (A-B) Network visualisation of protein-protein interactions for all transcription factor motifs found to be enriched at male-biased CpGs (A) and female-biased CpGs (B). Grey coloured boxes represent individual TFs located on autosomes, while purple-coloured boxes represent TFs encoded on the X chromosome and green coloured boxes represent TFs encoded for on the Y chromosome. Grey lines represent edges between transcription factors within the protein-protein network. (C-D) Subnetworks of the top 30 enriched TF motifs at male-biased CpGs (C) and females (D). Node colour represents the degree of connectivity. The scale from red to yellow represents the top 30 enriched TF motif rank from 1-30, with red indicating highest degree and yellow indicating lowest degree.

in the TF network, however the other 29 genes were encoded on the autosomes (see Figure 4.9C-D). Furthermore, for the TF motifs enriched at female biased CpGs, I identified MAPK1, JUN and BRCA1 and other autosomal genes to be hub TFs in the network revealing novel TFs involved in sex differences (see Figure 4.9C-D). Moreover, for those TFs enriched at male biased CpGs, I identified SP1, ESR1 and SMAD4 to be hub genes in this network. Interestingly, SP1 is a gene known to influence SRY expression [Harley et al., 2003b] and ESR is the gene that encodes the oestrogen receptor and lastly, SMAD4, has previously been described as a female germ cell determinant [Wu et al., 2016]. This analysis suggests that although I did identify some sex chromosome encoded TFs to act as hub genes in the TF network, it is unlikely that they are responsible for affecting CpG methylation at these saDMPs.

4.2.4 Relationship with gene expression

The 396 saDMPs were then further explored in association with the expression levels of their annotated genes using publicly available data for whole blood poly(A)+ RNA-seq (GSE120312). The majority of the differentially expressed genes (DEGs) are located on the sex chromosomes, but I did also observe differential expression between males and females for several autosomal genes (see Figure 4.9). I did not identify any significant sex biased gene expression patterns corresponding to differences in DNAm levels at these genes (see Figure 4.9). This is not surprising as it has been previously reported that autosomal sex differences in DNA methylation result in nominal or no differences in gene expression [Suderman et al., 2017],[Gatev et al., 2021], a trend also seen with age specific DNAm marks [Hernando-Herraez et al., 2019]. Moreover, while other studies claim that they identify DEGs on autosomes between males and females, corresponding to differences in DNA methylation, when adjusting for multiple testing, it appears that these no longer hold statistical significance [Inkster et al., 2021a]. It is also important to note that, the relationship between DNAm with gene expression is a complex one, although it is

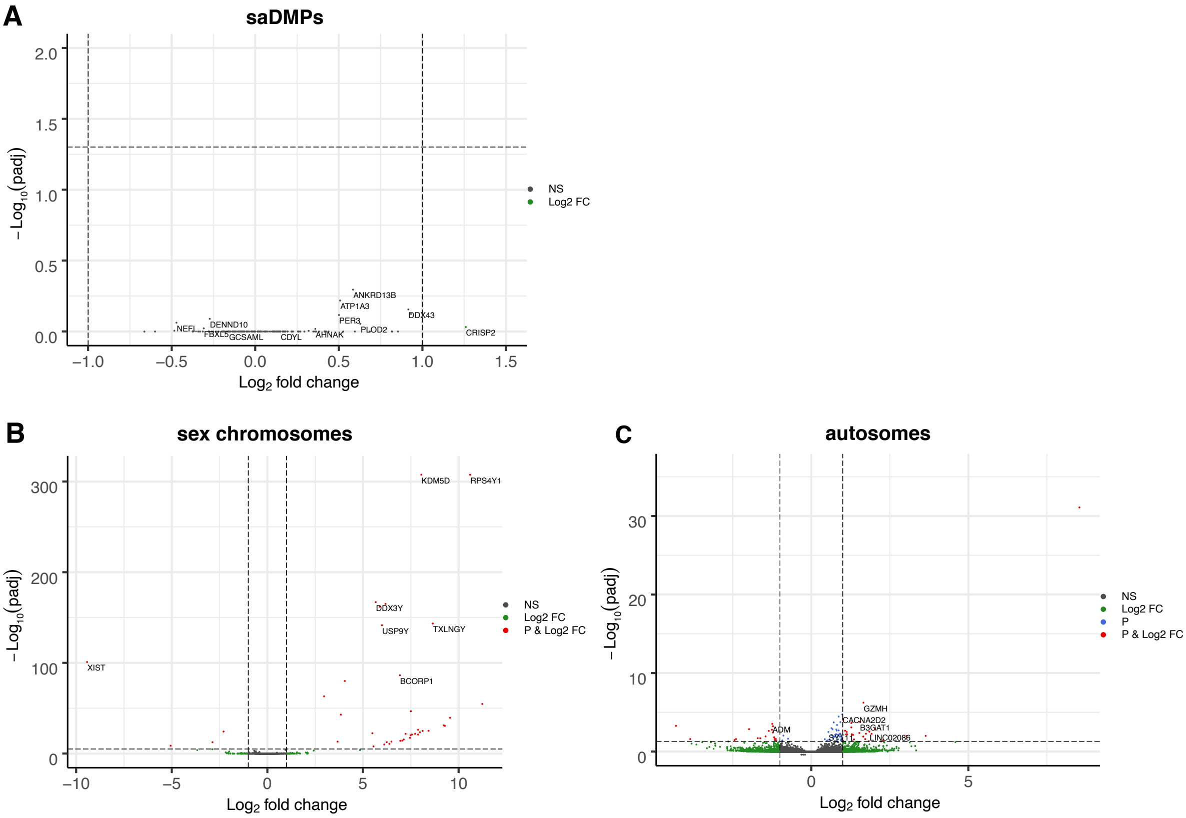


Figure 4.10: Volcano plot showing differential gene expression between males and females. I considered the case of: (A) genes annotated to the saDMPs, (B) sex chromosome linked genes and (C) autosomal genes. Points coloured in grey represent non differentially expressed genes. Green points represent genes which had a log2 Fold Change value greater than 1. Blue points represent genes which met the adjusted p value threshold (FDR <0.05). Points coloured in red represent genes which showed differential expression between males and females (adjusted FDR p value <0.05 and log2FC > 1).

generally thought that DNA methylation leads to gene repression, lots of literature reports methylation leading to active expression [MS et al., 2015],[Rauluseviciute et al., 2020],[Sadler et al., 2021] or that it is insufficient to repress transcription [Ford et al., 2017]. My results support the idea that differences in DNA methylation observed between males and females at these saDMPs do not lead to significant differences in gene expression.

4.2.5 Sex associated differentially methylated regions

Given that several genes harboured numerous saDMPs, I postulated whether some of the saDMPs were part of larger differentially methylated regions associated with sex. I therefore searched for differentially methylated regions associated with sex in our discovery and validation data set. Following adjustment for multiple testing (FDR) and adjustment for cell type proportions, batch effects and age, I identified 266 sex associated differentially methylated regions. I considered a saDMRs significant if it harboured at least 5 CpGs, had an FDR value smaller than 0.05 and had a methylation difference within the region greater than 0.05 in either direction and lastly, was present in both our discovery and validation data set. Following filtering of the list of saDMRs, I identified 266 significant sex associated DMRs on the autosomes between males and females located at 231 unique sets of genes. The number of CpGs within the DMRs ranged from 6 to 123 and had an average width of 2392 base pairs (bp) ranging from 178 to 14715 bp.

Figure 4.11 shows the beta values for males and females at 4 of the most significant saDMRs: The top hits in the saDMR list overlapped promoter regions of genes such as SDHD, TIMM8B, ATP5J, GABPA, GPN1, CCDC121, and PRKXP1. SDHD and TIMM8B are genes known to be influenced by oestrogen exposure [Bove et al., 2018] suggesting that sex hormones may underlie sex differences in autosomal DNA methylation, or alternatively that DNA methylation may mediate sex hormone levels. Moreover, ATP5J and GABPA are genes (male biased CpGs) which have previously been reported to be implicated in early onset of Alzheimer's disease [Kasuga et al., 2009],[Wiseman et al., 2015], a disease known to affect females more than males. Furthermore, ATP5J is a gene known to be a target gene of oestrogen, previously shown to serve an inhibitory role in the sex differences in hepatocellular carcinoma(Li et al., 2019). GPN1, CCDC121, ATP5J and GABPA have previously been shown to exhibit functions which are sex specific [Yousefi et al., 2011]. In addition, PRKXP1 is located on chromosome 15 and CpGs in this region have previously been associated with Crohns disease and intestinal inflammation, a disease which has previously been reported to be more prevalent in females [Somineni et al., 2016]. A saDMR harbouring 123 CpGs overlapped the promoter region of a gene called MCDC1, a gene known to direct chromosome wide silencing of the sex chromosomes in male germ cells, initiate meiotic sex chromosome inactivation (MSCI), and lead to XY body formation [Ichijima et al., 2011].

These findings are extremely important for epigenome wide association studies aiming to characterise sex specific effects in relation to exposures, a rising theme in the literature [Curtis et al., 2020],[I et al., 2020],[Koo et al., 2020],[Sunny et al., 2021],[Zhang et al., 2021].By providing a valuable resource for the community to disentangle whether particular sites or regions display sex differences in DNA methylation.

4.3 Summary and discussion

Here, I conducted the first study aiming to characterize autosomal sex differences in DNAm between males and females in whole blood using the IlluminaEPIC BeadChip, which interrogates 850,000 sites across the genome. Whilst I was able to identify thousands of autosomal CpGs displaying sex differences in DNA methylation, I focused the majority of my analysis on those autosomal CpGs displaying the largest sex differences in methylation levels (see Methods). I thereby, identified 396 sex associated differentially methylated positions on the autosomes. Previous work has reported contradicting results, some research report that there is higher methylation on autosomes in females [Yousefi et al., 2011],[Davegårdh et al., 2019],[Gatev et al., 2021] while other research reports identifying higher methylation on autosomes in males [Zhang et al., 2011],[Xu et al., 2014],[García-Calzón et al., 2018] and others reports no significant difference in DNAm on autosomes between males and females [Hall et al., 2014]. My results support the former and I found that 76% of these loci (293 CpGs) showed higher methylation in females compared to males.

Therefore, although the existence of these autosomal sex associated CpG sites is well established, a robust and consistent catalogue is yet to emerge. When comparing the saDMPs discovered in this study to findings previously reported

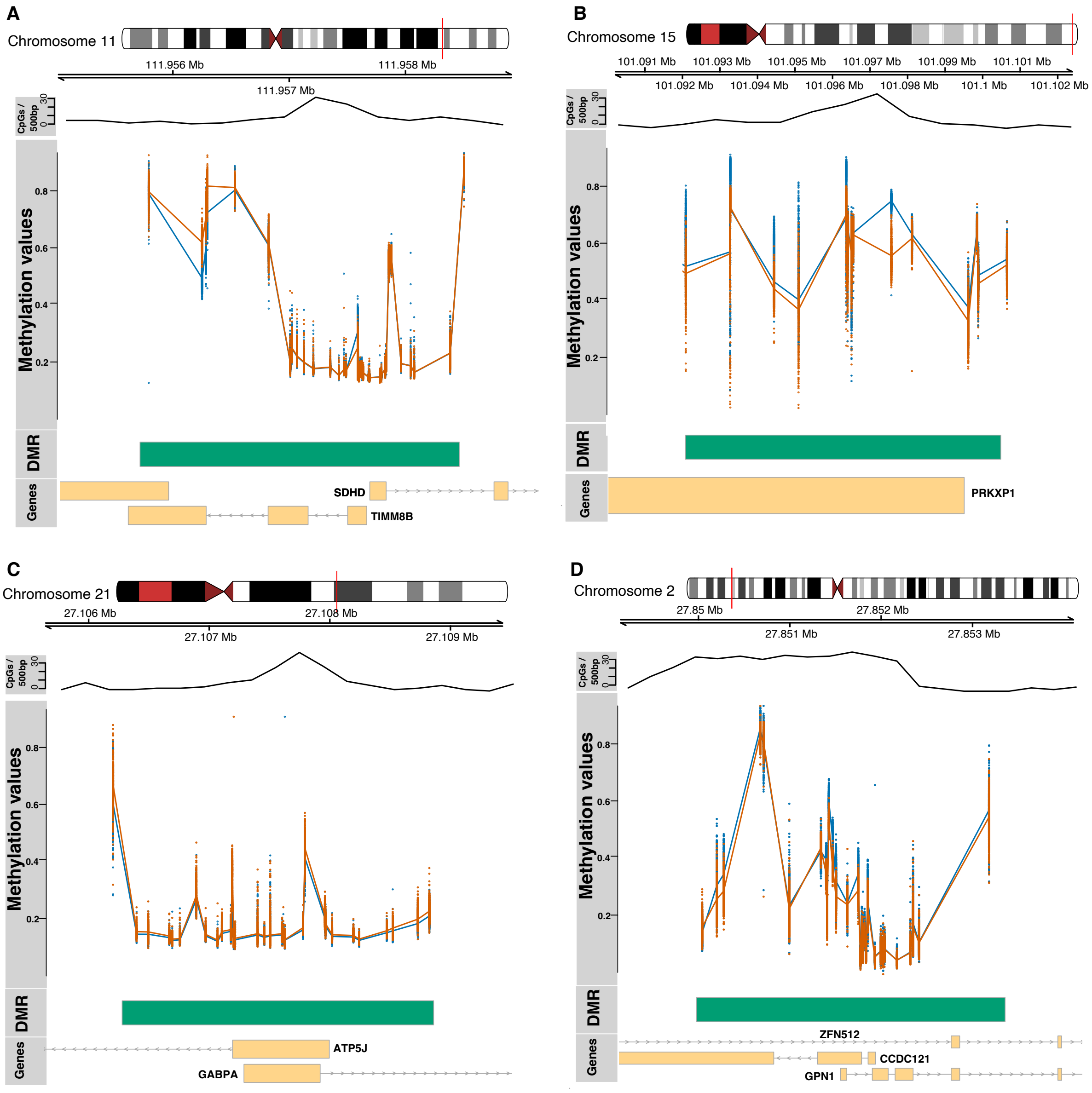


Figure 4.11: Plots of sex-associated differentially methylated regions (saDMR). I plotted regions: (A) SDHD and TIMM8B, (B) PRKXP1, (C) ATP5J and GABPA and (D) CCDC121 and GPN1. Yellow boxes represent appropriately labelled genes, green boxes represent the genomic region which the differentially methylated region spans. The scatterplots represent the beta values for males (orange) and females (blue) at CpG sites located within the differentially methylated region

in blood samples (where a full list of sex associated sites were available) I do observe some overlap, although it is limited (Table 4.3). For example, I identify 54% overlap of our identified saDMPs with previously reported sex associated sites in cord blood [Yousefi et al., 2011]. On the other hand, overlap from other research also investigating cord blood was lower, namely 36.87% [Maschietto et al., 2017]. Furthermore, I observe only 15% overlap of saDMPs identified in a previous study investigating sex differences in peripheral leukocytes [Inoshita et al., 2015]. I also additionally checked the overlap between our saDMPs identified in blood with those identified in other tissues (Table 4.3). Interestingly, I observe a 73% overlap with sex associated sites identified in placenta by Inkster and colleagues [Inkster et al., 2021b] and a 35% overlap with sex associated sites identified in post-mortem prefrontal cortex [Xu et al., 2014]. The existence of this overlap between different tissues shows that a portion of these saDMPs identified are conserved across tissues, and that some are tissue specific.

Table 4.2: Overlap of autosomal sex-associated differentially methylated positions reported in this study with previous literature in various tissues

Study	Tissue of interest	Sample size	Platform used	CpGs identified	% of saDMPs replicated (%)
Yousefi et al. (2014)	Cord blood	111 newborns	Illumina 450 k	3031	54
Mccarthy et al. (2014)	Meta analysis of 76 studies	6795	Illumina 27 k	184	0.54
Inoshita et al. (2015)	Peripheral leukocytes	117 adults	Illumina 450 k	292	15
Maschietto et al. (2015)	Cord blood	71 newborns	Illumina 450 k	2332	36
Xu et al. (2014)	Post-mortem prefrontal cortex	46 adults	Illumina 450 k	614	35
Hall et al. (2014)	Pancreas	87 adults	Illumina 450 k	470	18
Xia et al. (2021)	Post-mortem brain samples	1408 adults	Illumina 450 k	15,417	31
Inkster et al. (2021)	Placenta	293 adults	Illumina 450 k	162	73

However, alternative reasons for this limited overlap, may be attributed to differences in sample size, as the data sets used within this study are much larger than previously used thus increasing our ability to detect true sex differences in DNA methylation.

In addition, the the reduced overlap could be explained by the different normalisation methods applied to DNAm microarray data. Previous research has shown that the methylation levels of CpG sites on the X chromosome differ largely between males and females [Wang et al., 2021] and, thus, normalisation methods which normalise array data indiscriminately with CpG sites on the autosomes introduce large technical biases for autosomal CpGs [Wang et al., 2022]. Using normalisation methods which do not handle the technical bias introduced by sex chromosomes, will therefore lead to many autosomal CpG sites being falsely associated with sex and further, a higher number of autosomal CpGs being incorrectly identified as male biased CpGs. Our choice of normalisation method greatly reduced technical bias at autosomal CpGs for male and female samples.

Moreover, as thousands of autosomal CpG sites show differences in DNA methylation between males and females, differences in the methods for determining the definition of a sex associated site result in limited reproducibility between studies, a point also raised by Gatev and colleagues in their identification of sex associated regions [Gatev et al., 2021]. Here, I therefore proposed and applied stringent cut offs to define a sex associated site (FDR < 0.05 and effect size of at least 0.05 in either direction). Whilst I acknowledge that true but small differences in DNA methylation related to a phenotype may exist, in the interest of generating a reproducible and robust catalogue of saDMPs, I chose to apply effect size cut offs. Consistent with this, I was able to replicate 75% of my saDMPs identified in my validation data set in my discovery data set. Moreover, I found that 73% of the saDMPs I identified in this study were also identified by Inkster and colleagues [Inkster et al., 2021b], whom also applied effect size cut offs, demonstrating the reproducibility and robustness of our catalogue of saDMPs.

I further categorized these 396 saDMPs into two groups, those that were male biased CpGs (n=103) and those that were female biased CpGs (n=293). Several saDMPs found to be female biased CpGs overlapped the transcription start site (TSS) of genes not previously been reported to exhibit sex differences in DNAm including C19orf77, ATP10D and SHANK2. Interestingly, it has previously been shown that sex hormones can regulate SHANK expression leading to a sex differential expression in SHANK2 [Berkel et al., 2018]. Furthermore, this gene has previously been implicated in autism spectrum disorder, a disorder known to exhibit higher prevalence in males rather than females. In contrast, the most significant male biased CpG is located in the CpG island of a gene located on chromosome 21 called GABPA. GABP is a methylation sensitive transcription factor and has previously been shown to be a transcriptional activator of Cyp 2d-9, which is a gene encoding a male specific steroid in mice [N et al., 1995]. Sex differences in these regions have previously been identified by other studies investigating autosomal sex differences in DNA methylation. More specifically, Yousefi and colleagues also identified this region to be a top sex associated DMR in their analysis [Yousefi et al., 2011]. In addition, previous research investigating transcriptome wide sex differences using single cell RNA-seq data in mouse reports GABPA to be one of six TF families responsible for the majority of sex dimorphic transcriptional regulation activities [T and JC, 2020].

Interestingly, as well as GABPA being the gene annotated to the most significant saDMP (male biased CpGs), it was also the third most significant saDMR, suggesting these regions could account for important sex biases observed in some diseases. This is further supported by the fact that GABPA has also been associated with early onset of Alzheimers disease, Parkinsons disease, breast cancer and autism [Yokomori et al., 1995],[Lu and Mar, 2020],[Perdomo-Sabogal et al., 2016]. The saDMR harbouring the highest number of CpG sites (n=123) is located on chromosome 6, overlaps TUBB, MDC1 and XXbac. MDC1 is thought to play a crucial role in the production of male games, lead to XY body formation and also initiate meiotic sex chromosome inactivation. These functions are achieved through its interaction with DNA damage response (DDR) factors, ultimately leading to transcriptional silencing [EA et al., 2007],[Ichijima et al., 2011].These results collectively support the hypothesis that sex differences in autosomal DNAm may account for some of the sex differences seen in disease prevalence, onset and progression.

Moreover, I did identify saDMPs in genes known to exhibit sex differences in DNAm such as CRIPS2 and DDX43 which are involved in spermatogenesis and male fertility [Yousefi et al., 2011],[Lim et al., 2019]. Specifically, CRIPS2 harboured 8 significant saDMPs, all female biased CpGs, and is part of a group of proteins called CRISPs which show male biased expression in the male reproductive tract. CRIPS2 plays an important role in spermatogenesis, acrosome reaction and gamete fusion [Lim et al., 2019].Some of my saDMPs were located in genes known to show sex by age effects, such as PRR4, a gene associated with dry eye syndrome [N et al., 2016]. Despite this, recent research shows that the adult blood DNA methylome is largely affected by sex, but that these methylome sex differences do not change throughout adulthood and so are largely independent from age effects [Bergstedt et al., 2022].

The Illumina EPIC array has an increased coverage of the genome, including distal regulatory elements [Pidsley et al., 2016]. It was interesting that the 396 saDMPs were still found to be significantly enriched at CpG islands and CpG shores but depleted in open sea regions of the genome (see Figure 4.4). The genomic location of DNA methylation normally alters its function. Methylation in CpG islands normally functions to serve long term silencing of genes [Jones, 2012] and CpG island shore methylation is strongly related to gene expression [Irizarry et al., 2009], suggesting a potential functional role for these saDMPs. To further support these findings, I identified enrichment of these saDMPs at enhancers, 5'UTRs and promoters (see Figure 4.4). Enrichment at 5'UTRs is potentially suggestive that they may be acting as alternative promoters, though I did not test this hypothesis in this study. Despite this enrichment at regulatory regions, I found no correlation of these sites alone with significant differences in gene expression between males and females, suggesting that these saDMPs are not sufficient alone to predict gene expression. Further, this suggests that DNA methylation may potentially be acting as a passive reporter of sex specific transcription. Moreover, it is well established that DNA methylation differences do not always result in differences in gene expression but that these DNA methylation differences are likely to instead be part of larger gene regulatory networks, via acting distally or interacting with transcription factors [J et al., 2000],[Vaissière et al., 2008],[Ehrlich and Lacey, 2013],[Rauluseviciute et al., 2020],[Sadler et al., 2021]. Despite this, I acknowledge that one caveat of our study was that our DNA methylation data and gene expression data were obtained from different cohorts and have large differences in sample size. Expanding on this, the RNA seq data set specifically had had a small sample size, which I acknowledge is also a limitation of this study.

However, this potential link was identified in the TF motif analysis, where I found SRY (sex determining region Y) transcription factor motif, also known as the sex determining factor, to be enriched at female biased CpGs and further identified this gene to be acting as a hub in the TF network. SRY has been found to bind and repress WNT activation of ovarian genes, and has been shown to bind the promoter regions of many targets of involved in differentiation of the testis [Harley et al., 2003b],[Li et al., 2014b],[Song et al., 2017]. Furthermore, I also found ESR1 transcription factor motif enriched in the saDMPs female biased CpGs, a gene known to code for the oestrogen receptor. Whilst it is difficult to conclude for definite, I do hypothesise that although hormone levels may have some role in directing these DNA methylation patterns. it is also likely that some of these marks are established during development and early in life. I suspect this to be the case due to the presence of overlap between our identified saDMPs and those identified in newborns and placenta tissue.

It has previously been reported that 3D genome organisation can impact sex biased gene expression through direct and indirect effects of cohesion and CTCF looping on enhancer interactions with sex biased genes [Matthews and Waxman, 2020]. Recently, it was shown that with rising oestrogen levels, the female brain exhibits sex hormone driven plasticity and that chromatin changes underlie this [Rocks et al., 2021]. Interestingly, by annotating our saDMPs to distal genes using chromatin loops, I was able to identify contacts between saDMPs and three genes HIST1H3A, HIST1H4A and HIST1H4B which are core components of nucleosome, thereby responsible for playing a role in chromatin organisation. Note that the Hi-C data and DNA methylation data were not from matched samples, but two different cohorts. However, these results suggest that although I found DNAm to not be predictive of sex differences in gene expression (Figure S5), these saDMPs may interact with other genes, transcription factors and other epigenetic modifications to direct chromatin organisation and regulatory networks.

Moreover, it is important to note that it is difficult to truly evaluate how sex may contribute to health and disease due to the distinguishable relationship between sex and gender. As highlighted by Gatev and colleagues [Gatev et al., 2021], studies thus far, have not examined this relationship with regards to epigenetic differences. However, our previous work has shown that sex aneuploidies result in discordance in 'epigenetic sex' and 'physical sex' highlighting the importance for future work to examine this more carefully. Further, it accentuates the need for biomedical research to discriminate between sex and gender more carefully in research studies.

Lastly, I acknowledge limited overlap with previous studies yet conclude that this is due to our extremely large sample size (discovery, n=1171 and validation, n=2471) and improved handling of sex bias introduced by normalising such data with the sex chromosomes. Both factors contribute to our ability to detect true positives and obtain a more robust catalogue of true sex associated autosomal CpGs.

Here, I generated a resource that hopefully will provide useful in future research studies aiming to discriminate true associations in epigenome wide association studies from spurious associations.

Chapter 5

Identifying interindividual variation and stability of DNA methylation

5.1 Introduction

The combination of genetic, epigenetic and environmental variation between individuals is responsible for the large diversity observed in human phenotypes. Large efforts have previously been made to characterise genetic variation in humans, with important advances such as the characterisation of millions of single nucleotide polymoprhisms (SNPs), yet detailed catalogues of epigenetic variation are still not complete.

The need for these efforts are becoming increasingly clear, as more research suggests that changes to DNA sequence and exposure to environmental factors are unable to account for some phenotypic differences that can be observed amongst a population. This has been highlighted by studies in genetically identical organisms, commonly involving mono-zygotic twins but also some inbred animal studies [Wong et al., 2005]. Monozygotic twins (genetically identical twins) in almost all cases, are strikingly identical also in appearance, however are often discordant for particular diseases or phenotypes. Often, this discordance has been attributed to differences in environmental exposures which can have widespread and significant effects on human health. However, it is becoming increasingly accepted that epigenetic mechanisms may explain these findings in twin and animal studies. Further, this idea is perpetuated by studies identifying epigenetic differences between monozygotic twins whom of which are discordant for particular diseases such as amyotrophic lateral sclerosis, psoriasis and neurofibromatosis, thus suggesting that epigenetic variation could in fact explain differences in phenotype. [Young et al., 2017, Gervin et al., 2012, Vogt et al., 2011].

To this effect, previous studies have aimed to characterise a catalogue of loci showing highly variable DNA methylation in a range of tissues including peripheral blood, cord blood, saliva, placenta and colon [Bock et al., 2008, Hachiya et al., 2017, Garg et al., 2018, Costello et al., 2021, Palumbo et al., 2018, Derakhshan et al., 2022]. This is a difficult task due to the dynamic nature of the epigenome. Whilst genetic variation is minimal within individuals (intraindividual), yet extensive between individuals (interindidvidual), DNA methylation variation is extensive both within and between individuals. This is because individuals within a population can vary for a wide variety of reasons (stress levels, age, sex and smoking status) all of which in turn, can cause variation in the epigenome. Additionally, these varying patterns may differ between different tissues and cell types, unlike genetic variation.

Despite this, the Human Epigenome project is considered the best resource for mapping the human epigenome and some of their efforts were in fact directed towards characterising interindividual variation. However these efforts were focused mainly on an approximately 4KB region of the genome called the Major Histocompatibility Complex (MHC) rather than genome wide. Nevertheless, they drew some interesting conclusions, such as that DNAm variability is often tissue dependent, as the loci they investigated did not show concordant variation across tissues. Further, the authors also noted that these highly variable amplicons were mostly intragenic [Rakyan et al., 2004], a finding which was also identified in a study identifying inter-individual DMRs in monocytes [Schröder et al., 2017]. Additional work in human germ cells also supports this idea and further detected a large degree of variation within promoter CpG islands and pericentromeric satellites [Flanagan et al., 2006]. On the other hand, Garg and colleagues identify variably methylated regions (VMRs), defined as clusters of CpGs with high inter-individual epigenetic variation. They also found that these regions were enriched at enhancers and 3'UTRs and imprinted loci [Garg et al., 2018]. Despite the imperfect consistency of these results, the enrichment of these variably methylated sites and regions in these genomic regions does seem to suggest that they may play a functional role in gene expression or biological function.

Additionally, it is now also well known that DNA methylation is strongly influenced not only by environmental exposures, but also by genetic variation. A clear example is provided by the evidence that a rare change at DNMT3L, leads to significant DNA hypomethylation in subtelomeric regions of the genome, reported by [El-Maarri et al., 2009]. Further, we can consider a case where a SNP could be present at a CpG site that may change a cytosine base to a thymine base for example, which would mean that the lack of cytosine would result in an absence of DNA methylation, so the variation would result in either no DNA methylation or a fully methylated position. Another case may be where the absence of DNA methylation may facilitate transcription factor binding, however the binding of a transcription factor may or may not influence gene expression, introducing some randomness to this type of epigenotype. One example of this type of epigenotype can be seen in the agouti locus. This locus is responsible for coat colour in mice and its abnormal expression leads to obesity. Some of the agouti alleles carry IAP retrotransposon insertions. DNA methylation variation levels at these IAP insertions have been reported to differ depending on the allele, but also to correlate with gene expression [Argeson et al., 1996, Michaud et al., 1994, Perry et al., 1994]. Further, some of these loci have previously been reported to be 'epialleles'. Epialleles are sites which are variably expressed in genetically identical individuals induced by epigenetic modifications [Dolinoy et al., 2007]. Epialleles are most studies in plants, but studies in humans suggest that they have been associated with the colocalisation of transposable elements and are susceptible to influence from environmental and lifestyle exposures [Kessler et al., 2018].

This idea has been highlighted further, again through twin studies which show that mono-zygotic twins display more epigenetic similarity than di-zygotic twins [Kaminsky et al., 2009, Van Baak et al., 2018]. However, most evidence showing a relationship between genetic variation and epigenetic variation arises from work characterising the aforementioned methylation quantitative trait loci. Methylation quantitative trait loci can be described as genetic variants which influence CpG methylation [Gao et al., 2017]. mQTLs have previously been characterised in brain, blood, lung and adipose tissue [Hannon et al., 2018, Hannon et al., 2016, Gibbs et al., 2010, Drong et al., 2013, Gaunt et al., 2016, Olsson et al., 2014]. Additionally, these methylation quantitative trait loci (mQTLs) may also overlap variants which associate with gene expression levels. Thus, research linking these epigenetic signatures to genotypes is vital to provide more mechanistic insights into the interplay between genetics, disease and epigenetic variation.

Although, it should be noted that the relationship between DNA methylation and gene expression is a complicated one. Traditionally, DNAm has been considered as a mechanism of gene silencing or transcriptional repression, however the community is now beginning to appreciate the non linear relationship between DNAm and gene expression [de Mendoza et al., 2022] . To expand, the genomic context in which we find DNAm and its interplay with other epigenetic marks such as histone modifications and transcription factors also influence its impact on gene expression. Therefore, it is necessary to expand this line of research to consider not only where these highly variable and highly stable sites are located, but also the gene regulatory pathways they may be involved in, and what, if any, impact they have on gene expression.

Moreover, despite being a counter-intuitive hypothesis, more evidence is emerging suggesting that DNA methylation signatures may be stable across generations [Xavier et al., 2019, Zhang and Sirard, 2021]. Despite this, very little is known about inter-individual variation and stability of DNA methylation. Whilst answering this question was beyond the scope of this thesis, I am confident that providing the community with a robust catalogue of stably and variably methylated regions in whole blood will be useful in future studies exploring this line of research.

With this aim, I identify loci within the human genome with either high inter-individual variability or high inter-individual stability in DNA methylation in whole blood. I leveraged two large datasets from Understanding Society [Bao et al., 2022] investigating 850,000 CpG sites among a total of 3700 individuals (discovery dataset, n =1175 and validation dataset, n=2570). To our knowledge, this is the largest study using the Illumina EPIC BeadChip (allowing for interrogation of850,000 sites across the genome) to investigate variability and stability in DNA methylation at CpG sites in whole blood.

5.2 Results

5.2.1 Variably methylated probes and stably methylated probes are widespread across the genome and enriched in regulatory regions

I aimed to characterise inter-individual variation of DNA methylation in whole blood (DNAm) using a discovery and validation approach using two human cohorts (n=1171 and n=2471, respectively). This analysis was performed on the Illumina EPIC array which covers 850,000 sites across the human genome. Sites which were known SNP probes, cross hybridizing probes or sex chromosome linked probes were removed from our analysis. Therefore, after quality checks and data processing, a total of 747,302 CpGs were included in our analysis (see Material and Methods). I performed the analysis on discovery and validation datasets and of those CpGs, I identify 34,972 CpGs to be loci which show robust stable methylation in both our discovery and validation data sets (termed stably methylated probes (SMPs)) and 40,288 CpGs to have highly variable methylation levels in both our discovery and validation data sets across our population (which I termed variably methylated probes (VMPs)) (see Figure 5.1). Links to the full list of these sites is available in the appendix (see A). Interestingly, SMPs showed either low or high methylation, whereas VMPs showed predominantly intermediate methylation levels (see Figure 5.2A-B). Inter individual variation in DNA methylation is likely to be due to environmental differences, I therefore checked the overlap of VMPs and SMPs with known age, smoking and autosomal sex associated CpGs which were obtained from running EWAS for these phenotypes in the same cohort as explained in chapter 2. I did this as these are phenotypes are known to have strong associated epigenetic signatures. I found that although VMPs had a higher overlap with known sex, smoking or age associated CpGs, than SMPs, there are a still large portion of VMPs of which variation cannot be explained by these phenotypes alone (see Figure 5.2C-D). While I cannot rule out that other environmental differences may affect variation at these sites (such as air pollution exposure, diet or stress), testing all of the possible sources would be beyond the scope of this work.

To identify possible biological pathways that these probes were involved in, I performed KEGG ontology analyses which showed several enriched terms for these VMPs and SMPs, Table A.1 shows the top enriched terms for VMPs and SMPs. VMPs were enriched for terms such as Neuroactive ligand receptor interaction, olfactory transduction, cushing syndrome and morphine addiction. Enriched terms for the SMPs revealed several signalling pathways, such as metabolic pathways, MAPK, HIF-1 and FoxO signalling pathways along with other terms such as oxidative phosphorylation and lysine degradation.

VMPs were found in 9745 unique genes with 4780 of these genes harbouring several VMPs. The highest number of VMPs harboured by an individual gene was 87, located in PTPRN2 gene, a gene known to be encode an islet autoantigen in type 1 diabetes [Lee et al., 2019]. On the other hand, the SMPs were found in 11,597 unique genes with the majority of these harbouring more than one SMP. The highest number of SMPs harboured by an individual gene was 36, located in

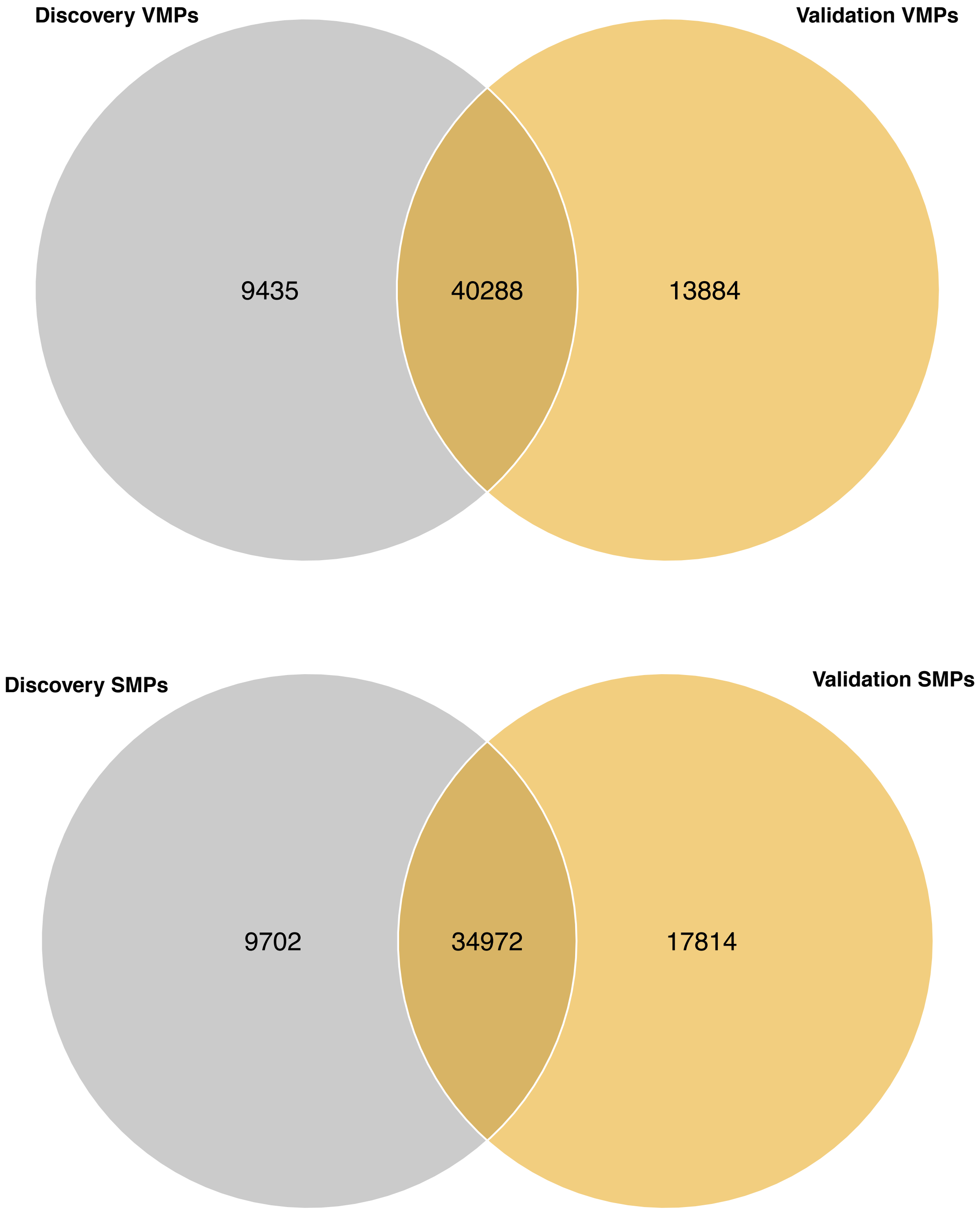


Figure 5.1: Venn diagram showing the number of VMPs and SMPs identified in discovery and validation data sets

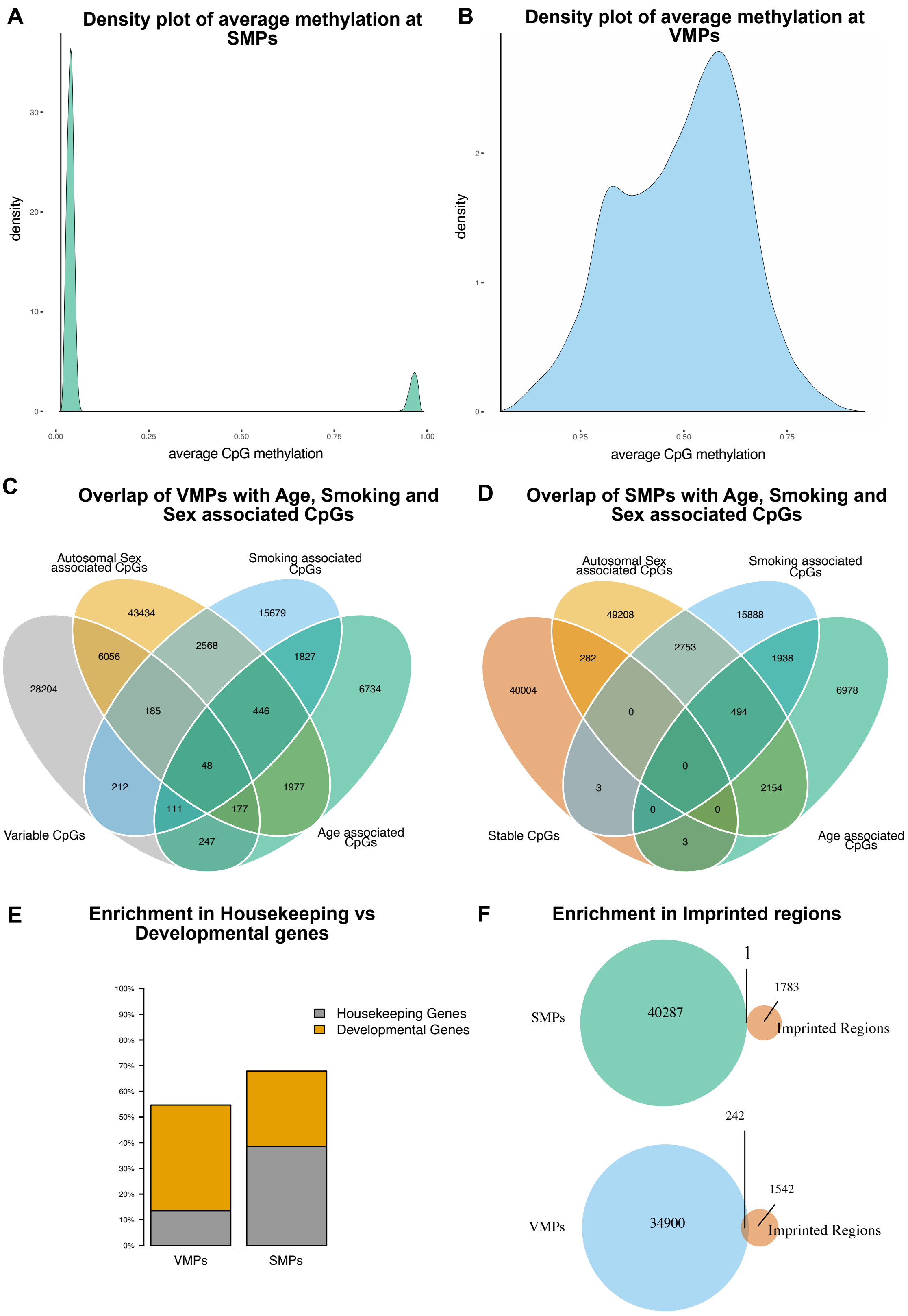


Figure 5.2: Density plots showing distribution of average CpG methylation at (A) SMPs and (B) VMPs. Venn diagrams showing overlap of (C) VMPs and (D) SMPs with known autosomal sex, smoking and age associated CpGs. (E) Barplot showing enrichment of VMPs and SMPs in housekeeping vs developmental genes. (F) Venn diagrams showing enrichment of VMPs and SMPs in imprinted regions.

a gene called MAD1L1 which is a checkpoint gene where dysfunction has been associated with chromosome instability [Tsukasaki et al., 2001].

Both VMPs and SMPs are distributed evenly across the genome but concentrated in CG dense regions (see Figure 5.3B). To help us gain more insight into the functional role of these variably and stably methylated probes, I aimed to characterise their genomic location with respect to CpG islands and functional regions. I observed that SMPs were highly enriched in CpG islands and depleted at CpG shelves and open sea regions of the genome. In contrast, I found that VMPs were slightly depleted at CpG islands and slightly enriched at CpG shores, however were also depleted at CpG shelves (see Figure 5.3C). Furthermore, SMPs were also highly enriched at 5' UTRs and promoter regions of the genome, and showed slight enrichment at enhancer and exon regions. On the other hand, VMPs were slightly enriched at intergenic regions but depleted at UTRs, promoters and transposable elements (see Figure 5.3D).

Therefore, I hypothesised that SMPs may be enriched at housekeeping genes. I find that SMPs are in fact highly significantly enriched in housekeeping genes compared to developmental genes and that VMPs were significantly depleted in housekeeping genes (see Figure 5.2E). I also investigated whether variable probes were enriched at imprinted regions of the genome and find that variable probes are significantly enriched here with 242 of our VMPs overlapping an imprinted gene, higher than I would expect by chance compared to the EPIC array background. On the other hand, stably methylated probes are depleted at imprinted regions (Permutation test, pvalue < 0.05) (see Figure 5.2F).

5.2.2 VMPs and SMPs located at promoters and enhancers are enriched for transcription factor motifs

Next, as DNA methylation may influence the affinity of transcription factor binding at regulatory regions I performed transcription factor (TF) binding site and gene ontology analyses for our VMPs and SMPs located at promoters and enhancers

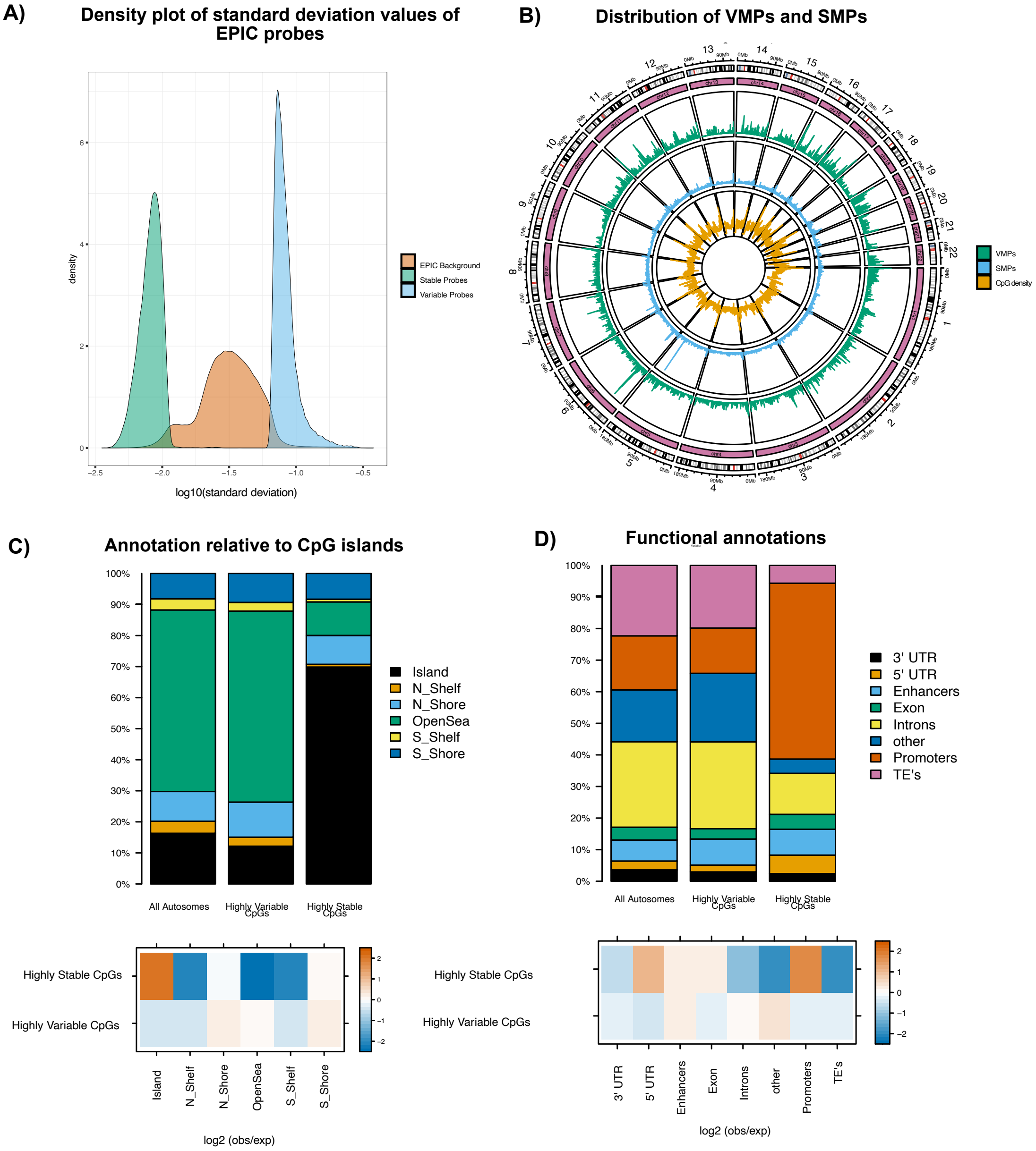


Figure 5.3: Identification and location of variably methylated probes (VMPs) and stably methylated probes (SMPs). (A) Density plot representing the standard deviation values of methylation across our samples for VMPs,SMPs and the autosomal EPIC background. (B) Circos plot representing distribution of VMPs and SMPs across the human genome. The outermost ring displays chromosome numbers and bands. The second ring represents CpG island density. The third and fourth ring displaying green and blue lines represents the distribution of SMPs and VMPs respectfully across the genome (50,000kb bins). (C) Top panel shows the annotation VMPs (n= 34972) and SMPs (n=40288) relative to CpG islands compared to the autosomal background. Bottom panel shows the log2 (obs/exp) annotations based on the autosomal EPIC background of the different annotations D) Top panel shows the annotation VMPs (n= 34972) and SMPs (n=40288) to genomic features compared to the autosomal background. Bottom panel shows the log2 (obs/exp) annotations based on the autosomal EPIC background of the different annotations. (E) Barplot representing the enrichment of VMPs and SMPs in housekeeping and developmental genes. (F) Venn diagrams representing enrichment of SMPs and VMPs in imprinted regions.

specifically. First, I evaluated whether the VMPs and SMPs located at promoters were enriched in motifs for TFs (50 bp window). For the VMPs I found 408 unique enriched TFs (p.value < 0.01) (see Figure 5.4A) with strongest evidence for TFAP2A and NHLH1. For the 103 SMPs I identified 86 enriched TFs, including SREBF1 and AHR (Figure 2A). The same analyses was repeated for those VMPs and SMPs located at enhancers, which revealed uniquely enriched TFs (see Figure 5.5A), specifically I identified 376 uniquely enriched TF motifs at VMPs located in enhancers and 74 at SMPs located at enhancers. The most strongly enriched TFs at enhancers differed to those at promoters, with ATF2 and FOS at VMPs and NNT and ODC1 at promoters (see Figure 5.4A).

To investigate whether the transcription factor motifs were enriched for terms related to biological processes or pathways, I performed pathway analyses using the GO and KEGG databases. I identified several enriched KEGG pathways for the TFBS enriched at VMPs at promoters, spanning a wide range of processes such as transcriptional misregulation in cancer, FoxO signalling pathways and signalling pathways regulating pluripotency of stem cells (see Figure 5.4B). I also identified several enriched terms for the TF motifs enriched at SMPs at promoters such as viral carcinogenesis, prion disease and Parkinson disease (see Figure 5.4B). Interestingly, for those TF motifs enriched at VMPs in enhancers, I find several replicated pathways, such as transcriptional misregulation in cancer (see Figure 5.5B).

I then used a cytoscape plug in tool called cytohubba to robustly characterise hub genes in the transcription factor network. For the TF motifs enriched at VMPs at promoters I identified MAPK1, GATA3 and SOX2 and other genes to be hub TFs in the network (see Figure 5.4C). Moreover, for those TFs enriched at SMPs at promoters I identified CYCS, P4HB and HSPA5 to be hub genes in this network (see Figure 5.4D). Similar to our enrichment analyses, I also found a large overlap between the hub genes identified at promoters and enhancers for both VMPs and SMPs. This suggests that enrichment of transcription factors

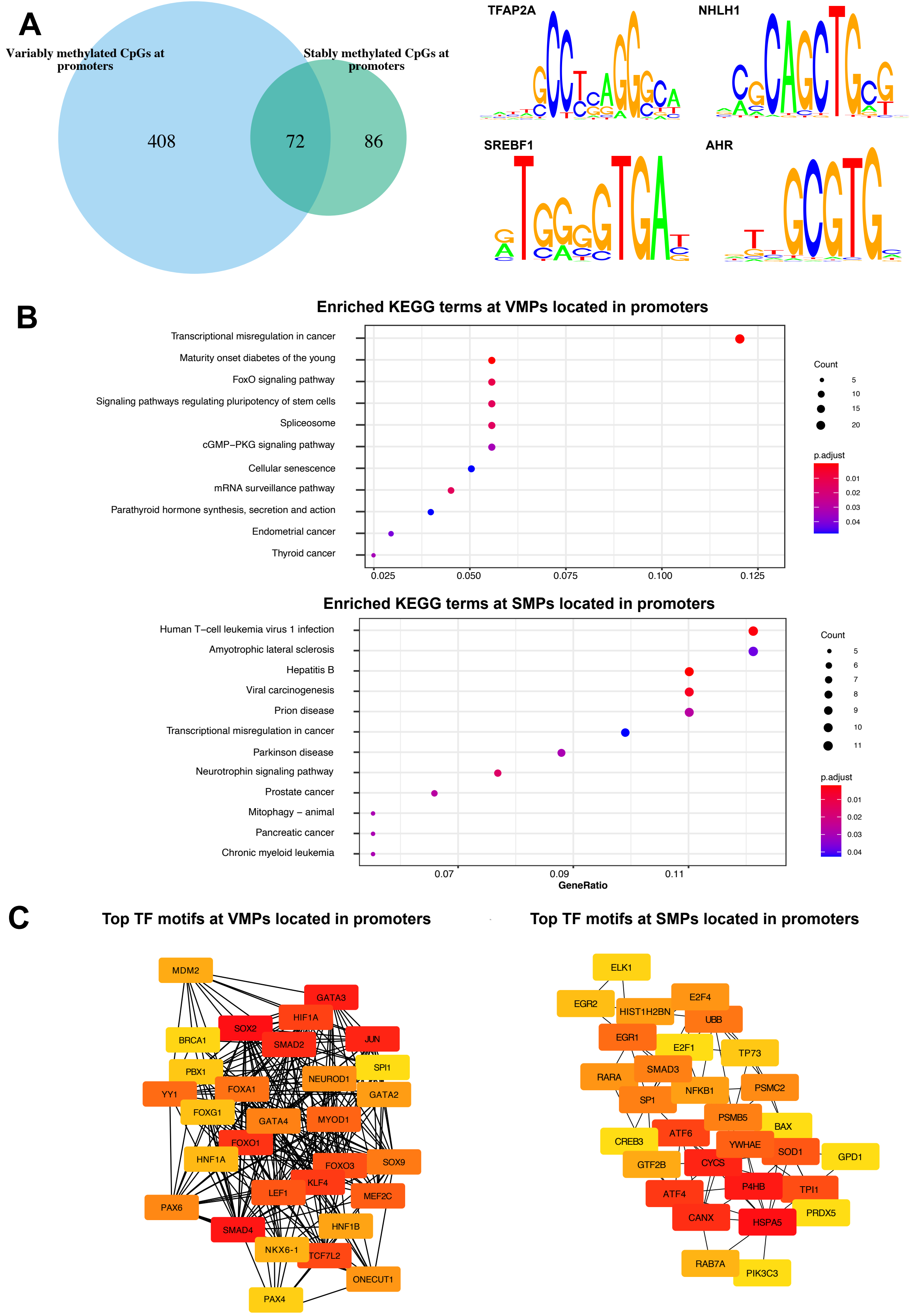


Figure 5.4: Transcription factor motif enrichment for VMPs and SMPs at promoters. (A) Overlap of enriched TF motifs for VMPs (blue) and SMPs (green). The top two motifs enriched in VMPs were TFAP2A and NHLH1 and in SMPs were SREBF1 and AHR. (B) KEGG analyses for the significantly enriched TF motifs at VMPs and SMPs. (C-D) Sub networks of the top 30 enriched TF motifs at (C) VMPs and (D) SMPs. Node colour represents the degree of connectivity. The scale from red to yellow represents the top 30 enriched TF motif rank from 1-30, with red indicating highest degree and yellow indicating lowest degree.

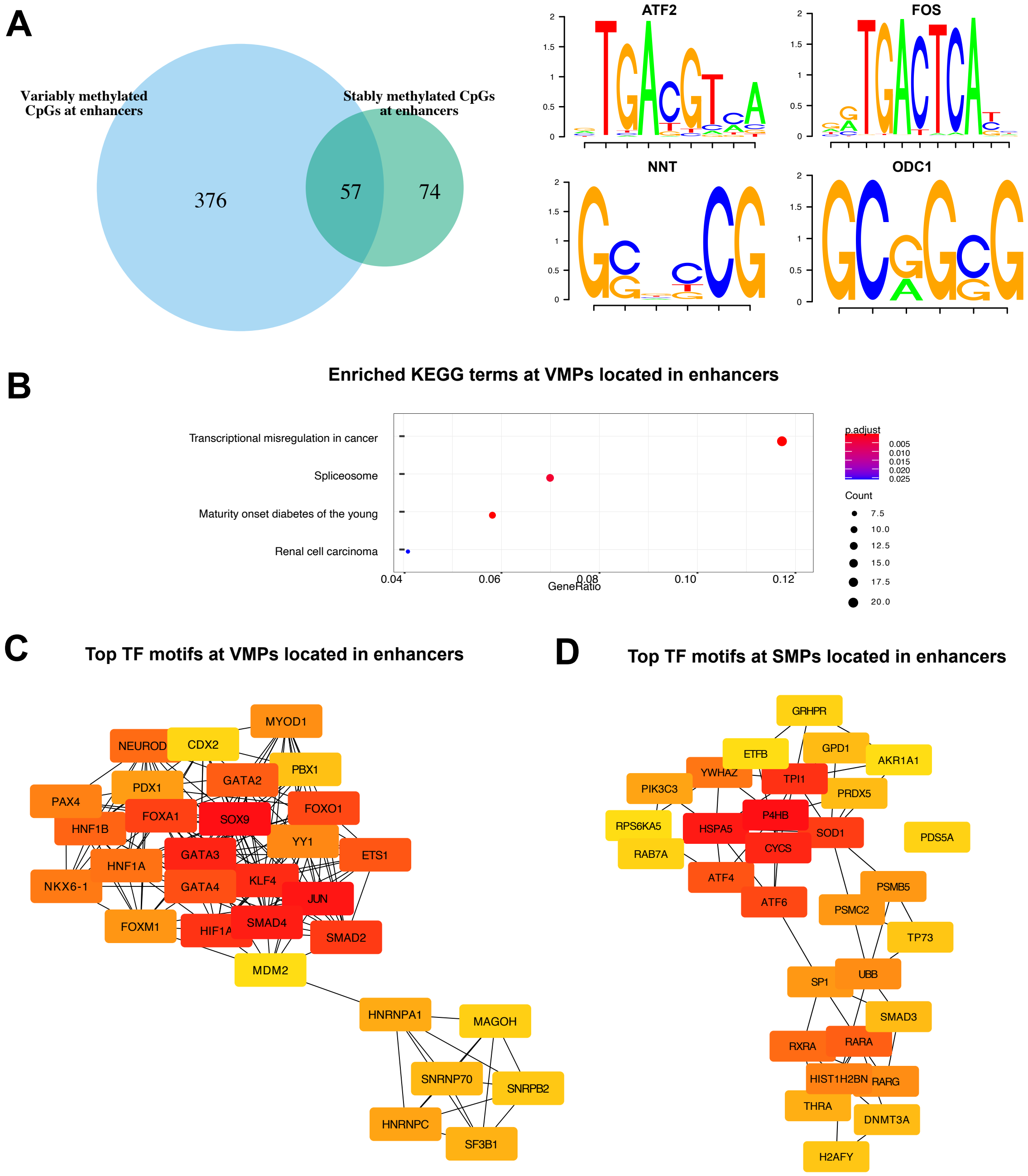
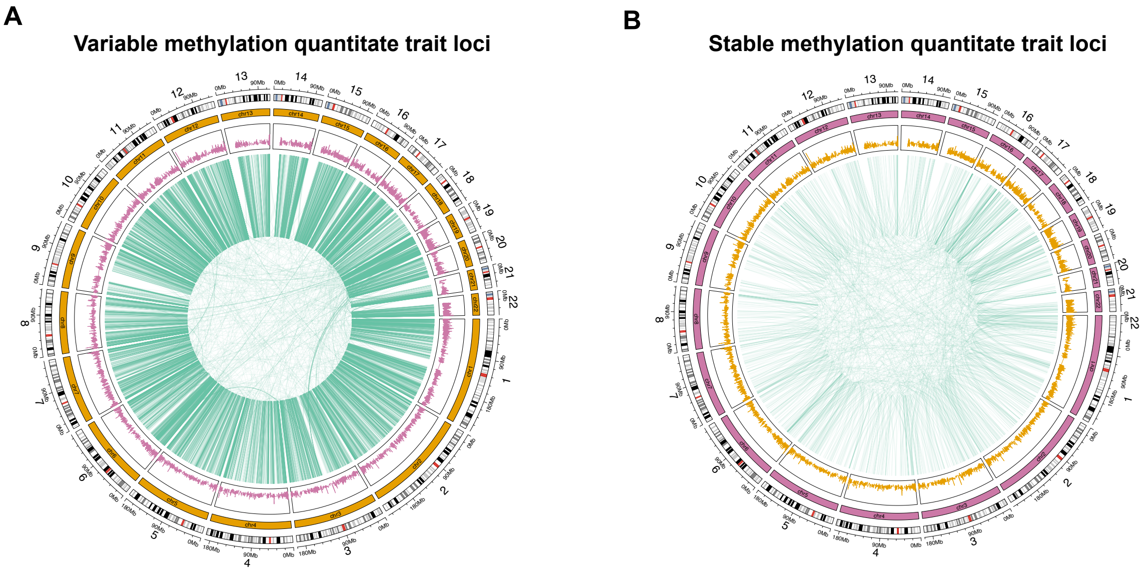


Figure 5.5: Transcription factor motif enrichment for VMPs and SMPs at promoters. (A) Overlap of enriched TF motifs for VMPs (blue) and SMPs (green). The top two motifs enriched in VMPs were ATF2 and FOS and in SMPs were NNT and ODC1. (B) KEGG analyses for the significantly enriched TF motifs at VMPs and SMPs at enhancers. (C-D) Sub networks of the top 30 enriched TF motifs at (C) VMPs at enhancers and (D) SMPs at enhancers. Node colour represents the degree of connectivity. The scale from red to yellow represents the top 30 enriched TF motif rank from 1-30, with red indicating highest degree and yellow indicating lowest degree.

differs slightly according to functional regions in the genome, but not significantly and that the hub transcription factors enriched near highly variable and highly stable methylation sites are robust.

5.2.3 VMPs are under higher genetic control than SMPs

To gain insight into the extent to which the variability or stability at the VMPs and SMPs could be explained by genetic causes, I obtained whole blood mQTL data. Of the VMPs, 44.9% were associated with SNPs (see Figure 5.6A). I further categorised these into cis and trans mQTLs where a cis mQTL was defined when the SNP and CpG were less than or equal to 500bp away from one another, a trans mQTL is defined when the SNP and CpG were more than 500bp away from one another. Of the VMP mQTLs, 21% were cis, 79% were trans mQTLs. One example is a mQTL pair located on chromosome 1 at the probe cg04315214 which is associated with two independent SNPs at the gene PRKCZ (FDR < 0.01). For SMPs, 3.27% of the SMPs were associated with SNPs (see Figure 5.6B). Of these, 8.1% were cis, 91.9% were trans. One example is a SNP-CpG pair on chromosome 1 between cg04093404 and 4 SNPs at the gene TAS1R1. These findings are in line with previously reported literature that mQTLs are more likely to affect more variable CpGs [Villicaña and Bell, 2021, Díez-Villanueva et al., 2021].



Variable methylation quantitate trait loci

Stable methylation quantitate trait loci

Figure 5.6: Methylation quantitative trait loci connect VMPs and SMPs with SNPs. (A) Circos plot illustrating cis and trans mQTLs in whole blood. The outermost ring displays chromosome numbers and bands. The second ring represents CpG island density. The third ring represents associations between VMPs and SNPs (green lines). (B) Circos plot illustrating cis and trans mQTLs in whole blood. The outermost ring displays chromosome numbers and bands. The second ring represents CpG island density. The third ring represents associations between SMPs and SNPs (green lines).

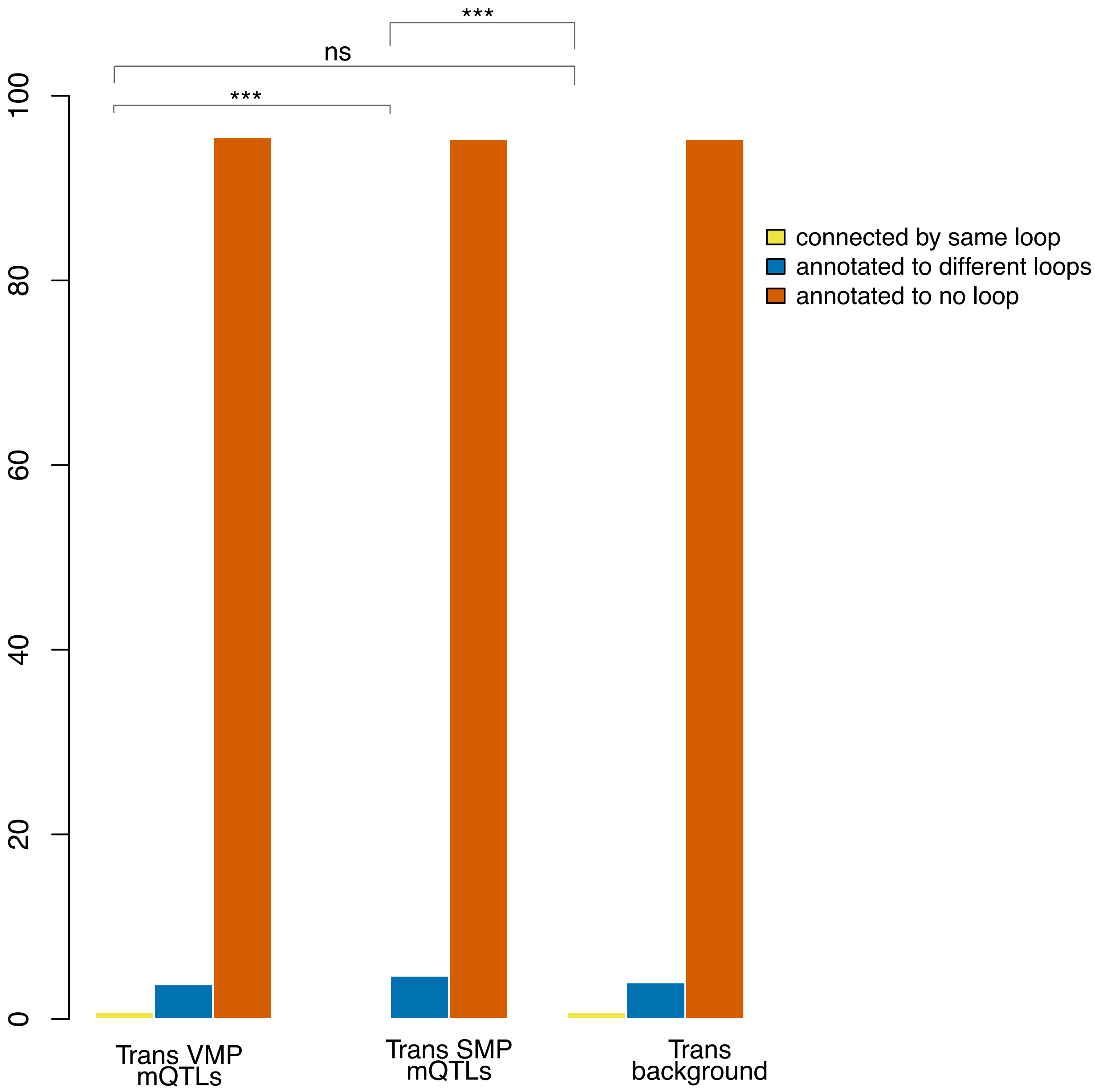


Figure 5.7: Barplot representing the percentage of trans mQTL pairs connected by chromatin loops. Fishers exact test was performed to determine statistical significance against the Illumina EPIC array background.

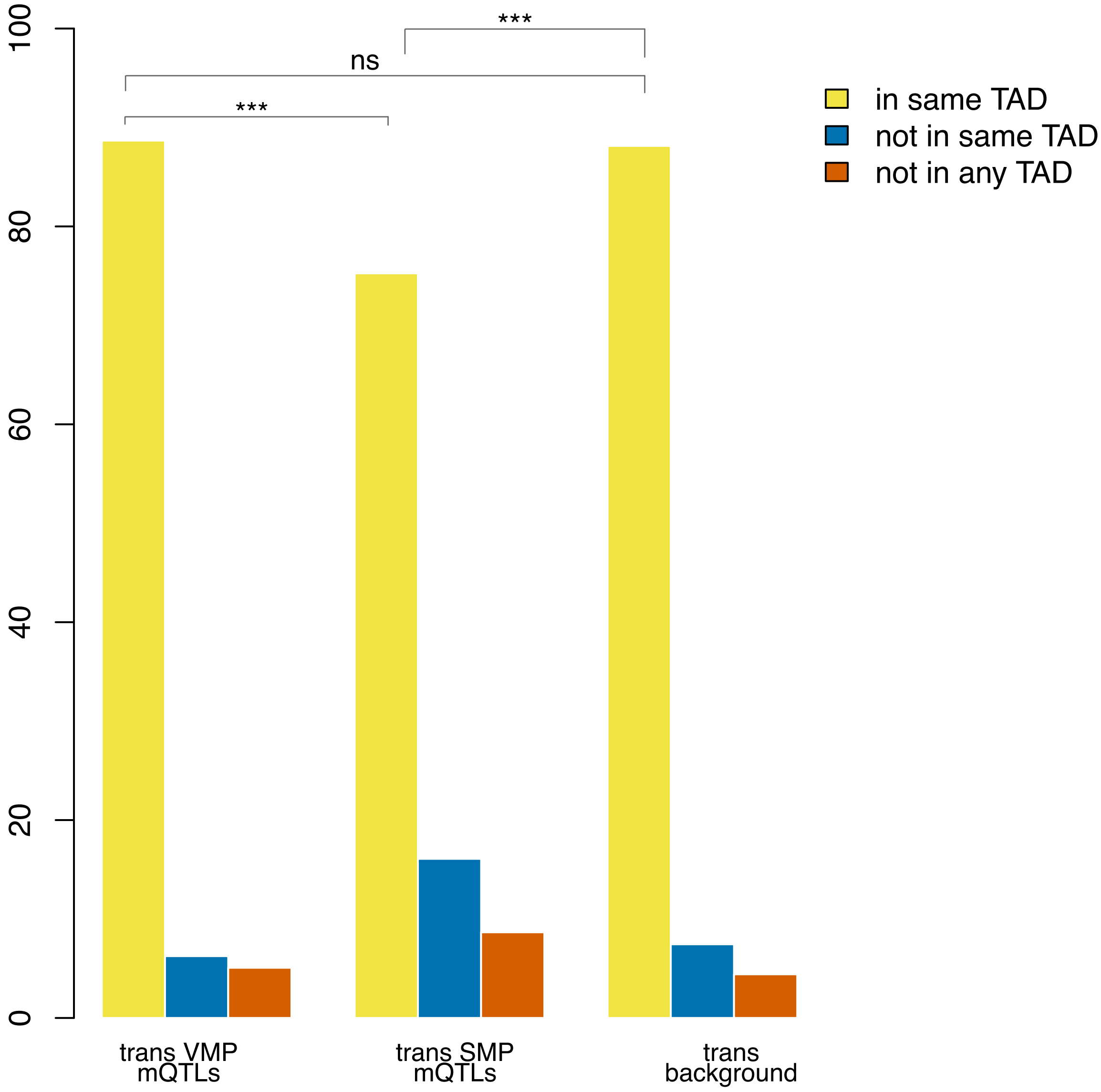


Figure 5.8: Barplot representing the percentage of trans mQTL pairs in the same topological associated domain. Fishers exact test was performed to determine statistical significance against the Illumina EPIC array background

.

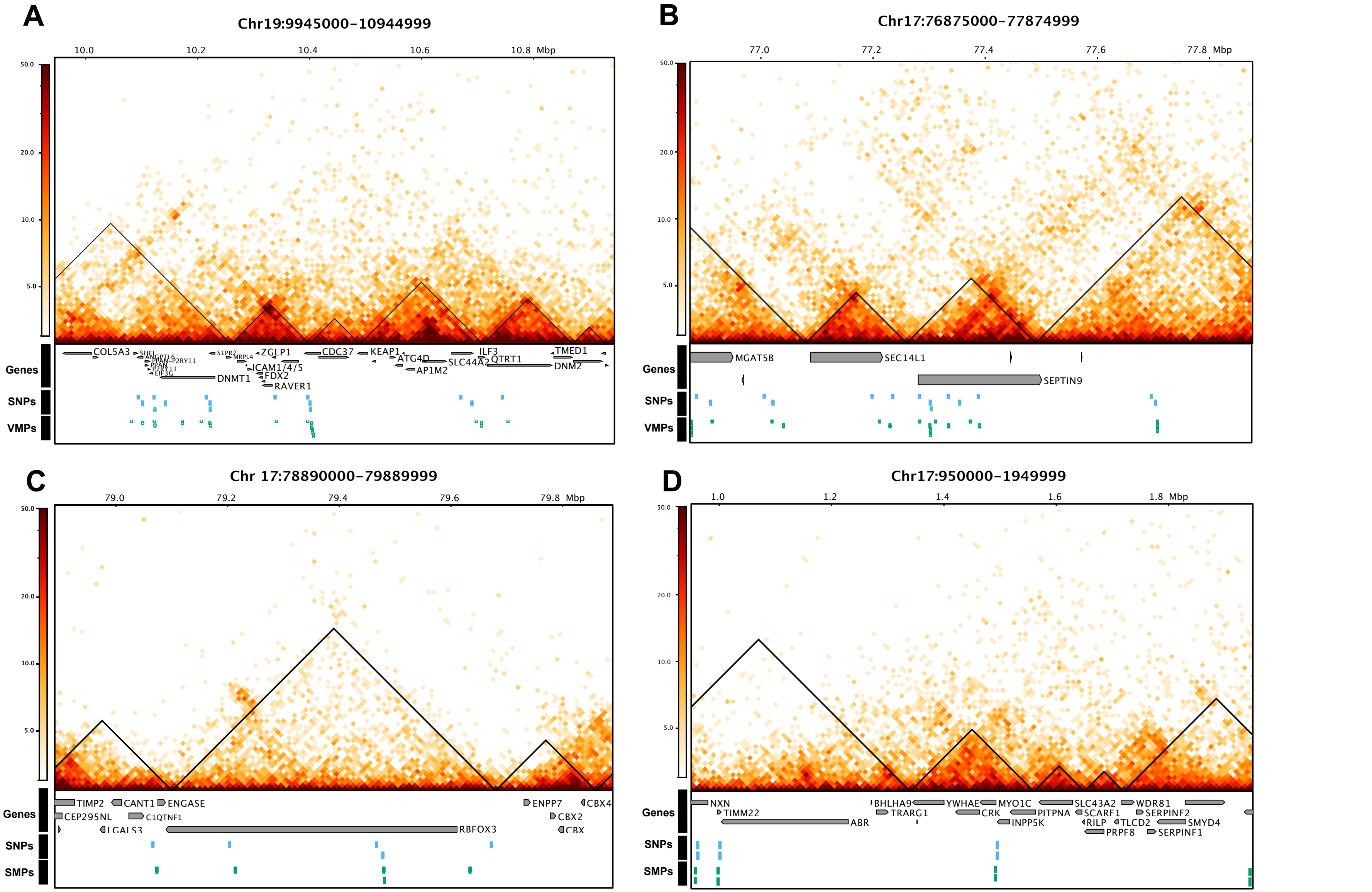


Figure 5.9: mQTL pairs occupy the same topologically associated domains. (A) and (B) illustrate cases where VMPs and their associated SNPs occupy the same TADs (C) and (D) illustrate cases where SMPs and their associated SNPs occupy the same TADs.

I also tested whether our SNP-CpG pairs were enriched in regulatory regions such as promoters or enhancers, however I found no significant enrichment in any particular regions that I could expect by chance. I did however, hypothesise that the trans SNP-CpG pairs may occur within topologically associated domains (TADs) and at chromatin loops as genetic and epigenetic interactions may lead to interindividual differences in chromatin organisation.

I found that the trans SMP-SNP pairs were annotated to a significantly lower percentage of loops compared to the VMP-SNP pairs and the EPIC background by chance (see Figure 5.7). Chromatin loops call very strong interactions, but TADs identify regions that interact more inside than outside thus allowing for me to capture finer connections between SNPs and VMPs. I hypothesised that I would also see a depletion of trans SMP-SNP pairs occupying the same TAD. I found that a lower percentage of SMP-SNP pairs occupied the same TAD than VMP-SNP pairs and than I would expect by chance alone (fishers exact test $<$ 0.05) (see Figure 5.8. Although I found that a larger amount of the VMP-SNP pairs occupied the same TAD (Figure 5.8), I did identify a subset of SMP-SNP pairs which also occupied the same TAD (see Figure 5.9). Several of the mQTLs overlapped multiple genes, including genes such as DNMT1, SEPTIN9 and ILF3, I therefore performed enrichment analyses on the genes annotated to our SNP pairs to try and gain some insight into biological function and pathways. Enrichment analyses of VMP-mQTL associated SNPs revealed several GO terms such as cell morphogenesis involved in neuron differentiation, small GTPase mediated signal transduction and cation transmembrane transporter activity (see Figure 5.10A). These SNPs were also enriched for few KEGG terms including Focal adhesion, axon guidance and cell ahesion molecules (see Figure 5.10C). The SMP-mQTL associated SNPs were enriched for fewer GO terms, but included pathways such as nuclear chromosome part, response to peptide and chromatin (see Figure 5.10B). However, were only enriched for one KEGG term, AMPK signalling pathway (see Figure 5.10D).

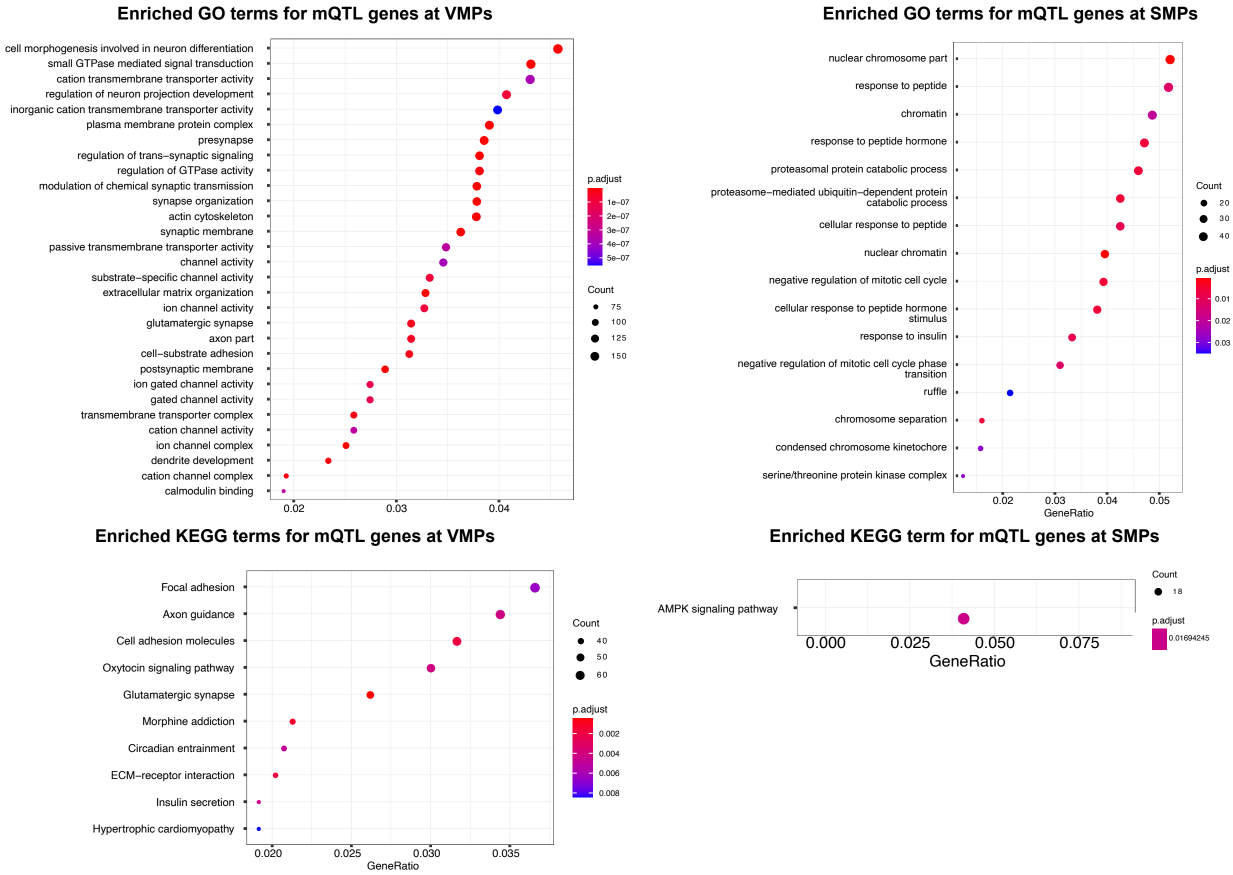


Figure 5.10: Enriched GO and KEGG terms for mQTLs. (A) GO and (C) KEGG analyses for the mQTL genes at VMPs. (B) GO and (D) KEGG analyses for the mQTL genes at SMPs

5.2.4 Identifying putative epialleles in human whole blood

Epialleles are sites which are variably expressed in genetically identical individuals induced by epigenetic modifications [Dolinoy et al., 2007]. I screened for epiallele like sites in human whole blood by applying a test for unimodality (using the hartigans dip test) in CpG sites which showed a variable intermediate methylation value. A total of 784 CpG sites met our criteria for epiallele like sites (see Figure 5.9A) and 405 of these CpG sites associated with a gene. 52 of these genes contained more than one epiallele like sites, with one gene PM20D1 containing 8 CpGs identified by our analysis as having an epiallele like status. PM20D1 is a gene which has previously been reported to be associated with obesity, insulin resistance and the progression of alzheimers disease [Wang et al., 2020b, Yang et al., 2022]. Moreover, it has previously been reported to contain a variably methylated region associated also with BMI [Feinberg et al., 2010]. Additionally, several genes involved in the major histocompatibility complex such as HLA-DRB1, HLA-DQA1, HLA-C and HLA-DRB6 also contained several CpGs identified as epialleles in our analysis. Of these CpGs,42% are found on the Illumina 450K array and 58% were unique to the Illumina EPIC array (see Figure 5.9B). Furthermore, I hypothesised that some of these epiallele sites may be controlled by genetic variation, despite my efforts made to clear up direct SNPs from our analysis. I therefore characterised how many of these epiallele like sites were linked to a SNP through our aforementioned mQTL analysis. I found that roughly 19% of these were in fact not linked to any genetic variants, but 81% were. Of these, the majority were trans mQTLs (65%) and the others were cis mQTLs (16%) (see Figure 5.9C) suggesting that these are unlikely to be an artefact of mutations. As I identified that some of these epiallele sites were potentially mediated by genetic variants, I hypothesised that if this was the case, I would observe a large overlap between different tissues. Therefore, I calculated these epiallele like sites in two other human tissues from different cohorts; skeletal muscle and brain. I identified limited overlap between all 3 tissues, suggesting that the majority of these epiallele like sites I identified are tissue specific. Further to this, as epialleles have been reported to be enriched in variable regions and also imprinted regions, I also checked the overlap of these epialles with previous studies. First, I overlapped our epiallele like sites with VMPs previously reported in whole blood in monozygotic twin pairs, and identify that 10% of our epiallele like sites were identified by this study. This suggests that a portion of these epialleles may have been established early in development, as expected. Secondly, I also checked the overlap of our epiallele like sites with hyper variable CpGs identified in multiple tissues by [Derakhshan et al., 2022] and found that 18% of our epiallele like sites were also identified in this study [Derakhshan et al., 2022]. This suggests that some of these epialleles, although not identified to be associated with a SNP by our stringent analysis, may be directed by underlying genetic influences. Lastly, I also checked the functional annotations of these epialleles, to try and identify a regulatory role for these sites. Interestingly, we find the epilleles to be enriched at enhancer and intergenic regions, but depleted at promoters, exonic and 3'UTR regions.

Functional annotations of epiallele like sites in human whole blood

As it has previously been reported that epigenetic variation at epialleles may result in gene expression variation, I next annotated epiallele like sites to their nearest target gene in order to investigate this further. I calculated the relationship between methylation variation and target gene expression variation in microarray data using the pearsons correlation method. In doing so, I was able to identify a positive correlation between methylation variation at epiallele like sites in 5' UTRs and their target gene expression variation (pval < 0.05). I note that is is interesting that I was able to find a statistically significant link despite not using matched datasets.

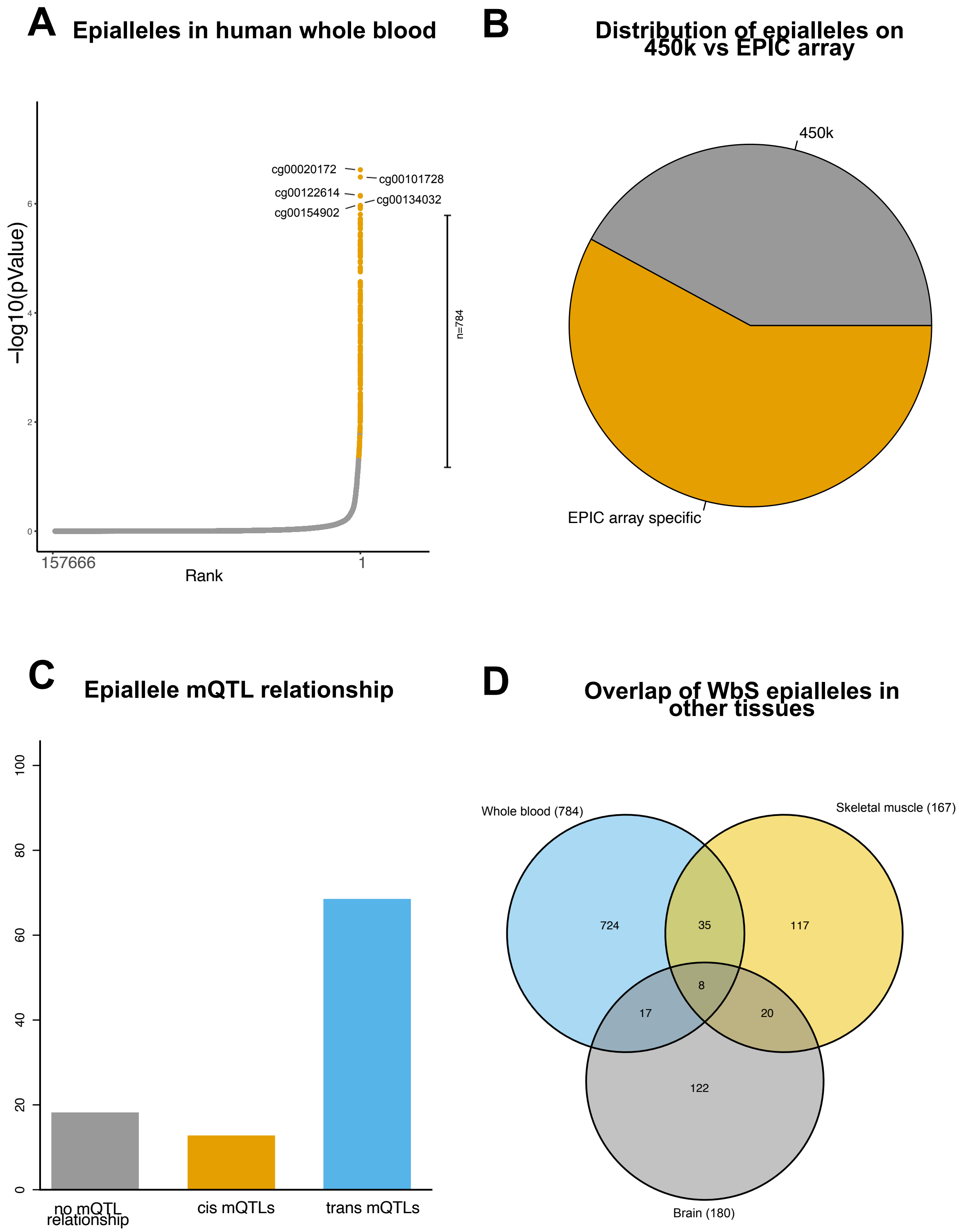


Figure 5.11: Epiallele like sites identified in whole human blood. (A) Rank plot showing proportion of CpG sites on the EPIC array which passed our threshold for what I considered to be an epiallele like site as determined by hartigans dip test. A total of 784 CpGs met this threshold, and the top 5 are annotated. (B) Pie chart showing distribution of epialleles on the Illumina 450k and Illumina EPIC array.(C) Barplot showing percentage of mQTL relationship of epiallele like sites. Y axis represents percentage. (D) Venn diagram indicating overlap of epiallele like sites identified in whole blood, brain and skeletal muscle tissue in humans.

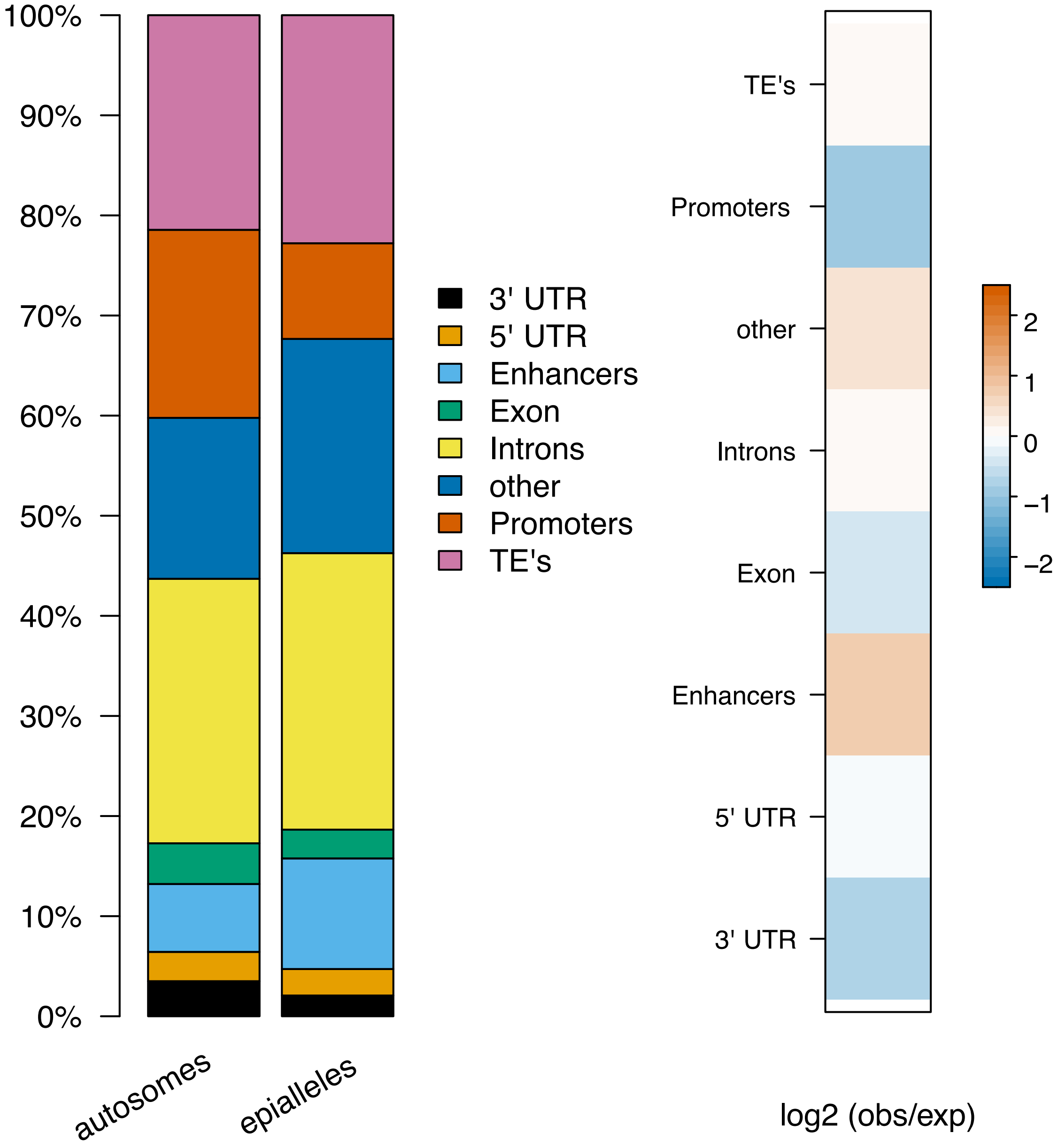


Figure 5.12: Functional annotations of epiallele like sites in human whole blood. Left panel shows the overlap of all epiallele like sites (n=784) with genomic features compared to the background. Right panel shows the log2 (obs/exp) based on the background of the different annotations.

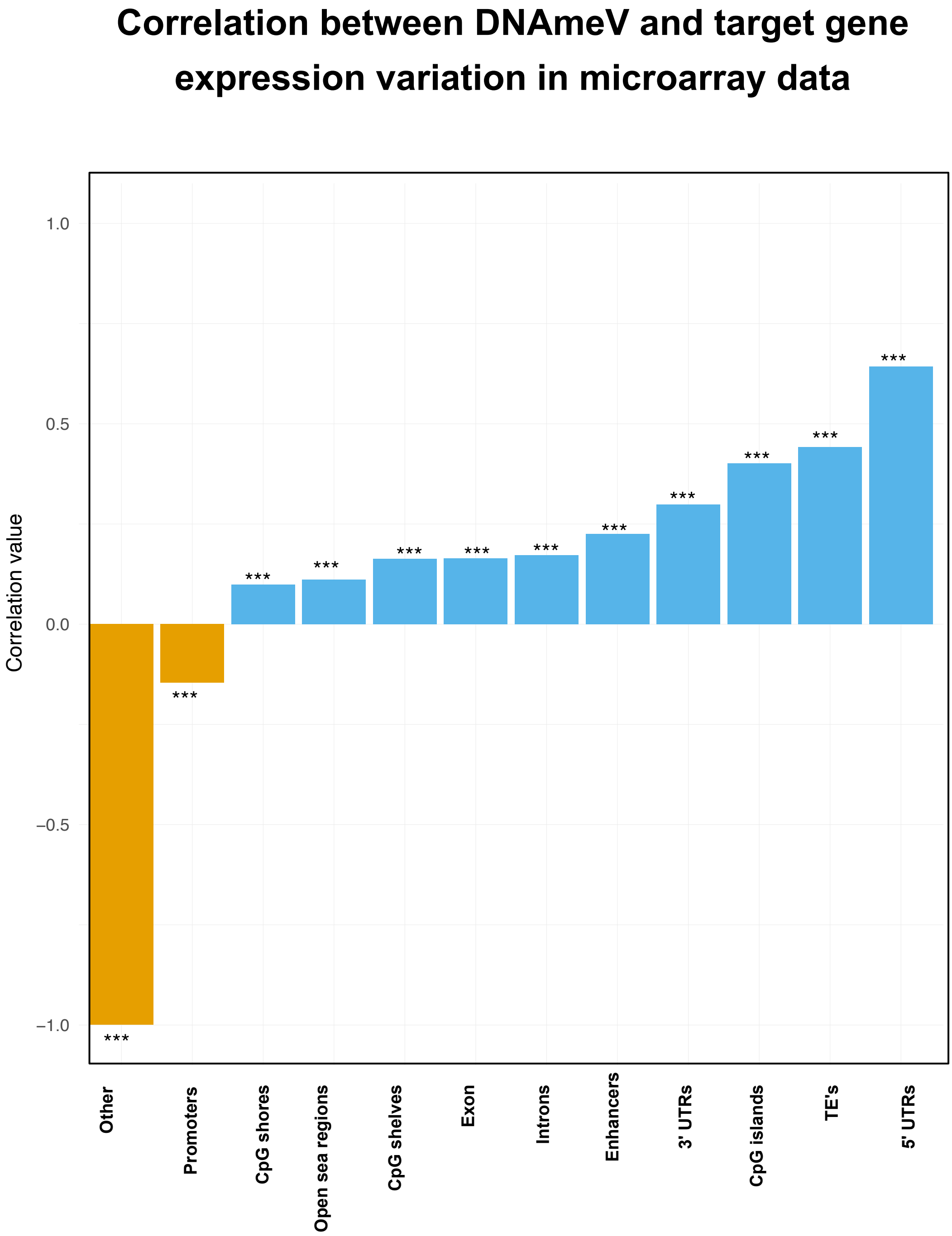


Figure 5.13: Barplot showing correlation between methylation variation and target gene expression variation in microarray data. Yellow bars indicate a negative correlation and blue bars indicate a positive correlation. The y-axis shows the correlation value with the corresponding p-value significance above each bar.

5.3 Summary and Discussion

Here, I analysed DNA methylation inter individual variability and stability amongst healthy individuals in Understanding Society data sets in order to better understand its role in diversity in human phenotypes and its relationship with gene expression variation and other genomic factors. I was able to establish the presence of epigenetic inter individual variation and stability across the human genome and showed that these loci were widespread across the genome. Previous work with a similar aim to identify regions or sites with high inter individual variability in DNA methylation has also reported some contradicting results with a detailed and robust catalogue of VMPs and additionally SMPs still needing to be categorised. When comparing the VMPs discovered in this study to findings previously reported in samples (where a full list of highly variable sites were available) I do observe some overlap. For example, I identify 43% overlap of our identified VMPs with previously reported VMPs in whole blood from 426 monozygotic twin pairs [Garg et al., 2018]. On the other hand, I observed an 85.28% overlap with hyper variable CpGs identified by [Derakhshan et al., 2022] in multiple tissues and ethnicity's. A perfect overlap between studies of any kind is rarely achieved and in this instance, it could be due to several factors. Firstly, some of the VMPs identified could be tissue specific, a finding I was able to reveal throughout this work, whereas others may be consistent across different tissues also confirmed throughout this work. Secondly, some VMPs identified in different studies may actually be cohort specific, as some of these VMPs or SMPs could be driven by rare alleles, which may be found only in specific cohorts. Therefore, it may be necessary for future work to consider cohort characteristics as well as the tissue of interest.

However, consistent with previous research [Lam et al., 2012], I demonstrated that these loci overlap regions known to be associated with biological age, biological sex and smoking status. Interestingly, I observed that all of the highly variable sites showed an intermediate methylation status, a finding previously identified by [Hachiya et al., 2017]. The role of intermediate methylation remains unclear, however, previous work suggests that it may be a conserved signature of gene regulation and exon usage [Elliott et al., 2015]. Regions of intermediate methylation were also shown to have similar intermediate levels of active chromatin marks and their target genes also having intermediate transcriptional activity. This suggests that these sites may have regulatory potential distinct from repressive or permissive states resulting from fully methylated or unmethylated sites. In contrast, the most stable sites show either high or low average DNA methylation values across our cohort, suggesting these loci might be under tight epigenetic control. In line with this, I identified an enrichment of SMPs at housekeeping genes. Housekeeping genes are thought to constitute a small set of genes required to maintain minimum basic cellular function, thus one could expect these genes to have consistent epigenetic regulation across individuals. On the other hand, I also identified enrichment of VMPs at developmental genes. These are genes which potentially could play a role in specialised functions within a cell or tissue and could potentially thereby serve as biomarkers. Therefore, I might expect high epigenetic variability at these genes, inline with the results I found. Expanding on this, it has also been reported that highly variable sites are found at imprinted regions [Derakhshan et al., 2022]. Through investigation of the genes annotated to our VMPs, I also found 21 VMPs identified in our analysis annotated to the gene HOXA5, a gene predicted to be a maternally imprinted gene. I therefore checked the enrichment of our identified VMPs and SMPs in imprinted regions, and found that the VMPs were significantly enriched in imprinted regions compared to SMPs (permutation test, pval < 0.05). This is also in line with previously reported similar results [Derakhshan et al., 2022, Garg et al., 2018, Zeng et al., 2019].

As the genomic context in which DNA methylation is found is important for its relationship with gene expression and biological function, I also sought to characterise the genomic location of our VMPs and SMPs. I found an enrichment of VMPs at CpG shores, intergenic regions and enhancers, with the latter having been previously reported by [Garg et al., 2018]. On the other hand, SMPs were enriched at 5'UTRs, promoters and exons. These results collectively suggest that these loci may play a role in gene regulation or biological processes, due to the functional role of these annotations.

However, the relationship between DNA methylation and gene expression is a complex one. For example, it is traditionally thought that DNA methylation at promoter regions is linked directly to transcriptional repression or gene silencing. Despite this, recent work which synthetically methylated 1000's of promoters in the human genome failed to identify a link between promoter DNA methylation and gene expression. Yet, they suggest that the context specific roles of DNA methylation are highly influenced by transcription factor binding affinities [de Mendoza et al., 2022]. Therefore, I investigate the regulatory networks by searching for the presence of transcription factors among VMPs and SMPs enriched at important regulatory regions such as promoters and enhancers. Interestingly, I identify the presence of transcription factors which have indeed been previously reported to be methylation sensitive. For example, I identify TFAP2A as the most signficantly enriched motif at VMPs at promoters. It has previously been reported that the presence of DNA methylation leads to an increased binding of TFAP2A to B1, leading to suppressed gene expression of NRBP1 gene [Zhu et al., 2017]. Additionally, I find SREBF1 to be the most significantly enriched TF motif at SMPs at promoters which has also been reported to be sensitive to CpG methylation [Krause et al., 2020]. Further, our hub transcription factor motif analysis for VMPs and SMPs located at enhancers and promoters also revealed more methylation sensitive TFs such as JUN, ATF4 BRCA1 and FOXA1 [Luo et al., 2021, Héberlé and Bardet, 2019, Zhu et al., 2017]. This indicates that interindividual differences in DNA methylation would result in differences in TF binding and consequently differences in gene regulation.

Expanding on this, I have previously shown that although DNA methylation may not always be predictive of gene expression [Grant et al., 2022], it is possible that DNA methylation may be directed by or direct inter individual differences in chromatin organisation via chromatin looping or toplogically associated domains. Furthermore, variance or stability of DNA methylation may also be directed by genetic differences, as previously reported [Hannon et al., 2018]. In line with this, I demonstrated that VMPs seem to be under higher genetic control than SMPs, suggesting that genetic differences may in part drive epigenetic inter individual variability. As previously described in Section 5.1, there are several ways that genetic differences may result in epigenetic differences. The most simple cases are observed when the mQTL pairs are in cis, one example may be where a cytosine base is changed to a thymine base, meaning the lack of cytosine will result in the absence of DNA methylation. This could result in increased or decreased binding of specific transcription factors, leading to inter individual differences in gene expression levels. However, the mechanism by which mQTL pairs work in trans remains vague. I therefore hypothesised that inter individual differences in chromatin organisation and DNA methylation are tightly linked based on previous work indicating that TADs play a role in gene regulation [Chathoth et al., 2022]. With this aim, I showed that SMP mQTL pairs are depleted at chromatin loops, whereas VMPs are found at chromatin loop anchors no greater than I would expect by chance based on the Illumina EPIC array. This seems to suggest that SMP mQTL pairs in trans are directed by another mechanism beyond chromatin looping, but exploring this was beyond the scope of this thesis. However, it does suggest that chromatin looping is in part responsible for the relationship between SNPs and variability in DNA methylation at sites across the genome in trans, but this is no more than I would expect by chance. Moreover, I next hypothesised that perhaps the trans mQTL pairs may be located in the same topologically associated domains (TADs). TADs are segments of DNA which vary from 100's of kb to a few million bases and are considered the unit of chromatin organisation. As I could not validate chromatin looping as a mechanism for SMP mQTL pairs in trans, I considered that large TADs may harbour SNPs and CpG sites, putting them in physical contact with one another. Here, I found a similar trend observed with chromatin loops, trans SMP mQTL pairs were identified to be in different TADs or no annotated TADS more often than they were in the same TAD. However, VMP mQTL pairs were located in the same TAD no more than I would expect by chance too. This indicates that mQTL pairs are in part influenced by or influence chromatin organisation, suggesting that 3D genome organisation may in part explain the inter individual variability in DNA methylation. Furthermore, it suggests that stable DNA methylation is less affected or has less of an effect on chromatin organisation too. However, it is important to note that although the design of the Illumina EPIC array was curated to include more distal regulatory elements such as enhancers, it is possible that the design of the array was not ideal to answer this research question. Therefore, future work aiming to explore the interplay between DNA methylation and chromatin organisation should aim to use methods which have increased coverage of enhancers and open sea regions of the genome, such as whole genome bisulphite sequencing.

The results collected thus far in this research suggest that the relationship between inter individual differences in DNA methylation, gene expression and phenotype is highly complex. Further, it is highly influenced by several factors, including chromatin organisation, epigenetic interplay and the presence of transcription factors. I suggest that unravelling the rules and role of DNA methylation in gene regulation is a highly demanding task, and one which is too complex to do well when looking genome wide. In order to understand this relationship, I believe it is necessary to focus on a specific set of loci or alternatively, a specific disease or disorder to be able to reveal meaningful biological results. Therefore, I chose to focus the final work in this chapter on a subset of loci known as epialleles. Epialleles are described as regions at which the epigenetic state varies amongst individuals within a population [Finer et al., 2011]. DNA methylation at epialleles occurs during embryonic development and leads to prominent inter individual variation [Harris et al., 2013]. Further, intermediate methylation states have previously been reported as a feature of epialleles in humans [Kessler et al., 2018] with a bimodal distribution of either very high or low methylation. Therefore, I leveraged our catalogue of VMPs as a means of identifying epiallele like sites as described in Chapter 2. Using this approach, I was able to identify 784 epiallele like sites in human whole blood which showed intermediate methylation levels across our sample, yet upon closer inspection demonstrated bimodal distribution, with either low or high methylation within individuals. I observed that just over half of these are unique to the Illumina EPIC array, meaning I was able to reveal novel epiallele like sites. At this point, it is important to note that these are not definite epialleles, for one to make that conclusion, they would need to be experimentally validated. However, we present them as sites which behave like epialleles and could be an interesting avenue for future research. The establishment of these epiallele like sites remains unclear, but may result from several factors. One possibility is that the differences in methylation state at these sites is due to genotype. Whilst I screened out CpGs related to SNPs as described in section 2, some epiallele sites may be influenced by mQTLs. Here, I was able to identify enrichment for mQTLs at epiallele like sites and these were cis and trans relationships. However, there were still a portion of epiallele like sites which were not linked to any SNPs via our mQTL analysis, suggesting that stochastic or environmental exposures may play a role in the establishment of these epiallele like sites in whole blood. I further demonstrated that some of these epiallele like sites were consistent across tissues, but most were tissue specific. This supports the idea that some of these epiallele like sites are caused by the aforementioned stochastic or environmental effects.

Epiallelic variation is functionally characterised in terms of its influence over gene transcription [Finer et al., 2011], however, the direction of effect remains debated in the literature. For example, one study suggests that differential epigenetic modification of two distinct variants is associated with tissue-specific changes in adjacent gene expression [Kessler et al., 2018]. Therefore, I explored this idea further by focusing not only on the relationship between DNA methylation variance and target gene expression variance at these epiallele like sites, but also considered the genomic context in which DNAm was found, to try and unravel this relationship further. Interestingly, it is observed that DNA methylation variance and target gene expression variance are significantly positively correlated at 5' UTRs . The 5 untranslated region (5' UTR) lies upstream of the coding sequence and plays a vital role in directing gene expression. Encoded within 5' UTR DNA sequences are numerous cis regulatory elements that can interact with the transcriptional machinery to regulate messenger RNA (mRNA) abundance [Lim et al., 2021]. Methylation in this region has previously been described to be associated with changes to gene expression [Rauluseviciute et al., 2020]. This suggests that inter individual differences at epiallele like sites at 5' UTRs are linked to corresponding variance in expression of their target genes. It is important to note, that this analysis was performed in non matching data sets (i.e. the data sets were collected from individual cohorts). Nevertheless, this does make it particularly interesting that I was able to identify this link. Future work should aim to replicate these findings in matched data sets, in order to rule out the idea that I missed potential links due to differing alleles between the cohorts used in this research, and also to validate the link I identified.

The identification of epiallele like sites that are linked to gene expression and are susceptible to genetic and environmental exposures may be indicative of an adaptive mechanism as previously described [Feinberg and Irizarry, 2010, Kessler et al., 2018], making these sites important for research exploring adaptive responses to environmental exposures.

Lastly, I acknowledge that one caveat of this research was that I was unable to expand this line of research to consider the role of cell type specificity with regards to DNA methylation variation. Therefore, future work should aim to consider the presence of differing cell types within bulk tissues. However, I am confident that providing a catalogue of VMPs, SMPs and epiallele like sites will be extremely useful to the community working on epigenome wide association studies and such in human whole blood.

Chapter 6

General discussion

Since the adoption and popularity of genetic epidemiology, thousands of genetic variants have been identified that are able to explain some of the diversity observed in human phenotypes and disease. Despite this, a large proportion of this diversity has not yet been explained by genetic variation. Where human phenotypic diversity is not able to be explained by genetic factors alone, it is attributed to environmental and lifestyle exposures. Moreover, complex diseases and phenotypes are likely to be explained by both between genetic and environmental factors. This has resulted in the growing focus on the field of epigenetics. Studying epigenetics allows researchers to study the interplay between genetics and environment, and take a closer look at the complex biological pathways underlying human disease and variation. During this thesis, I aimed to characterise several factors which influence the human epigenome, and to also investigate the biological and gene regulation pathways they influence. Several conclusions have resulted from this thesis;

6.0.1 Key findings and their implications in the field

There are no significant detectable DNA methylation changes in response to exposure to background levels of ambient air pollution in the UK

To address the aim of chapter 3 of this thesis to investigate how background levels of air pollution may modulate the human epigenome. I leveraged data from Understanding Society and performed an epigenome wide association study of the four main pollutants (Nitrogen dioxide, Ozone and Particulate matter 10 and 2.5). This involved using a measure of average air pollution exposure over a five year period prior to blood collection and DNA methylation measurement (2007-2011). In the first instance I found no significant detectable DNA methylation changes related to air pollution exposure for PM10, NO2 or O3. However, I did originally identify 43 CpGs significantly associated with PM2.5 following FDR correction, which I found to be positioned in or near genes which have previously been reported in pollution related EWAS'. I concluded that these results were being driven by two outliers in our cohort, and following removal of these samples, I concluded that there were also no significant detectable DNA methylation changes in relation to PM2.5 either. These data suggest that considering the study region in research aiming to explore how air pollution influences the epigenome may be neccessary. In other words, it may be more biologically insightful to focus on vulnerable populations, such as those living in highly polluted rather than areas with lower air pollution levels like the UK. Although, it is important to note that just because this research did not reveal any significant DNA methylation changes in response to air pollution in the UK, this does not mean that the levels of air pollution in the UK do not have any adverse health effects through epigenetic mechanisms. Epigenetic changes caused by lower levels of air pollution exposure in the UK may be small enough to not be detected by EWAS but could have a cumulative effect, resulting in adverse health effects over time. This offers a potentially interesting avenue for future work, discussed in section 6.0.2.

A distance to road measure gives insight into how exposure to traffic related air pollution may modify DNA methylation signatures across the genome

In line with our second aim of chapter 3 of this thesis, which was to investigate how traffic related air pollution may influence DNA methylation patterns in human whole blood, I performed EWAS using a distance to road measure. This consisted of a measure of how close a participant lives to a busy road. I identified 531 significant CpG sites associated with traffic related air pollution. Of these, 266 CpGs were hypomethylated in response to TRAP exposure and 265 were hypermethylated in response to TRAP exposure. Gene ontology analyses showed no enriched terms for these 531 CpGs following FDR adjustment. Similarly, KEGG analyses also revealed no enriched terms either. However, I found that many of these CpG sites were intragenic, specifically, these were enriched in 3'UTRs, enhancers, and CpG shores indicating a regulatory potential for these CpGs. Expanding on this, I was able to identify an enrichment of 65 transcription factors, some of which have previously been linked to exposure to air pollution meaning I was able to validate previously reported links and also produce some novel links. Lastly, I also reveal 3 novel differentially methylated regions linked to exposure to traffic related air pollution overlapping two genes, MYOM3 and FRA10AC11. This research supports the previously reported idea that ambient air pollution and traffic related air pollution may have distinct adverse health effects, and further provides the new idea that this may occur through separate epigenetic mechanisms. Additionally, this could also mean that different types of air pollution may be attributed to distinct diseases, but as aforementioned, this work needs to be validated in additional cohorts before such far fetched conclusions can be drawn. Further, this work could have benefited from including an additional measure such as lung function measurements to validate the associations made in this work and our pollution measures. Lastly, I note that it is unclear whether using a 5 year average mean for pollution exposure obscured any potential association.

As the data included in this measure was not longitudinal, I had no information regarding how long the individuals had lived at the address used in our models. As some epigenetic signatures have previously been reported to be reversible, it is possible that this could be an issue in this data. Nevertheless, this work was able to validate several loci already established within the literature, and also provide the community with novel loci for future work.

Autosomal epigenetic differences exist between males and females in whole blood which have implications for sex specific gene regulatory networks

Sex differences are known to play a role in disease aetiology, progression and outcome. Previous studies have revealed autosomal epigenetic differences between males and females in some tissues, including differences in DNA methylation patterns. In chapter 4 of this thesis, I characterised autosomal sex differences in DNAm using the Illumina EPIC array in human whole blood by performing a discovery (n=1171) and validation (n=2471) analysis. I identified and validated 396 sex-associated differentially methylated CpG sites (saDMPs) with the majority found to be female-biased CpGs (74%). The catalogues provided from this chapter will be important for future studies such as epigenome wide association studies aiming to question sex specific effects. I also demonstrated that these saDMP's are enriched in CpG islands and CpG shores and located preferentially at 5'UTRs, 3'UTRs and enhancers. This work confirmed trends previously reported in the literature, and motivated the additional research presented in chapter 4. Thereby, I additionally identified 266 significant sex-associated differentially methylated regions overlapping genes, which have previously been shown to exhibit epigenetic sex differences, and novel genes. These regions displayed a large overlap with previously reported sex DMRs reported by [Gatev et al., 2021]. Furthermore, transcription factor binding site enrichment revealed enrichment of transcription factors related to critical developmental processes and sex determination such as

SRY and ESR1 indicating that these CpG sites have implications for sex specific gene regulatory networks. This idea was confirmed by my analysis of Hi-C data too. This chapter reports a reliable catalogue of sex-associated CpG sites and elucidates several characteristics of these sites using large-scale discovery and validation data sets. This research provided the community with a valuable resource and advice for defining a sex associated site or region in order to increase replication across studies. The results from this chapter will also benefit future studies aiming to investigate sex specific epigenetic signatures and further our understanding of the role of DNA methylation in sex differences in human whole blood.

Interindividual differences in DNA methylation are widespread across the genome, and a subset of CpG sites show high variability or stability in DNA methylation.

The study presented in chapter 5 of this thesis, was motivated by the fact that inter individual genetic variability is well characterised, yet I are still lacking a complete catalogue of autosomal loci displaying variable and stable epigenetic patterns across the human epigenome. In this research, I report a catalogue of loci across the human whole blood epigenome displaying either variable or stable inter- individual DNA methylation by analysing the DNA methylation patterns in 3642 individuals (n =1171 for discovery and n = 2471 for validation). This research showed that 35,142 CpGs display robust stable methylation (stably methylated probes SMPs) and 40,288 CpGs display highly variable methylation levels (variably methylated probes VMPs). The results from this chapter are important for studies investigating differential methylation in human whole blood as it provides a reference resource for researchers interested in DNA methylation variability and stability. I report that SMPs are highly enriched in CpG islands and depleted at CpG shelves and open sea regions of the genome, which are novel findings in the field. Moreover, I demonstrate that the majority of the VMPs are not controlled by age, sex or smoking status. I also report highly variable and stable CpG sites enriched at methylation sensitive TFs, highlighting that there is potentially a strong relationship between DNA methylation and transcription factor binding, which will be important for gene regulation. In addition, I found that the VMPs were under higher genetic control than the SMPs and further analysis revealed that trans mQTL pairs are often located in the same TAD or connected by chromatin loops. This finding allowed us to confirm theories in the field that mQTL relationships may be influenced by or influence the 3D chromatin organisation.

Identification of a relationship between epigenetic variation and gene expression variation among individuals

A subset of VMPs (n=784) located in 5'UTRs exhibit a link with gene expression. These results suggest to the community that when investigating the relationship between DNA methylation and gene expression, it is important to consider genomic context, as the location of DNA methylation is important for its relationship with gene expression. These results confirm the idea proposed by others [de Mendoza et al., 2022], that this relationship is not a linear one and is very context dependent.

6.0.2 Limitations and future research directions

It is important to note that the research discussed in this thesis also possess several methodological limitations. Each chapter discusses a unique and complete set of limitations, but several limitations were shared amongst chapters, of which I will discuss here.

Sample size and collection

Several of the studies discussed in this thesis are based on samples which are fairly limited. For example, RNA-seq data used in chapter 4 was based on only 20 human samples (10 males and 10 females). The reason for using this particular data set, is that I focused the analysis in this thesis on whole blood which is a bulk tissue.

I therefore searched for RNA-seq data which was also based on whole blood, and this was the most suitable resource. Future studies should aim to work with larger sample sizes, and also with matched data sets as this would provide more robust and invaluable conclusions about the relationship between DNA methylation and target gene expression. Further to this, as I did not have access to gene expression or other genomic data for our participants involved in Understanding Society, I leveraged several publicly available data sets in order to investigate gene expression changes and 3D chromatin organisation. Future work will benefit largely from being able to leverage matched data sets in order to provide more robust findings in regards to genomic interplay.

Moreover, I was able to obtain reasonable sample sizes for Chapter 4 and Chapter 5 of this thesis as the data I was accessing had no privacy restrictions due to the nature of the study. However, for Chapter 3 of this thesis, I was unable to obtain a validation data set within the timeframe of this PhD due to the data access process. Therefore, in the time period for this thesis, I was unable to obtain a suitable validation data set for this chapter. Future aims include obtaining this validation data set, in order to validate our findings presented in Chapter 3 of this thesis.

Also, I acknowledge that several of our findings may be limited or lack generalisability due to the fact that all of our participants were of white ethnicity. I especially acknowledge that our work aiming leveraging genetic data may be limited because of the potential lack of diverse genetic diversity. Therefore, it is important that future work considers the use of more diverse populations in such research.

Lastly, it is important to note that I focused our analyses on DNA obtained from whole blood samples, which is a bulk tissue comprised of individual cell types. This means that our results may not be relevant for other tissues. As mentioned in the Introduction of this thesis, it is also important to note that, especially for Chapter 3 of this thesis, blood may not have been the most relevant tissue for the research question. Other tissues which may have been more relevant may have been oesophagus or lung tissue. Despite this, it is also important to remember that there are several advantages to studying whole blood. The first being that it means that sample collection is a lot easier, less invasive and a more cost effective method. It also has clinical benefits, meaning that identified DMPs and DMRs resulting from work presented in this thesis may have the potential to later act as biomarkers. Therefore, future work should consider these limitations and strengths in their study design.

Technology

With respect to DNA methylation measurements, I focused the majority of our analysis on one main technology which comes with particular limitations. The Illumina EPIC array used in Chapters 3-5 of this thesis is limited to measuring approximately 850,000 CpG sites across the genome, which is only a small fraction of the total (approximately 28 million) CpG sites in the human genome. Although this technology has been carefully designed to cover more distal regulatory elements and relevant protein coding genes, it still lacks coverage of many distal regions of the genome. Future work may benefit from intergrating other methods with increased coverage of the genome such as WGBS.

Causality

Analyses presented in each chapter in this thesis are cross sectional, meaning they are all based off of epigenetic information collected at a single time point. Studies of this nature make drawing conclusions about causality difficult. Chapter 3 of this thesis benefited from a coincidental 'natural experiment' as I was able to collect pollution measurements from a 5 year period prior to blood collection. However, it is impossible for us to rule out other factors such as socioeconomic factors confounding our results. Furthermore, Chapter 4 and Chapter 5 of this thesis present some exciting results regarding particular genes, biological pathways and gene regulation. However, it is also impossible for us to attribute for example, gene expression changes to DNA methylation changes, rather than the other way round. In other words, I also cannot definitively say that DNA methylation changes arise as a consequence of disease or whether DNA methylation changes result as a consequence of the disease. Future work should aim to produce more complex maps of disease, incorporating more data to reveal more about epigenetic regulation of complex diseases and phenotypes.

6.0.3 Final words

Epigenomics is a growing field with numerous branches and recent developments. The field is beginning to appreciate the link between DNA methylation, other epigenetic factors and gene regulation. This thesis aimed to explore and describe some of the factors that influence and are influenced by DNA methylation patterns in an effort to advance our understanding of the dynamic nature of the human epigenome. The findings produced in this thesis provide the community with important resources that will be beneficial in future studies aiming to explore external factors influencing DNA methylation patterns which will help advance our understanding of complex human diseases and phenotypes.

Appendix A

Appendix

Additional files A full list of all additional files referenced throughout this thesis are provided at https://github.com/livygrant97/thesisAppendices.git.

1. Chapter 3 folder includes the full list of raDMPs identified in the distance to road EWAS.
2. Chapter 4 folder includes the list of saDMPs, saDMRs and enriched TF motifs identified by the analyses performed in chapter 4.
3. Chapter 5 folder includes the list of VMPs and SMPs identified by the analyses performed in chapter 5 of this thesis.

Table A.1: Enriched KEGG terms for VMPs and SMPs

Enriched KEGG terms for SMPs

Description	N	DE	P.DE	FDR
Amyotrophic lateral sclerosis	351	265	6,28122063667143e-15	2,21727088474501e-12
Ribosome	134	108	3,83876639434824e-14	6,77542268602464e-12
Spliceosome	132	112	7,91161858131672e-13	9,30933786401601e-11
Parkinson disease	252	193	1,47771751366563e-12	1,30408570580992e-10
Cell cycle	127	110	2,95604276231233e-12	2,0869661901925e-10
Pathways of neurodegeneration - multiple diseases	462	333	8,10586829637703e-12	4,76895251436849e-10
Alzheimer disease	370	271	1,50070757741972e-11	7,56785392613087e-10
Huntington disease	292	217	3,51626417177965e-11	1,55155156579777e-09
Shigellosis	244	188	5,79018166509307e-11	2,27103791975317e-09
Protein processing in endoplasmic reticulum	169	137	7,54673414710612e-11	2,66399715392846e-09
mTOR signaling pathway	155	129	8,78760976760718e-11	2,82002386178667e-09
Oxidative phosphorylation	121	97	2,2300125731345e-10	6,55995365263732e-09
Thermogenesis	219	165	3,40217144992078e-10	9,23820401401567e-09

Nucleocytoplasmic transport	107	91	4,1923778216466e-10	1,05707812217232e-08
Metabolic pathways	1520	960	5,4918214935511e-10	1,29240865814902e-08
Prion disease	259	188	1,95789911948829e-09	4,31961493237104e-08
Ubiquitin mediated proteolysis	142	117	2,72565075958291e-09	5,06397220069877e-08
Endocytosis	250	193	2,67319877609672e-09	5,06397220069877e-08
Human papillomavirus infection	330	243	2,58302454061205e-09	5,06397220069877e-08
Salmonella infection	248	186	9,65910682471857e-09	1,70483235456283e-07
Lysosome	132	107	1,45490779197195e-08	2,44563071698142e-07
p53 signaling pathway	73	65	2,07368240990917e-08	3,32731768499062e-07
RNA degradation	79	67	3,64316881515636e-08	5,59147213804433e-07
Non-alcoholic fatty liver disease	151	115	6,07384643024182e-08	8,93361579114735e-07
Aminoacyl-tRNA biosynthesis	43	41	6,62322928392709e-08	9,35199974890504e-07
Cellular senescence	155	123	7,03241228125542e-08	9,5478520587814e-07
Hepatocellular carcinoma	168	132	7,56743751176594e-08	9,89372385797547e-07
Chronic myeloid leukemia	76	67	1,45689697077046e-07	1,83673082386419e-06
mRNA surveillance pathway	95	77	1,65484563259968e-07	2,01434658037133e-06
Proteasome	46	42	1,7525083919525e-07	2,0621182078641e-06

Ribosome biogenesis in eukaryotes	76	62	2,25833106396968e-07	2,57158343735902e-06
Proteoglycans in cancer	204	155	4,38633309553444e-07	4,83867369601143e-06
Diabetic cardiomyopathy	189	137	7,35093955530803e-07	7,86327776673859e-06
Viral carcinogenesis	193	141	1,19313842083981e-06	1,23875841928369e-05
Autophagy - animal	140	109	3,23573425894806e-06	3,26346912402475e-05
Neurotrophin signaling pathway	119	95	3,34862279106797e-06	3,28351068124165e-05
Pathways in cancer	529	355	9,22607496892237e-06	8,80217422710701e-05
Insulin signaling pathway	137	105	1,22318057380546e-05	0,000113627037514034
Hepatitis B	162	117	2,0755261774501e-05	0,000187861728369201
Protein export	23	22	2,26024468994379e-05	0,000199466593887539
Bacterial invasion of epithelial cells	77	64	2,33406429394967e-05	0,000200957242869325
Pancreatic cancer	76	63	2,82469897306439e-05	0,000237409223212316
Basal cell carcinoma	63	53	2,92865109557689e-05	0,000240421822497358
AMPK signaling pathway	120	93	3,36940611651206e-05	0,000268453281014652
Pathogenic Escherichia coli infection	196	140	3,42220896477602e-05	0,000268453281014652
Colorectal cancer	86	69	4,37258388979409e-05	0,000335548285455938
Yersinia infection	135	102	4,99864012234574e-05	0,000375429779401712

Chemical carcinogenesis - reactive oxygen species	209	143	5,33989998021755e-05	0,000392705144378499
Longevity regulating pathway	89	71	5,5421314174805e-05	0,000399259671504208
Vibrio cholerae infection	50	42	7,35047831314703e-05	0,000508768400890373
Human cytomegalovirus infection	225	154	7,24606929402117e-05	0,000508768400890373
EGFR tyrosine kinase inhibitor resistance	78	64	8,74667469087359e-05	0,000593764647284303
Epstein-Barr virus infection	199	139	0,000102372493468261	0,000681839437628229
Base excision repair	33	29	0,000107771033517345	0,000704503237622645
Prostate cancer	97	76	0,00011431415727259	0,000733689045767715
Viral life cycle - HIV-1	63	50	0,000123251924138016	0,000776927307512851
Human immunodeficiency virus 1 infection	211	143	0,000129305197858292	0,000800784821824158
Non-small cell lung cancer	72	59	0,000135720973630878	0,000826025925718964
Apoptosis	135	99	0,000143715595053966	0,000859857712780506
Necroptosis	157	104	0,000167335341466477	0,00098448959229444
Hippo signaling pathway	157	117	0,000179647339085533	0,00103959853601956
Notch signaling pathway	58	48	0,00019787611707035	0,00112661724719086
Coronavirus disease - COVID-19	231	142	0,000213466251292555	0,00119351054566879
Glioma	75	60	0,000216387181084426	0,00119351054566879

Spinocerebellar ataxia	142	105	0,000253491517351228	0,00137665393269206
Endocrine resistance	96	75	0,000266263901250229	0,00142410844153531
N-Glycan biosynthesis	50	41	0,000303622142344774	0,00157615612128978
Insulin resistance	108	81	0,000301393919488287	0,00157615612128978
Breast cancer	147	108	0,000316740602661474	0,00162042656144203
Human T-cell leukemia virus 1 infection	219	154	0,000367491521422416	0,00185320724374447
Renal cell carcinoma	68	55	0,000386847161957786	0,00192333870663519
Cushing syndrome	155	113	0,000466201088024551	0,00228568033434259
FoxO signaling pathway	131	95	0,000508173248435117	0,0024573309136657
Gastric cancer	149	108	0,000548927288692065	0,00261853152578782
Various types of N-glycan biosynthesis	39	33	0,000669719262003778	0,00315214532649778
DNA replication	36	30	0,000763439097795538	0,00353583605703785
Hepatitis C	157	106	0,000771273020940267	0,00353583605703785
Hedgehog signaling pathway	56	46	0,000821624312240712	0,00371837669514066
Kaposi sarcoma-associated herpesvirus infection	194	129	0,000886477288587348	0,0039610947198903
Tight junction	169	118	0,00102978025007159	0,00448780775648484
Endometrial cancer	58	47	0,00102958143960272	0,00448780775648484

Adherens junction	71	56	0,00110537116140705	0,00470115686718903
Small cell lung cancer	92	70	0,00110152679335838	0,00470115686718903
PD-L1 expression and PD-1 checkpoint pathway in cancer	89	66	0,00150752709905882	0,00633520316628289
Mitophagy - animal	71	54	0,00154180161208694	0,00640301140078458
Oocyte meiosis	126	88	0,00172246532470681	0,00707011929792446
Sphingolipid signaling pathway	118	86	0,00200549545610396	0,00813724018396206
Focal adhesion	200	142	0,00204912873718499	0,00821980050257159
Terpenoid backbone biosynthesis	23	20	0,00249490928256442	0,00989553906455327
Phosphatidylinositol signaling system	97	72	0,00314494744729239	0,0123351827654913
Epithelial cell signaling in Helicobacter pylori infection	70	52	0,00328711781293906	0,0127511273403021
Regulation of actin cytoskeleton	228	155	0,00370379827737956	0,0142113129555977
RNA polymerase	34	25	0,00388269689833619	0,0146706444756503
MAPK signaling pathway	294	198	0,00390663054025816	0,0146706444756503
TNF signaling pathway	112	78	0,00397285117408299	0,0147622785731715
Signaling pathways regulating pluripotency of stem cells	142	100	0,0042084709121646	0,0154748982499386
Thyroid hormone signaling pathway	121	88	0,00481004263864616	0,0175045881591969
Thyroid cancer	37	30	0,00509737708914352	0,0183609603313027

Choline metabolism in cancer	98	72	0,00535996099187389	0,0191117801023382
Other glycan degradation	18	16	0,00603438399320163	0,0213013754960018
Fc gamma R-mediated phagocytosis	96	69	0,00621317655651187	0,0217153596480068
Ferroptosis	41	32	0,00638083588192775	0,0220826967286323
Fanconi anemia pathway	53	39	0,0068124578658109	0,0233475497731189
Nucleotide excision repair	46	34	0,00827867615657796	0,028099737339154
Growth hormone synthesis, secretion and action	120	85	0,00868533398513894	0,0291992656833719
Selenocompound metabolism	17	15	0,00884730238042067	0,0294631862291368
Parathyroid hormone synthesis, secretion and action	106	76	0,00933748154921164	0,0308049624941281
Lysine degradation	63	47	0,00955265201093819	0,0312230199987146
Basal transcription factors	44	32	0,0104113051102941	0,0337173459076497
Non-homologous end-joining	13	12	0,0106399110727049	0,034144441896953
Valine, leucine and isoleucine degradation	48	36	0,0112333783321689	0,0357241671284289
Prolactin signaling pathway	70	51	0,0119061030084194	0,0375254853747505
One carbon pool by folate	20	17	0,0121476550317687	0,0379479843027819
Melanogenesis	101	71	0,0125586623645185	0,0388877878480267
Wnt signaling pathway	169	116	0,0129618352621462	0,0397871986742402

Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	26	20	0,0139020096461422	0,0409728164774185
Platinum drug resistance	73	51	0,0137100191288479	0,0409728164774185
HIF-1 signaling pathway	109	75	0,0136467688844695	0,0409728164774185
VEGF signaling pathway	59	44	0,0139284361962896	0,0409728164774185
Glucagon signaling pathway	105	71	0,0139079206134487	0,0409728164774185
Central carbon metabolism in cancer	70	51	0,0141732165106753	0,0413483093245321
Measles	139	88	0,0149435620874029	0,0432383394824035

Enriched KEGG terms for VMPs

Description	N	DE	P.DE	FDR
Olfactory transduction	407	162	1.76027586223357e-09	6.21377379368449e-07
Neuroactive ligand-receptor interaction	363	195	3.53430651674902e-06	0.000623805100206203

Grant *et al. Clinical Epigenetics* (2022) 14:62
https://doi.org/10.1186/s13148-022-01279-7

Clinical Epigenetics

RESEARCH

Open Access

Characterising sex differences of autosomal DNA methylation in whole blood using the Illumina EPIC array

Olivia A. Grant[1,2,3], Yucheng Wang[1,4], Meena Kumari[2], Nicolae Radu Zabet[1,3*] and Leonard Schalkwyk[1*]

Abstract

Background: Sex differences are known to play a role in disease aetiology, progression and outcome. Previous studies have revealed autosomal epigenetic differences between males and females in some tissues, including differences in DNA methylation patterns. Here, we report for the first time an analysis of autosomal sex differences in DNAme using the Illumina EPIC array in human whole blood by performing a discovery ($n = 1171$) and validation ($n = 2471$) analysis.

Results: We identified and validated 396 sex-associated differentially methylated CpG sites (saDMPs) with the majority found to be female-biased CpGs (74%). These saDMP's are enriched in CpG islands and CpG shores and located preferentially at 5'UTRs, 3'UTRs and enhancers. Additionally, we identified 266 significant sex-associated differentially methylated regions overlapping genes, which have previously been shown to exhibit epigenetic sex differences, and novel genes. Transcription factor binding site enrichment revealed enrichment of transcription factors related to critical developmental processes and sex determination such as SRY and ESR1.

Conclusion: Our study reports a reliable catalogue of sex-associated CpG sites and elucidates several characteristics of these sites using large-scale discovery and validation data sets. This resource will benefit future studies aiming to investigate sex specific epigenetic signatures and further our understanding of the role of DNA methylation in sex differences in human whole blood.

Keywords: Epigenetics, DNA methylation, Gene regulation, Autosomes, Sex differences, Sex, Illumina EPIC array

Introduction

Sex is an important covariate in all epigenetic research due to its role in the incidence, progression and outcome of many phenotypic characteristics and human diseases [1, 2]. There is an increasing interest as to which role epigenetic modifications (such as DNA methylation) may play in the underpinnings for relationships between environmental exposures and disease onset. In addition, sex has previously been shown to have a strong influence on DNA methylation variation [3–7]. However, the idea that DNA methylation variation between males and females may underlie the sex biases observed in diseases has not been well documented thus far.

Sex differences in disease prevalence are sometimes explained at the molecular level and rooted in genetic differences between males and females. Differences in sex chromosome complement have independently been shown to direct differences in gene expression and chromatin organisation [8–11]. Furthermore, these differences in sex chromosome complement are sufficient to explain sex bias seen in some diseases. For example, X chromosome number has previously been shown to

*Correspondence: r.zabet@qmul.ac.uk; lschal@essex.ac.uk

[1] School of Life Sciences, University of Essex, Colchester CO4 3SQ, UK
[3] Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK
Full list of author information is available at the end of the article

impact immune cell population and occasionally therefore the development of diseases such as autoimmunity [12, 13].

Previous research has also revealed sex differences in gene expression of autosomal genes as well as sex chromosome linked genes [14]. It is worth noting that most of the differences in gene expression on the autosomes are small differences [15]. However, small expression differences may still be associated with great effects on phenotypic characteristics and disease incidence and onset. Others also identified sex differences in chromatin accessibility and histone modifications, thus suggesting that different epigenetic factors contribute to gene expression sex biases seen in some diseases [16].

Sex specific gene expression and levels of sex hormones may be mediated by epigenetic mechanisms, including DNA methylation. Several genome wide association methylome studies (or Epigenome Wide Association Studies, EWAS) have highlighted differences in DNA methylation patterns linked to sex differences in genes on the autosomes [15–18]. Previous studies have reported sites and regions showing varying methylation due to sex differences in several tissues such as saliva, placenta, brain, pancreatic islets and whole blood [15, 17, 19–28]. These studies highlight the presence of autosomal loci displaying sex-biased DNA methylation patterns across the genome for several tissues. In order to determine their role in disease and developmental processes, these loci warrant further exploration.

However, due to X chromosome inactivation in females, large differences in methylation levels of X chromosomes can be observed between males and females [29]. Recent research suggests that normalising methylation data with the sex chromosomes introduces a large technical bias to many autosomal CpGs [30]. This technical bias has been reported to result in many autosomal CpG sites being falsely associated with sex even when male and female samples are normalised independently of each other, a method employed by some studies in the field. Moreover, it also leads to many autosomal CpGs being incorrectly identified to be more methylated in male samples compared to female samples. Therefore, the breadth of autosomal DNA methylation variation between males and females is still not well understood and requires further clarification. Extra steps were therefore employed in this study by applying a normalisation method which aims to reduce bias introduced to autosomal CpGs [30] to uncover true biological differences and determine patterns of global DNA methylation levels between males and females.

Additionally, it is worth noting that thousands of autosomal CpGs do show very small differences in DNA methylation patterns between males and females. However, a robust and well-annotated catalogue of sites showing the largest differences still needs to be characterised.

Here, we use the EPIC BeadChip to assess autosomal sex differences in DNA methylation levels from whole blood at individual sites and genomic regions. All individuals involved in this study were part of Understanding Society: The UK Household longitudinal study [31]. Additionally, we adequately handle the technical bias introduced by sex chromosomes. To our knowledge, this is the largest study using the Illumina EPIC BeadChip (allowing for interrogation of ~ 850,000 sites across the genome) to investigate autosomal sex differences in DNA methylation at CpG sites in whole blood.

Results

Females show higher methylation at a subset of autosomal loci

Analysis of DNA methylation (DNAme) differences between males and females on the autosomes was performed using linear regression for the Illumina EPIC BeadChip for 1171 individuals (682 females and 489 males) for discovery and repeated in a validation data set of 2471 participants (1339 females and 1132 males). After data processing and cleaning, $n = 747{,}302$ CpGs were analysed (see *Material and Methods*). Sites which are known SNP probes, cross hybridising or X/Y linked probes were excluded. Moreover, since whole blood is a bulk tissue, we calculated the estimated cell type proportions for whole blood between our male and female samples to assess whether any differences in cell type proportions would potentially be reflected in our results resulting in false positives. Using Wilcoxon test, we found no significant difference in the proportions of Granulocytes between males and females, but we did find statistically significant differences in proportions of CD4T, CD8T, Natural killer, B cells and monocytes (Additional file 6: Figure S1B and S1D). We therefore included cell type proportions in our models for identifying sex-associated differentially methylated probes and regions. After adjusting for multiple testing using the Benjamini–Hochberg FDR method (FDR $p < 0.05$), we identified 54,261 autosomal CpGs associated with sex in our discovery and validation data set (Additional file 6: Figure S1C). Of those CpGs, 60% (33,103 CpGs) were more highly methylated in females and the remaining 40% (21,788 CpGs) were more methylated in males. Gene ontology analyses showed several enriched terms for these 54,261 autosomal CpGs (Table 1) which included terms related to mammalian sex determination and gonad development, specifically several signalling pathways such as Ras signalling, MAPK signalling, Wnt and Hippo signalling [32–1]. Other terms included pathways related to cancer

Table 1 Enriched GO terms among the 54,261 CpGs identified to be significantly associated with sex

Path	Description	*N*	DE	P.DE	FDR
path:hsa04020	Calcium signalling pathway	240	192	4.10E−09	1.41E−06
path:hsa04015	Rap1 signalling pathway	210	169	1.15E−07	1.98E−05
path:hsa05200	Pathways in cancer	531	382.8333	1.91E−07	2.19E−05
path:hsa04014	Ras signalling pathway	232	180	2.97E−07	2.56E−05
path:hsa04010	MAPK signalling pathway	294	223.33333	2.24E−06	0.00015382
path:hsa04360	Axon guidance	182	148.5	3.46E−06	0.00019826
path:hsa04072	Phospholipase D signalling pathway	148	121	4.51E−06	0.00022149
path:hsa04310	Wnt signalling pathway	166	129.5	7.50E−05	0.00322397
path:hsa04371	Apelin signalling pathway	139	107.5	0.00011119	0.00363348
path:hsa04724	Glutamatergic synapse	114	93	0.00011675	0.00363348
path:hsa04390	Hippo signalling pathway	157	123.5	0.00012001	0.00363348
path:hsa01521	EGFR tyrosine kinase inhibitor resistance	79	67.5	0.00013811	0.00363348
path:hsa04071	Sphingolipid signalling pathway	119	94.5	0.00013944	0.00363348
path:hsa05226	Gastric cancer	149	115.5	0.00014787	0.00363348
path:hsa04550	Signalling pathways regulating pluripotency of stem cells	143	108.5	0.00059275	0.01359379
path:hsa04151	PI3K-Akt signalling pathway	354	245	0.00076003	0.01634056
path:hsa05224	Breast cancer	147	111.5	0.00092102	0.01863701
path:hsa04725	Cholinergic synapse	113	88	0.00148428	0.02836619
path:hsa04961	Endocrine and other factor-regulated calcium reabsorption	53	44	0.00209794	0.03798375
path:hsa05225	Hepatocellular carcinoma	168	123.5	0.00284431	0.04892215

N indicates the number of genes in the KEGG term. DE refers to the number of genes annotated to the sex-associated DMPs which are differentially methylated. P.DE indicates the *P* value for over representation of the KEGG term in this data set. FDR indicates the false discovery rate (using the Benjamini and Hochberg method)

and cellular proliferation (Table 35). This is not surprising though, as there is overwhelming evidence that sex influences cancer risk, progression, and treatment response [36–38]. It is also now well accepted that sex differences may significantly impact on the cell biology of cancer [39]. Further, epigenetic dysregulation is also now accepted widely as a mechanism for cancer initiation and progression. This may be through transcriptional activation or repression of specific autosomal loci through means of DNA methylation. Therefore, one can hypothesise that sex specific patterns may influence the ability of cancer cells to adopt a stem cell like phenotype. This enables us to draw a link between epigenetic signatures and cancer pathways. It is likely that these sex differences in DNA methylation in part cause or are caused by differing levels of sex hormones such as androgen or oestrogen. This idea is supported by previous literature highlighting that DNA methylation transcriptionally represses masculinising genes and that this depends on gonadal hormones during development [39].

The lambda value of the Q-Q plots is slightly high (Additional file 6: Figure S1A and S1C) indicating slight inflation of test statistics and, in order to ensure we detect true sex differences, we selected CpGs that displayed large differences in methylation. Further, as a high proportion of CpG sites across the genome, we were also interested in investigating further, those CpG sites which show the largest differences between sexes. Thus, we further filtered our list of 54,261 CpGs by only considering those probes that displayed the largest sex differences, determined by a ΔBeta value (absolute difference between average Beta values in male and female samples) greater than 0.05. A total of 396 CpGs met this criterion (called sex-associated DMPs or saDMPs) in both our validation and discovery data sets and, from here on, are the focus of this manuscript (Additional file 6: Figure S1C). CpG sites which we identified to have higher methylation in females are from here, referred to as 'female-biased CpGs' and CpG sites which have higher methylation in males are here on referred to as 'male-biased CpGs'. We found that these saDMPs were distributed across all autosomes (Fig. 1A) with 74% of the saDMPs being female-biased CpGs (293 CpGs) and 26% being male-biased CpGs (103 CpGs) (Fig. 1B) (see Additional file 1 for the full list).

Since we had such stringent parameters to define what we considered a significantly associated saDMP for males and females, we performed principal component analysis (PCA) to see how male and female beta values clustered in PC space and to evaluate the effect of DNAme at the saDMPs. As shown in Fig. 1C, male and female samples formed distinct clusters based on the beta values of the significant sex-associated DMPs (396 CpGs). PC1 explained 16.1% of the variance and PC2 explained 4.2% of the variance. Based on Fig. 1C we can conclude that these saDMPs are sufficient to contribute to the clear separation of male and female samples in PC space.

Characterisation of sex-associated DMPs

The saDMPs were found in 174 unique genes with 48 of these genes harbouring several saDMPs (Fig. 1D). The number of saDMPs harboured by individual genes ranged from 1 to 8. CRISP2, a gene known to be involved in sperm function and male fertility [40], harboured the largest number of saDMPs, 8, which interestingly were all found to be female-biased CpGs. We performed GO and KEGG analyses but did not identify any significantly enriched biological processes or pathways for these genes. Nevertheless, the genes which did harbour saDMPs are biologically interesting, as many are genes known to be involved in sexual development and processes, such as SOX18 [41]. Further, some genes are already known to exhibit sex specific methylation patterns, such as PRR4 and PTPRN2 [42, 43]. Despite this, we were able to identify some novel genes which have not previously been reported to exhibit sex differences in DNA methylation such as GCK, HIP1R and KANK1.

To help us gain more insight into the functional role of these saDMPs, we characterised their genomic location and further compared this with the autosomal EPIC background. We found that saDMPs are preferentially located in CpG islands and CpG shores and depleted in open sea regions compared to the autosomal background (Fig. 1E). Moreover, female-biased CpGs are enriched at promoters and exons, with male-biased CpGs being enriched at 5'UTRs (Fig. 1F). Interestingly, we observed that all saDMPs display enrichment at enhancers, which, together with their presence at promoters, indicates that they could play a role in gene regulation. Lastly, we also note that all saDMPs were depleted at transposable elements and introns compared to the autosomal EPIC background.

Enrichment of saDMPs at enhancers suggests that some of the saDMPs could potentially regulate distal genes [44, 45]. We further annotated the saDMPs to genes by identifying if their contacts with promoters are mediated by 3D chromatin loops detected in Hi-C data. Following this, we further annotated the saDMPs to 37 additional genes, 28 of them being annotated to female-biased CpGs and 8 to male-biased CpGs (see Additional file 7: Figure S2A, B and Additional file 5).

Of the 8 genes linked to male-biased CpGs, we found three histones (HIST1H3A, HIST1H4A and HIST1H4B), which are known to interact with CDYL. Chromodomain Y-like protein (CDYL) is a chromatin reader binding to heterochromatin (H3K9me3, H3K27me2 and H3K27me3) that is crucial for spermatogenesis, male fertility and X chromosome inactivation [46]. In addition, ODF2L; outer dense fibre of sperm tails 2 like is linked to saDMPs female-biased CpGs and has previously been shown to interact with PRSS23, which is involved in ovulation [47].

Next, to evaluate whether the genes controlled by the saDMPs are part of the same regulatory network, we merged all proximal and distal genes and produced protein–protein interaction networks to visualise the networks of these genes. Following this, we were able to identify the top 30 hub genes by evaluating each gene by its network connectivity. The results for these analyses are produced in Additional file 7: Figure S2C, D. The top hub gene (ranked by the maximum clique centrality method) for the male-biased CpGs in males was HIST1H4B and the top hub gene for female-biased CpGs was SLC17A7.

Enrichment of saDMPs in transcription factor binding sites

To identify common features among the sex-associated DMPs, we performed transcription factor (TF) binding site and gene ontology analyses. First, we evaluated whether the saDMPs were enriched in motifs for TFs (100 bp window). For the 293 female-biased CpGs, we found 315 unique enriched TFs (*p* value < 0.05) (Fig. 2A and Additional file 3) with strongest evidence for FOXB1, TIA1 and XRCC1. These are genes not previously reported to exhibit any sex differences or be enriched at areas exhibiting any sex differences. We did however find

(See figure on next page.)

Fig. 1 Location and characterisation of saDMPs. **A** Manhattan plot for EWAS analysis of sex. CpG sites which met a threshold of FDR < 0.05 and had an average beta change of > 0.05 and found in both discovery and validation data sets were considered significant and are represented by darker colours. **B** Volcano plot for saDMPs. CpGs which are not significant in both the discovery and validation data sets are represented in grey, replicated saDMPs male-biased CpGs are in orange and replicated saDMPs female-biased CpGs in blue. Grey points displayed beyond the cut off points represent CpG sites which were met the criteria in the discovery data set (FDR < 0.05 and deltaBeta value > than 0.05 in any direction) but were not replicated in the validation data set. **C** Principal component analysis of beta values at the significant saDMPs. Male samples are indicated in orange while female samples are indicated in blue. **D** Number of saDMPs harboured by individual genes. **E** Top panel shows the annotation of all saDMPs (*n* = 396), female-biased CpGs (*n* = 293) and male-biased CpGs (*n* = 103) relative to CpG island regions compared to the autosomal background. Bottom panel shows the $\log_2$ (obs/exp) annotations based on the autosomal background of the different annotations. **F** Top panel shows the overlap of all saDMPs (*n* = 396), female-biased CpGs (*n* = 293) and male-biased CpGs (*n* = 103) with genomic features compared to the autosomal background. Bottom panel shows the $\log_2$ (obs/exp) annotations based on the autosomal background of the different annotations

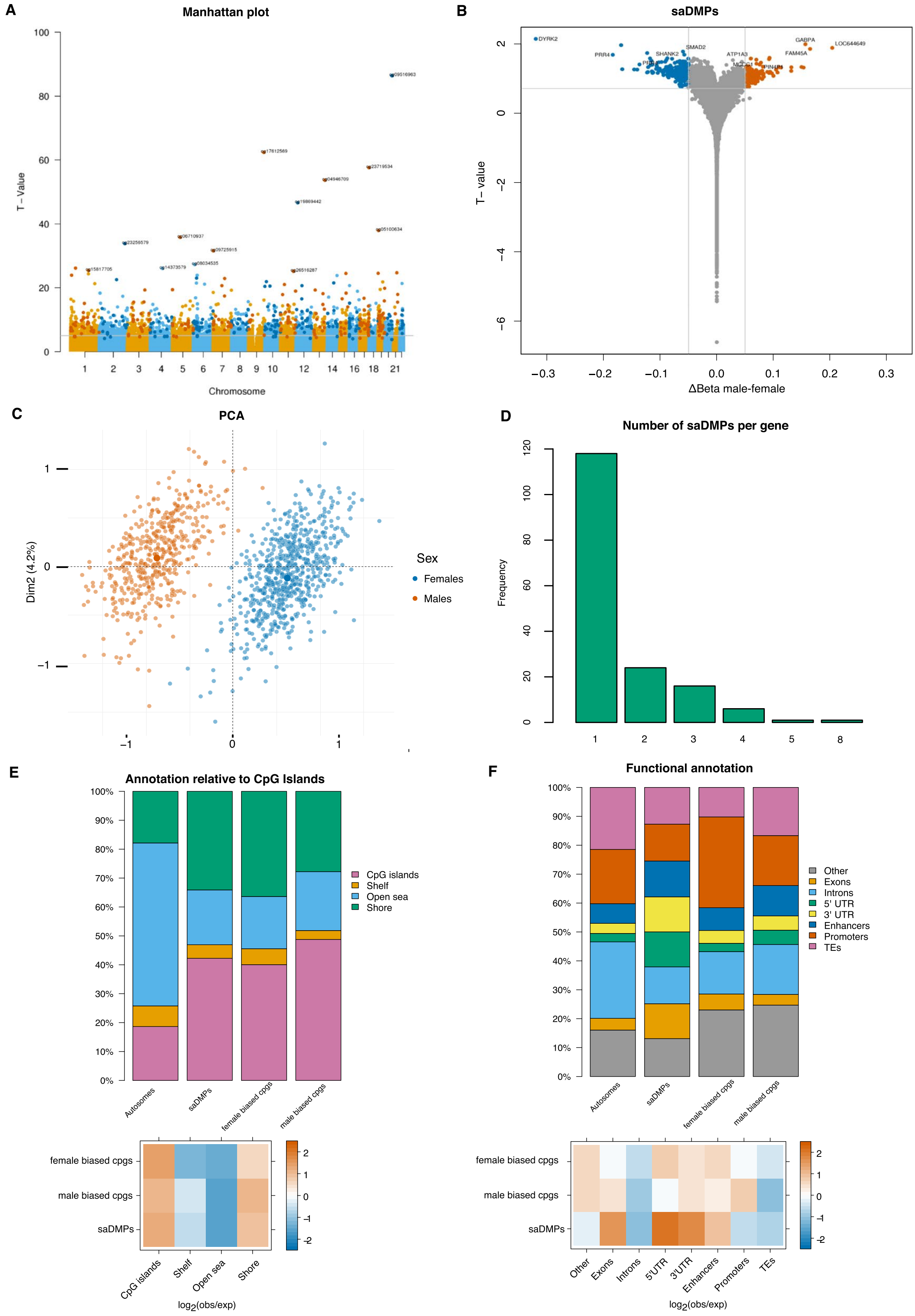


Fig. 1 (See legend on previous page.)

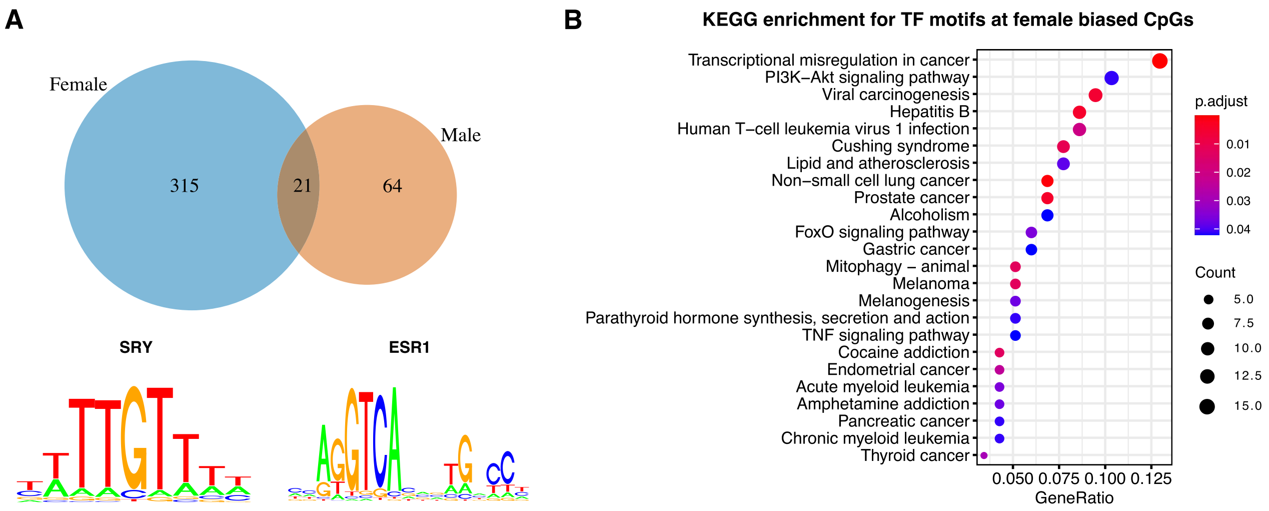


Fig. 2 Transcription factor motif enrichment analysis. **A** Overlap of enriched TF motifs for female-biased CpGs (blue) and male-biased CpGs (orange). saDMPs were enriched in TF binding motifs including SRY and ESR1. **B** KEGG analyses for the significantly enriched TF motifs at female-biased CpGs

some TF motifs enriched which have previously been shown to play a role in sexual development and hormone levels. For example, we found SOX13, SOX21 and SRY TF motif to be enriched in female-biased CpGs, which are known to be involved in male sex determination (Fig. 2A) [2, 48]. For the 103 male-biased CpGs, we identified 64 enriched TFs, including ESR1 which encodes the oestrogen receptor, TCEAL6 HLCS and GPD1 (Fig. 49A and Additional file 4).

To analyse whether the TF motifs were enriched for annotation to biological processes or pathways, we performed pathway analyses using the GO and KEGG databases in order to learn more about how potential sex specific regulatory pathways affected different pathways. We identified several enriched KEGG pathways for the TFBS enriched at female-biased CpGs, spanning a wide range of processes such as transcriptional misregulation in cancer, several specific cancer pathways, PI3K-Akt signalling and more (Fig. 2B). In addition, we also found 39 enriched GO terms ranging from transcription factor activity, E-box binding, transcription coactivator activity and interestingly, bHLH transcription factor binding (Additional file 8: Figure S3B). Nevertheless, we found no enriched KEGG terms for the TFs enriched at male-biased CpGs, likely due to the small number of enriched TFs. However, we identified several enriched GO terms such as NAD, NADP binding and oxidoreductase activity (Additional file 8: Figure S3A).

As we identified enrichment for some transcription factors encoded on the sex chromosomes (e.g. SRY), we hypothesised that sex chromosome encoded transcription factors may influence CpG methylation at the saDMPs directly or indirectly by acting as hub genes in the enriched TF motif network. To assess this, we firstly produced protein–protein interaction networks to visualise the networks of these TFs (Additional file 9: Figure S4). Although we identified some enriched motifs for several TFs encoded on the X chromosomes in the male-biased CpGs such as ELK1, TGIF2LX and TCEAL6 (Additional file 9: Figure S4A), we observed that they were not central nodes in the network. Nevertheless, we did identify several central TF motifs encoded on the sex chromosomes for the female-biased CpGs (Additional file 9: Figure S4B). These included 15 TFs encoded on the X chromosome and 2 on the Y chromosome including SRY and KDM5D.

Secondly, we further utilised cytohubba a plug-in tool in cytoscape to robustly identify if these TFs were in fact hub genes in the network. This revealed that one TF encoded on the X chromosome (RPS4X) did in fact act as a hub gene in the TF network; however, the other 29 genes were encoded on the autosomes (Additional file 9: Figure S4C). Furthermore, for the TF motifs enriched at female-biased CpGs, we identified MAPK1, JUN and BRCA1 and other autosomal genes to be hub TFs in the network revealing novel TFs involved in sex differences (Additional file 9: Figure S4D). Moreover, for those TFs enriched at male-biased CpGs, we identified SP1, ESR1 and SMAD4 to be hub genes in this network (Additional file 9: Figure S4C). Interestingly, SP1 is a gene known to influence SRY expression [49] and ESR is the gene that encodes the oestrogen receptor and lastly, SMAD4, has previously been described as a female germ cell determinant [50]. This analysis suggests that although we did identify some sex chromosome encoded TFs to act as hub genes in the TF network, it is unlikely that they are responsible for affecting CpG methylation at these saDMPs.

Relationship with gene expression

The 396 saDMPs were then further explored in association with the expression levels of their annotated genes using publicly available data for whole blood poly(A) + (GSE120312). The majority of the differentially expressed genes (DEGs) are located on the sex chromosomes, but we also did observe differential expression between males and females for several autosomal genes (Additional file 10: Figure S5B-S5C). We did not identify any significant sex-biased gene expression patterns corresponding to differences in DNAme levels at these genes (Additional file 10: Figure S5). This is not surprising as it has been previously reported that autosomal sex differences in DNA methylation result in nominal or no differences in gene expression [26, 28], a trend also seen with age specific DNAme marks [51]. Moreover, while other studies claim that they identify DEGs on autosomes between males and females, corresponding to differences in DNA methylation, when adjusting for multiple testing, it appears that these no longer hold statistical significance [27]. It is also important to note that, the relationship between DNAme with gene expression is a complex one, although it is generally thought that DNA methylation leads to gene repression, lots of literature reports methylation leading to active expression [52–54] or that it is insufficient to repress transcription [55]. These results support the idea that differences in DNA methylation observed between males and females do not lead to significant differences in gene expression.

Sex-associated differentially methylated regions

Given that several genes harboured numerous saDMPs, we postulated whether some of the saDMPs were part of larger differentially methylated regions associated with sex. We therefore searched for differentially methylated regions associated with sex in our discovery and validation data set. Following adjustment for multiple testing (FDR) and adjustment for cell type proportions, batch effects and age, we identified many sex-associated differentially methylated regions. We therefore considered a sex associated differentially methylated regions (saDMRs) as significant if it harboured at least 5 CpGs, had an FDR value smaller than 0.05, had a methylation difference within the region greater than 0.05 in either direction and was present in both our discovery and validation data set. Following filtering of the list of saDMRs, we identified 266 significant sex-associated DMRs on the autosomes between males and females located at 231 unique sets of genes (Additional file 2). The number of CpGs within the DMRs ranged from 6 to 123 and had an average width of 2392 base pairs (bp) ranging from 178 to 14,715 bp.

Figure 3 shows the beta values for males and females at 4 of the most significant saDMRs: The top hits in the saDMR list overlapped promoter regions of genes such as SDHD, TIMM8B, ATP5J, GABPA, GPN1, CCDC121, AND PRKXP1. SDHD and TIMM8B are genes known to be influenced by oestrogen exposure [56] suggesting that sex hormones may underlie sex differences in autosomal DNA methylation, or alternatively that DNA methylation may mediate sex hormone levels. Moreover, ATP5J and GABPA are genes (male-biased CpGs) which have previously been reported to be implicated in early onset of Alzheimer's disease [57, 58], a disease known to affect females more than males. Furthermore, ATP5J is a gene known to be a target gene of oestrogen, previously shown to serve an inhibitory role in the sex differences in hepatocellular carcinoma [59]. GPN1, CCDC121, ATP5J and GABPA have previously been shown to exhibit functions which are sex specific [21]. Furthermore, PRKXP1 is located on chromosome 15 and CpGs in this region have previously been associated with Crohn's disease and intestinal inflammation, a disease which has previously been reported to be more prevalent in females [60]. A saDMR harbouring 123 CpGs overlapped the promoter region of a gene called MCDC1, a gene known to direct chromosome wide silencing of the sex chromosomes in male germ cells, initiate meiotic sex chromosome inactivation (MSCI), and lead to XY body formation [61].

These findings are extremely important for epigenome wide association studies aiming to characterise sex specific effects in relation to exposures, a rising theme in the literature [6, 42, 62–65]. Our study provides a valuable resource for the community to disentangle whether particular sites or regions display sex differences in DNA methylation.

Discussion

Here, we conducted the first study aiming to characterise autosomal sex differences in DNAme between males and females in whole blood using the Illumina EPIC BeadChip, which interrogates ~ 850,000 sites across the genome. While we were able to identify thousands of autosomal CpGs displaying sex differences in DNA methylation, we focused the majority of our analysis on those autosomal CpGs displaying the largest sex differences (see Methods). We, thereby, identified 396 sex-associated differentially methylated positions on the autosomes. Previous work has reported contradicting results, some research report that there is higher methylation on autosomes in females [5, 19, 21], while other research reports identifying higher methylation on autosomes in males

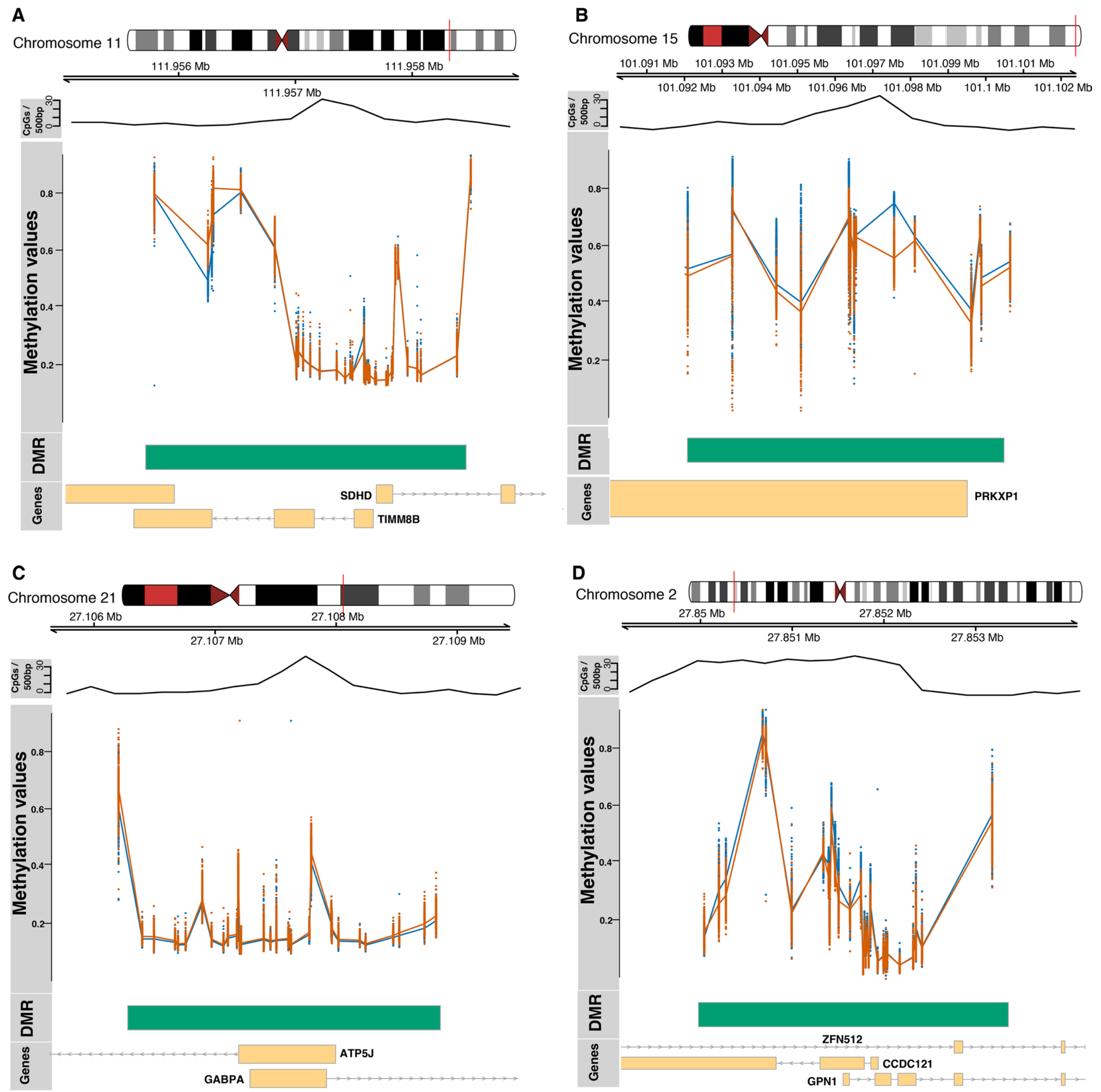


Fig. 3 Plots of sex-associated differentially methylated regions (saDMR). We plotted regions: **A** SDHD and TIMM8B, **B** PRKXP1, **C** ATP5J and GABPA and **D** CCDC121 and GPN1. Yellow boxes represent appropriately labelled genes, green boxes represent the genomic region which the differentially methylated region spans. The scatterplots represent the beta values for males (orange) and females (blue) at CpG sites located within the differentially methylated region

[28–66] and others reports no significant difference in DNAme on autosomes between males and females [68]. Our results support the former and we found that 76% of these loci (293 CpGs) showed higher methylation in females compared to males.

Therefore, although the existence of these autosomal sex-associated CpG sites is well established, a robust and consistent catalogue is yet to emerge. When comparing the saDMPs discovered in this study to findings previously reported in blood samples (where a full list of sex-associated sites were available), we do observe some overlap, although it is limited (Table 2). For example, we identify 54% overlap of our identified saDMPs with previously reported sex-associated sites in cord blood [21].

Table 2 Overlap of autosomal sex-associated differentially methylated positions reported in this study with previous literature in various tissues

Study	Tissue of interest	Sample size (number of individual samples in study)	Platform used for DNA methylation assessment	Number of autosomal probes identified	Percentage of saDMPs replicated (%)	References
Yousefi et al. (2014)	Cord blood	111 newborns	Illumina 450 k	3031	54	[21]
Mccarthy et al. (2014)	Meta analysis of 76 studies	6795	Illumina 27 k	184	0.54	[70]
Inoshita et al. (2015)	Peripheral leukocytes	117 adults	Illumina 450 k	292	15	[24]
Maschietto et al. (2015)	Cord blood	71 newborns	Illumina 450 k	2332	36	[69]
Xu et al. (2014)	Post-mortem pre-frontal cortex	46 adults	Illumina 450 k	614	35	[66]
Hall et al. (2014)	Pancreas	87 adults	Illumina 450 k	470	18	[19]
Xia et al. (2021)	Post-mortem brain samples	1408 adults	Illumina 450 k	15,417	31	[7]
Inkster et al. (2021)	Placenta	293 adults	Illumina 450 k	162	73	[27]

On the other hand, overlap from other research also investigating cord blood was lower, namely 36.87% [69]. Furthermore, we observe only ~ 15% overlap of saDMPs identified in a previous study investigating sex differences in peripheral leukocytes [24]. We also additionally checked the overlap between our saDMPs identified in blood with those identified in other tissues (Table 2). Interestingly, we observe a 73% overlap with sex-associated sites identified in placenta by Inkster et al. [27] and a 35% overlap with sex-associated sites identified in post-mortem prefrontal cortex [66]. The existence of this overlap between different tissues shows that a portion of these saDMPs identified are conserved across tissues, and that some are tissue specific. This highlights an important avenue for future work.

However, alternative reasons for this limited overlap exist; firstly, this may be attributed to differences in sample size, as the datasets used within this study are much larger than previously used thus increasing our ability to detect true sex differences in DNA methylation.

Secondly, the differing normalisation methods applied to DNAme microarray data. Previous research has shown that the methylation levels of CpG sites on the X chromosome differ largely between males and females [29] and, thus, normalisation methods which normalise array data indiscriminately with CpG sites on the autosomes introduce large technical biases for autosomal CpGs [30]. Using normalisation methods, which do not handle the technical bias introduced by sex chromosomes, will therefore lead to many autosomal CpG sites being falsely associated with sex and further, a higher number of autosomal CpGs being incorrectly identified as male-biased CpGs. Our choice of normalisation method greatly reduced technical bias at autosomal CpGs for male and female samples.

Moreover, as thousands of autosomal CpG sites show differences in DNA methylation between males and females, differences in the methods for determining the definition of a sex-associated site result in limited reproducibility between studies, a point also raised by Gatev and colleagues in their identification of sex-associated regions [28]. Here, we therefore proposed and applied stringent cut offs to define a sex-associated site (FDR < 0.05 and effect size of at least 0.05 in either direction). While we acknowledge that true but small differences in DNA methylation related to a phenotype may exist, in the interest of generating a reproducible and robust catalogue of saDMPs, we chose to apply effect size cut offs. Consistent with this, we were able to replicate 75% of our saDMPs identified in our validation data set in our discovery data set. Moreover, we found that 73% of the saDMPs we identified in this study were also identified by Inkster and colleagues [27], whom also applied effect size cut offs, demonstrating the reproducibility and robustness of our catalogue of saDMPs.

We further categorised these 396 saDMPs into two groups, those that were male-biased CpGs ($n = 103$) and those that were female-biased CpGs ($n = 293$). Several saDMPs found to be female-biased CpGs overlapped the transcription start site (TSS) of genes not previously been reported to exhibit sex differences in DNAme including C19orf77, ATP10D and SHANK2. Interestingly, it has previously been shown that sex hormones can regulate SHANK expression leading to a sex differential expression in SHANK2 [71]. Furthermore, this gene has previously been implicated in autism spectrum disorder, a disorder known to exhibit higher prevalence in males rather than females [21]. In contrast, the most significant male-biased CpG is located in the CpG island of a gene located on chromosome 21 called GABPA. GABP is a methylation sensitive transcription factor and has previously been shown to be a transcriptional activator of Cyp 2d-9, which is a gene encoding a male specific steroid in mice [72]. Sex differences in these regions have previously been identified by other studies investigating autosomal sex differences in DNA methylation, specifically Yousefi et al. [73] also identified this region to be a top sex-associated DMR in their analysis. In addition, previous research investigating transcriptome wide sex differences using single cell RNA-seq data in mouse reports GABPA to be one of six TF families responsible for the majority of sex dimorphic transcriptional regulation activities [74].

Interestingly, as well as GABPA being the gene annotated to the most significant saDMP (male-biased CpGs), it was also the third most significant saDMR, suggesting these regions could account for important sex biases observed in some diseases. This is further supported by the fact that GABPA has also been heavily associated with early onset of Alzheimer's disease, Parkinson's disease, breast cancer and autism [71, 73–75].

The saDMR harbouring the highest number of CpG sites ($n = 123$) is located on chromosome 6, overlaps TUBB, MDC1 and XXbac. MDC1 is thought to play a crucial role in the production of male games, lead to XY body formation and also initiate meiotic sex chromosome inactivation. These functions are achieved through its interaction with DNA damage response (DDR) factors, ultimately leading to transcriptional silencing [61, 76].

These results collectively support the hypothesis that sex differences in autosomal DNAme may account for some of the sex differences seen in disease prevalence, onset and progression. Moreover, we did identify saDMPs in genes known to exhibit sex differences in DNAme such as CRIPS2 and DDX43 which are involved in spermatogenesis and male fertility [21, 40], Specifically, CRIPS2 harboured 8 significant saDMPs, all female-biased CpGs, and is part of a group of proteins called CRISPs which show male-biased expression in the male reproductive tract. CRIPS2 plays an important role in spermatogenesis, acrosome reaction and gamete fusion [40]. Some of our saDMPs were located in genes known to show sex by age effects, such as PRR4, a gene associated with dry eye syndrome [77]. Despite this, recent research shows that the adult blood DNA methylome is largely affected by sex, but that these methylome sex differences do not change throughout adulthood and so are largely independent from age effects [78].

The Illumina EPIC array has an increased coverage of the genome, including distal regulatory elements [79]. It was interesting that the 396 saDMPs were still found to be significantly enriched at CpG islands and CpG shores but depleted in open sea regions of the genome (Fig. 1E). The genomic location of DNA methylation normally alters its function. Methylation in CpG islands normally functions to serve long term silencing of genes [80] and CpG island shore methylation is strongly related to gene expression [81], suggesting a potential functional role for these saDMPs. To further support these findings, we identified enrichment of these saDMPs at enhancers, 5'UTRs and promoters (Fig. 1F). Enrichment at 5'UTRs is potentially suggestive that they may be acting as alternative promoters, though we did not test this hypothesis in this study. Despite this enrichment at regulatory regions, we found no correlation of these sites alone with significant differences in gene expression between males and females, suggesting that these saDMPs are not sufficient alone to predict gene expression. Further, this suggests that DNA methylation may potentially be acting as a passive reporter of sex specific transcription. Moreover, it is well established that DNA methylation differences do not always result in differences in gene expression but that these DNA methylation differences are likely to instead be part of larger gene regulatory networks, via acting distally or interacting with transcription factors [52, 53, 82–84]. Despite this, we acknowledge that one caveat of our study was that our DNA methylation data and gene expression data were obtained from different cohorts (they are unmatched) and have large differences in sample size (RNA-seq data have a significantly smaller sample size).

However, this potential link was identified in our TF motif analysis, where we found SRY (sex determining region Y) transcription factor motif, also known as the sex determining factor, to be enriched at female-biased CpGs and further identified this gene to be acting as a hub in the TF network. SRY has been found to bind and repress WNT activation of ovarian genes, and has been shown to bind the promoter regions of many targets of involved in differentiation of the testis [48, 49, 85]. Furthermore, we also found ESR1 transcription factor motif enriched in the saDMPs female-biased CpGs, a gene known to code for the oestrogen receptor.

It has previously been reported that 3D genome organisation can impact sex-biased gene expression through direct and indirect effects of cohesion and CTCF looping on enhancer interactions with sex-biased genes [10]. Recently, it was shown that with rising oestrogen levels, the female brain exhibits sex hormone driven plasticity and that chromatin changes underlie this [86]. Interestingly, by annotating our saDMPs to distal genes using chromatin loops, we were able to identify contacts between saDMPs and three genes HIST1H3A, HIST1H4A and HIST1H4B which are core components of nucleosome, thereby responsible for playing a role in chromatin organisation. Note that the Hi-C data and DNA methylation data were not from matched samples, but two different cohorts. However, these results suggest that although we found DNAme to not be predictive of sex differences in gene expression (Additional file [87]: Figure S5), these saDMPs may interact with other genes, transcription factors and other epigenetic modifications to direct chromatin organisation and regulatory networks.

Lastly, we acknowledge limited overlap with previous studies yet conclude that this is due to our extremely large sample size (discovery, $n = 1171$ and validation, $n = 2471$) and improved handling of sex bias introduced by normalising such data with the sex chromosomes. Both factors contribute to our ability to detect true positives and obtain a more robust catalogue of true sex-associated autosomal CpGs.

Material and methods

Participants

Whole blood Illumina Infinium MethylationEPIC BeadChip DNAme data were collected from participants involved in Understanding Society: The UK Household Longitudinal Study [31]. In wave 3 of the study (2011–12), blood samples were collected from a portion of the study participants. Individuals were considered eligible to give a blood sample if they were over the age of 16, consented to blood sampling and genetic analysis, and participated in all annual interviews between 1999 and 2011 as previously reported in Hughes et al. [88]. Our study population was restricted to participants of white ethnicity. A full description of the dataset and data processing has been described by [88]. Following quality checks of the data, our final data set consisted of 1171 participants (males = 489, females = 686) for discovery and 2471 (males = 1135, females = 1345) participants for validation. The age ranges for each data set were 28–98 years old and 16–99 years old, respectively.

DNA methylation data

Samples of whole blood DNA from participants were obtained following the protocol described in [88]. Raw signal intensities were processed using the R package bigmelon [29] and watermelon [89] from idat files. Prior to normalisation of the data, outlier samples were identified using principal component analysis and subsequently removed from the data set. The reported age of each sample was compared to predicted age using the epigenetic age method implemented by *agep* in the R package bigmelon [89]. Further, the reported sex of the samples was checked using a DNA methylation-based sex classifier [90] which predicts sex based on the methylation difference of X and Y chromosomes. 4 samples were subsequently removed from our discovery data set and 9 samples were removed from our validation data set, as reported and predicted sex did not match. The data were then normalised via the *interpolatedXY* adjusted *dasen* method implemented in the R package, wateRmelon [91]. Following normalisation of the data, SNP probes, cross hybridising probes 27 and X or Y linked probes were removed from the data set. The final discovery and validation data set consisted of 1171 and 2471 samples, respectively, and 747,302 DNA methylation sites.

As whole blood is a heterogenous tissue and contains different cell types, individual samples will have different cell type proportions which may confound analyses. Often, this manifests itself as many false positives being identified. The estimation is based on epigenetic data and expected DNA methylation signatures at specific loci in each cell types are used to estimate cell type composition. To ensure that whole blood cell composition did not differ significantly by sex and would not introduce bias to our results, the relative proportions of Granulocytes, mononuclear, natural killer, CD4T, CD8T and B cells were estimated for all samples using the *estimateCellCounts* function implemented in bigmelon [89]. Furthermore, to assess whether the sex differences we observed were age independent, we performed a Mann–Whitney U test between the age distribution of males and females. Our results confirmed that there is no statistical difference in age between our male and female samples for our discovery data set (p value 0.07; median values of 60 and 58, respectively) and also for our validation data set (p value 0.26; median values of 52 and 51, respectively).

Identifying sex-associated autosomal differential methylation.

Sex-associated autosomal differentially methylated positions (saDMPs) were identified by performing linear modelling using the limma package in R [92] using sex annotation and Beta values while adjusting for age, cell type proportions and batch effects. Correction for multiple testing was performed with the Benjamini–Hochberg false discovery rate method (FDR values). We further used the Bayesian method for controlling p value inflation using the R package *bacon* for both our discovery and validation data sets [93]. A probe was considered significantly differentially methylated if the difference in Beta values between males and females was greater than 0.05 in either direction and the FDR value was smaller than 0.05. We considered a saDMP to be validated if it met these two criteria in both the discovery and validation data set. We further characterised differentially methylated regions (DMRs) by applying the *DMRcate* function from the R package ChAMP to detect DMRs between males and females on the autosomes [94]. A DMR was considered to be significantly associated with sex (saDMR) if it consisted of at least 5 CpG sites with a maximum difference in beta values between males and females greater than 0.05.

Genomic annotation of CpG sites

We annotated the autosomal CpG's using the manufacturer supplied annotation data (MethylationEPIC_v-1-0_B2 manifest file). Annotation was completed in the R package Minfi [95]. Several categories were used as annotations in relation to CpG islands and divided into the following categories: CGIs, CGI shores (S and N), CGI shelfs (S and N) and open sea regions. Further, we also annotated the autosomal CpGs to several genomic features, including exons, introns, 5' UTR, 3'UTR, enhancers, promoters and transposable elements (TEs) using data from UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables).

Gene ontology analyses

GO analyses were conducted using the *gometh* function in the missMethyl package [96] which tests gene ontology enrichment for significant CpGs while accounting for the differing number of probes per gene present on the EPIC array. For GO ontology analyses of enriched TFBS, we used *enrichGO* from the clusterProfiler package in R [97], which performs FDR adjustment .

Enrichment of saDMPs in transcription factor binding motifs and integration with gene expression

The enrichment analysis of known motifs in sex-associated DMPs was performed using the R package PWMEnrich [98] using the MotifDb collection of TF motifs [99]. Specifically, the DNA sequences within a 100 bp range from the saDMP which were female-biased CpGs were extracted from the genome and compared to the saDMPs which were male-biased CpGs as the background to reveal unique enriched motifs (adjusted p value < 0.05). RNA-seq data for 20 healthy donors (10 males and 10 females) from publicly available data from GEO (GSE120312) were used in our analysis. We used the pre-processed count matrices with DESeq2 [100] to calculate differentially expressed genes between males and females with an adjusted p value of 0.05 and $\log_2$ fold change of 1. DESeq2 does apply an automatic filtering step to remove genes with low counts but we did also apply our own independent filtering to this data by removing genes that have counts of at least 10 in all samples.

Overlap of saDMP's with chromatin loops

We examined whether any of the sex-associated DMPs made 3D contacts with distal genes using Hi-C data available from the GEO under accession number (GSE124974) for white blood cells and neutrophils. Hi-C library preparation was performed using the Arima-HiC kit and pre-processing of the data was performed using Juicer command line tools [101]. Reads were aligned to the human (hg38) genome using BWA-mem [97] and then pre-processed using the Juicer pre-processing pipeline. We called chromatin loops using the HICCUPS tool from Juicer using a 10 Kb resolution. We then constructed GenomicInteractions objects to annotate saDMPs to loop anchors using the *findOverlaps* function from the GenomicRanges package using a *maxgap* of 10,000. Following this, we then annotated the corresponding anchor to the relevant gene ID. These steps then allowed us to perform network analysis in Cytoscape [102] and GO and KEGG analyses in clusterProfiler [103].

Protein–protein network visualisation and hub gene identification

We searched all of the genes annotated to our saDMPs using the Search Tool for the Retrieval of Interacting Genes (STRING) (https:://string-db.org) database to generate our networks. We extracted protein–protein interactions with a combined score of > 0.4. Following this, we utilised the cytoscape plugin tool Cytohubba [104] in order to identify hub genes within the networks. This was done by employing the local based method called maximum clique centrality (MCC). The same analysis was applied to the enriched transcription factors found at saDMPs.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13148-022-01279-7.

Additional file 1. Significant sex-associated autosomal DMPs. Illumina Manifest annotations for all 396 significant CpG sites associated with sex on the autosomes.

Additional file 2. Significant sex-associated autosomal DMRs. Test results for all significant DMR's associated with sex on the autosomes ordered by FDR value.

Additional file 3. Enrichment statistics for the TF motifs enriched at female-biased CpGs.

Additional file 4. Enrichment statistics for the TF motifs enriched at male-biased CpGs.

Additional file 5. Genes annotated to saDMPs via Hi-C analysis.

Additional file 6. Figure S1: (A) QQ plot and lambda values (discovery data) distribution of the adjusted p values against the null distribution for EWAS of sex in the understanding society cohort. Genomic inflation lambda score is indicated in the QQ plot to indicate statistical inflation of p values. (B) Boxplots of estimated whole blood cell type proportions for males (orange) and females (blue) in the discovery data set, estimated using the estimateCellCounts function from bigmelon. We performed a Mann-Whitney U test (p value: n.s. 0.05, *p value < 0.05, **< 0.01 and ***< 0.001). (C) QQ plot and lambda values (validation data) distribution of the adjusted p values against the null distribution for EWAS of sex in the understanding society cohort. Genomic inflation lambda score is indicated in the QQ plot to indicate statistical inflation of p values. (D) Boxplots of estimated whole blood cell type proportions for males (orange) and females (blue) in the validation data set, estimated using the estimateCellCounts function from bigmelon. We performed a Mann–Whitney U test (p value: n.s. 0.05, *p value < 0.05, **< 0.01 and ***< 0.001). (E) Venn diagram showing overlap of differentially methylated positions identified in our validation and discovery data set before and after filtering of the list of saDMPs.

Additional file 7. Figure S2: (A) Integrated genomics viewer track of chromatin loop on chromosome 6 showing two male-biased CpGs contacting H1/H4/H3/H2V/H2A. (B) Integrated genomics viewer track of chromatin loop on chromosome 1 showing a female-biased CpG contacting the ODF2L gene. Blue lines represent the chromatin loops, with black lines showing the loop anchors. Orange vertical lines represent the male-biased CpGs and blue vertical lines represent the female-biased CpGs. Purple annotations represent genes. (C-D) Subnetworks of the top 30 genes annotated to male-biased CpGs (C) and females (D). Node colour represents the degree of connectivity. The scale from red to yellow represents the top 30 enriched genes rank from 1-30, with red indicating highest degree and yellow indicating lowest degree.

Additional file 8. Figure S3: GO terms overrepresented for the significantly enriched TF motifs at male-biased CpGs (A) and female-biased CpGs (B).

Additional file 9. Figure S4: (A-B) Network visualisation of protein-protein interactions for all transcription factor motifs found to be enriched at male-biased CpGs (A) and female-biased CpGs (B). Grey coloured boxes represent individual TFs located on autosomes, while purple-coloured boxes represent TFs encoded on the X chromosome and green coloured boxes represent TFs encoded for on the Y chromosome. Grey lines represent edges between transcription factors within the protein-protein network. (C-D) Subnetworks of the top 30 enriched TF motifs at male-biased CpGs (C) and females (D). Node colour represents the degree of connectivity. The scale from red to yellow represents the top 30 enriched TF motif rank from 1-30, with red indicating highest degree and yellow indicating lowest degree.

Additional file 10. Figure S5: Volcano plot showing differential gene expression between males and females. We considered the case of: (A) genes annotated to the saDMPs, (B) sex chromosome linked genes and (C) autosomal genes. Points coloured in grey represent non differentially expressed genes. Green points represent genes which had a log2 Fold Change value greater than 1. Blue points represent genes which met the adjusted p value threshold (FDR <0.05). Points coloured in red represent genes which showed differential expression between males and females (adjusted p value<0.05 & log2FC > 1).

Acknowledgements
The authors acknowledge the use of the High Performance Computing Facility at the University of Essex and would like to thank Stuart Newman for his support. We would also like to thank Tyler Gorrie-Stone, Alexandria Andrayas and Liudmila Mikheeva for their useful discussions on the analysis.

Author contributions
All authors read and approved the final manuscript.

Funding
OAG is funded by the University of Essex. M.K. was supported by the University of Essex and ESRC (grant RES-596-28-0001). L.S. was supported by Medical Research Council grant K013807. N.R.Z. was supported by the Queen Mary University of London. Measurement of DNA methylation in Understanding Society: The UK Household Longitudinal Study was funded through an enhancement to Economic and Social Research Council (ESRC) grant ES/N00812X/1. The analysis was facilitated by access to the Ceres high-performance computing cluster at the University of Essex.

Availability of data and materials
The code to perform this analysis is available on GitHub https://github.com/livygrant97/ASD_DNAme. RNA-seq data used in the differential gene expression analysis is publicly available on GEO (Gene expression Omnibus) under the accession number GSE120312.

Declaration

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Life Sciences, University of Essex, Colchester CO4 3SQ, UK. [2]Institute of Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. [3]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK. [4]School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK.

Received: 29 October 2021 Accepted: 18 April 2022
Published online: 14 May 2022

References
1. Beery AK, Zucker I. Sex bias in neuroscience and biomedical research. Neurosci Biobehav Rev. 2011;35:565–72.
2. Credendino SC, Neumayer C, Cantone I. Genetics and epigenetics of sex bias: insights from human cancer and autoimmunity. Trends Genet. 2020;36:650–63.
3. Hartman RJG, Huisman SE, den Ruijter HM. Sex differences in cardiovascular epigenetics—a systematic review. Biol Sex Differ. 2018;9:19. https://doi.org/10.1186/s13293-018-0180-z.
4. Qin X, Li J, Wu T, Wu Y, Tang X, Gao P, et al. Overall and sex-specific associations between methylation of the ABCG1 and APOE genes and ischemic stroke or other atherosclerosis-related traits in a sibling study of Chinese population. Clin Epigenetics. 2019;11:189. https://doi.org/10.1186/s13148-019-0784-0.
5. Davegårdh C, Hall Wedin E, Broholm C, Henriksen TI, Pedersen M, Pedersen BK, et al. Sex influences DNA methylation and gene expression in human skeletal muscle myoblasts and myotubes. Stem Cell Res Ther. 2019;10:26. https://doi.org/10.1186/s13287-018-1118-4.
6. Koo HK, Morrow J, Kachroo P, Tantisira K, Weiss ST, Hersh CP, et al. Sex-specific associations with DNA methylation in lung tissue demonstrate smoking interactions. Epigenetics. 2020. https://doi.org/10.1080/15592294.2020.1819662.
7. Xia Y, Dai R, Wang K, Jiao C, Zhang C, Xu Y, et al. Sex-differential DNA methylation and associated regulation networks in human brain implicated in the sex-biased risks of psychiatric disorders. Mol Psychiatry. 2021;26:835–48. https://doi.org/10.1038/s41380-019-0416-2.
8. Smith-Bouvier DL, Divekar AA, Sasidhar M, Du S, Tiwari-Woodruff SK, King JK, et al. A role for sex chromosome complement in the female bias in autoimmune disease. J Exp Med. 2008;205:1099–108.
9. Wijchers PJ, Festenstein RJ. Epigenetic regulation of autosomal gene expression by sex chromosomes. Trends Genet. 2011;27:132–40.
10. Werner RJ, Schultz BM, Huhn JM, Jelinek J, Madzo J, Engel N. Sex chromosomes drive gene expression and regulatory dimorphisms in mouse embryonic stem cells. Biol Sex Differ. 2017;8:1–18.
11. Link JC, Chen X, Arnold AP, Reue K. Metabolic impact of sex chromosomes. Adipocyte. 2013;2:74–9. https://doi.org/10.4161/adip.23320.

12. Fish EN. The X-files in immunity: sex-based differences predispose immune responses. Nat Rev Immunol. 2008;8:737–44.
13. Rubtsova K, Marrack P, Rubtsov AV. Sexual dimorphism in autoimmunity. J Clin Investig. 2015;125:2187–93.
14. Andrews S, Yang IJ, Froehlich K, Oskotsky T, Sirota M. Large-scale placenta DNA methylation mega-analysis reveals fetal sex-specific differentially methylated CpG sites and regions. https://doi.org/10.1101/2021.03.04.433985
15. Lopes-Ramos CM, Chen C-Y, Kuijjer ML, Glass K, Quackenbush J, Demeo DL. Sex differences in gene expression and regulatory networks across 29 human tissues. Cell Rep. 2020. https://doi.org/10.1016/j.celrep.2020.107795.
16. Sugathan A, Waxman DJ. Genome-wide analysis of chromatin states reveals distinct mechanisms of sex-dependent gene regulation in male and female mouse liver. Mol Cell Biol. 2013;33:3594–610.
17. Liu J, Morgan M, Hutchison K, Calhoun VD. A study of the influence of sex on genome wide methylation. PLoS ONE. 2010;5:e10028.
18. Numata S, Ye T, Hyde TM, Guitart-Navarro X, Tao R, Wininger M, et al. DNA methylation signatures in development and aging of the human prefrontal cortex. Am J Hum Genet. 2012;90:260–72.
19. Hall E, Volkov P, Dayeh T, Esguerra JL, Salö S, Eliasson L, et al. Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets. Genome Biol. 2014;15:522.
20. Sun L, Lin J, Du H, Hu C, Huang Z, Lv Z, et al. Gender-specific DNA methylome analysis of a Han Chinese longevity population. Biomed Res Int. 2014;2014:1–9.
21. Yousefi P, Huen K, Davé V, Barcellos L, Eskenazi B, Holland N. Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. 2011.
22. Price ME, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin. 2013;6:4.
23. Singmann P, Shem-Tov D, Wahl S, Grallert H, Fiorito G, Shin SY, et al. Characterization of whole-genome autosomal differences of DNA methylation between men and women. Epigenetics Chromatin. 2015;8:1–13.
24. Inoshita M, Numata S, Tajima A, Kinoshita M, Umehara H, Yamamori H, et al. Sex differences of leukocytes DNA methylation adjusted for estimated cellular proportions. Biol Sex Differ. 2015;6:1–7.
25. Martin E, Smeester L, Bommarito PA, Grace MR, Boggess K, Kuban K, et al. Sexual epigenetic dimorphism in the human placenta: implications for susceptibility during the prenatal period. Epigenomics. 2017;9:267–78.
26. Suderman M, Simpkin A, Sharp G, Gaunt T, Lyttleton O, McArdle W, et al. Sex-associated autosomal DNA methylation differences are wide-spread and stable throughout childhood. bioRxiv. 2017. http://europepmc.org/article/PPR/PPR32347
27. Inkster AM, Yuan V, Konwar C, Matthews AM, Brown CJ, Robinson WP. A cross-cohort analysis of autosomal DNA methylation sex differences in the term placenta. Biol Sex Differ. 2021;12(1):1–14.
28. Gatev E, Inkster AM, Negri GL, Konwar C, Lussier AA, Skakkebaek A, et al. Autosomal sex-associated co-methylated regions predict biological sex from DNA methylation. Nucleic Acids Res. 2021. https://doi.org/10.1093/nar/gkab682/6353815.
29. Wang Y, Hannon E, Grant OA, Gorrie-Stone TJ, Kumari M, Mill J, et al. DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy. BMC Genom. 2021;22:1–11. https://doi.org/10.1186/s12864-021-07675-2.
30. Wang Y, Gorrie-Stone TJ, Grant OA, Andrayas AD, Zhai X, McDonald-Maier KD, et al. interpolatedXY: a two-step strategy to normalise DNA methylation microarray data avoiding sex bias. bioRxiv. 2021. https://doi.org/10.1101/2021.09.30.462546v1.
31. G K. Understanding society—UK household longitudinal study: wave 1–5, User Manual. Colchester, United Kingdom. 2015.
32. Windley SP, Wilhelm D. Signaling pathways involved in mammalian sex determination and gonad development. Sex Dev. 2015;9:297–315.
33. Nef S, Vassalli J-D. Complementary pathways in mammalian female sex determination. J Biol. 2009;8:1–3. https://doi.org/10.1186/jbiol173.
34. Jiménez R, Burgos M, Barrionuevo FJ. Sex maintenance in mammals. Genes. 2021;12:999.
35. Ottolenghi C, Pelosi E, Tran J, Colombino M, Douglass E, Nedorezov T, Cao A, Forabosco A, Schlessinger D. Loss of Wnt4 and Foxl2 leads to female-to-male sex reversal extending to germ cells. Hum Mol Genet. 2007;16:2795–804.
36. Rubin JB. The spectrum of sex differences in cancer. Trends Cancer. 2022. http://www.cell.com/article/S2405803322000206/fulltext.
37. Zhu C, Boutros PC. Sex differences in cancer genomes: much learned, more unknown. Endocrinology. 2021;162:bqab170.
38. Lopes-Ramos CM, Quackenbush J, DeMeo DL. Genome-wide sex and gender differences in cancer. Front Oncol. 2020;10:2486.
39. Rubin JB, Lagas JS, Broestl L, Sponagel J, Rockwell N, Rhee G, et al. Sex differences in cancer mechanisms. Biol Sex Differ. 2020;11:1–29.
40. Lim S, Kierzek M, O'Connor AE, Brenker C, Merriner DJ, Okuda H, et al. CRISP2 is a regulator of multiple aspects of sperm function and male fertility. Endocrinology. 2019;160:915–24.
41. Prior HM, Walter MA. Sox genes: architects of development. 1996.
42. Yusipov I, Bacalini MG, Kalyakulina A, Krivonosov M, Pirazzini C, Gensous N, Ravaioli F, Milazzo M, Giuliani C, Vedunova M, Fiorito G. Age-related DNA methylation changes are sex-specific: a comprehensive assessment. Aging. 2020;12:24057–80.
43. Kochmanski J, Kuhn NC, Bernstein AI. Parkinson's disease-associated, sex-specific changes in DNA methylation at PARK7 (DJ-1), ATXN1, SLC17A6, NR4A2, and PTPRN2 in cortical neurons. bioRxiv. 2021. https://doi.org/10.1101/2021.09.08.459434v1.
44. Chathoth KT, Zabet NR. Chromatin architecture reorganization during neuronal cell differentiation in Drosophila genome. Genome Res. 2019;29:613–25.
45. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, et al. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021;593:238–43.
46. Xia X, Zhou X, Quan Y, Hu Y, Xing F, Li Z, et al. Germline deletion of Cdyl causes teratozoospermia and progressive infertility in male mice. Cell Death Dis. 2019;10:1–13.
47. Dimas AS, Nica AC, Montgomery SB, Stranger BE, Raj T, Buil A, et al. Sex-biased genetic effects on gene regulation in humans. Genome Res. 2012;22:2368–75.
48. Li Y, Zheng M, Lau YFC. The sex-determining factors SRY and SOX9 regulate similar target genes and promote testis cord formation during testicular differentiation. Cell Rep. 2014;8:723–33. https://doi.org/10.1016/j.celrep.2014.06.055.
49. Harley VR, Clarkson MJ, Argentaro A. The molecular action and regulation of the testis-determining factors, SRY (sex-determining region on the Y chromosome) and SOX9 [SRY-related high-mobility group (HMG) Box 9]. Endocr Rev. 2003;24:466–87.
50. Wu Q, Fukuda K, Kato Y, Zhou Z, Deng C-X, Saga Y. Sexual fate change of XX germ cells caused by the deletion of SMAD4 and STRA8 independent of somatic sex reprogramming. PLoS Biol. 2016;14:e1002553.
51. Hernando-Herraez I, Evano B, Stubbs T, Commere P-H, Jan Bonder M, Clark S, et al. Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. Nat Commun. 2019;10:1–11.
52. Sadler MC, Auwerx C, Porcu E, Kutalik Z. Quantifying mediation between omics layers and complex traits. bioRxiv. 2021. https://doi.org/10.1101/2021.09.29.462396v1.
53. Rauluseviciute I, Drabløs F, Rye MB. DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. BMC Med Genom. 2020;13:1–15. https://doi.org/10.1186/s12920-020-0657-6.
54. Geybels MS, Zhao S, Wong CJ, Bibikova M, Klotzle B, Wu M, Ostrander EA, Fan JB, Feng Z, Stanford JL. Epigenomic profiling of DNA methylation in paired prostate cancer versus adjacent benign tissue. Prostate. 2015;75:1941–50.
55. Ford E, Grimmer MR, Stolzenburg S, Bogdanovic O, de Mendoza A, Farnham PJ, et al. Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation. bioRxiv. 2017. https://doi.org/10.1101/170506v3.
56. Bove RM, Patrick E, Aubin CM, Srivastava G, Schneider JA, Bennett DA, et al. Reproductive period and epigenetic modifications of the

oxidative phosphorylation pathway in the human prefrontal cortex. PLoS ONE. 2018;13:e0199073.

57. Wiseman FK, Al-Janabi T, Hardy J, Karmiloff-Smith A, Nizetic D, Tybulewicz VLJ, et al. A genetic cause of Alzheimer disease: mechanistic insights from down syndrome. Nat Rev Neurosci. 2015;16:564–74.
58. Kasuga K, Shimohata T, Nishimura A, Shiga A, Mizuguchi T, Tokunaga J, et al. Identification of independent APP locus duplication in Japanese patients with early-onset Alzheimer disease. J Neurol Neurosurg Psychiatry. 2009;80:1050–2.
59. Li Y, Xu A, Jia S, Huang J. Recent advances in the molecular mechanism of sex disparity in hepatocellular carcinoma (review). Oncol Lett. 2019;17:4222–8. https://doi.org/10.3892/ol.2019.10127/abstract.
60. Somineni HK, Venkateswaran S, Kilaru V, Marigorta UM, Mo A, Okou DT, et al. Blood-derived DNA methylation signatures of Crohn's disease and severity of intestinal inflammation. Gastroenterology. 2019;156:2254-2265.e3.
61. Ichijima Y, Ichijima M, Lou Z, Nussenzweig A, Daniel Camerini-Otero R, Chen J, et al. MDC1 directs chromosome-wide silencing of the sex chromosomes in male germ cells. Genes Dev. 2011;25:959–71.
62. Sunny SK, Zhang H, Relton CL, Ring S, Kadalayil L, Mzayek F, et al. Sex-specific longitudinal association of DNA methylation with lung function. ERJ Open Res. 2021;7:00127–2021.
63. Zhang L, Young JI, Gomez L, Silva TC, Schmidt MA, Cai J, et al. Sex-specific DNA methylation differences in Alzheimer's disease pathology. Acta Neuropathol Commun. 2021;9:1–19. https://doi.org/10.1186/s40478-021-01177-8.
64. Curtis SW, Gerkowicz SA, Cobb DO, Kilaru V, Terrell ML, Marder ME, et al. Sex-specific DNA methylation differences in people exposed to polybrominated biphenyl. Epigenomics. 2020;12:757–70.
65. Koo H-K, Morrow J, Kachroo P, Tantisira K, Weiss ST, Hersh CP, et al. Sex-specific associations with DNA methylation in lung tissue demonstrate smoking interactions. Epigenetics. 2021;16:692.
66. Xu H, Wang F, Liu Y, Yu Y, Gelernter J, Zhang H. Sex-biased methylome and transcriptome in human prefrontal cortex. Hum Mol Genet. 2014;23:1260–70.
67. Zhang FF, Cardarelli R, Carroll J, Fulda KG, Kaur M, Gonzalez K, et al. Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. Epigenetics. 2011;6:623–9.
68. García-Calzón S, Perfilyev A, de Mello VD, Pihlajamäki J, Ling C. Sex differences in the methylome and transcriptome of the human liver and circulating HDL-cholesterol levels. J Clin Endocrinol Metab. 2018;103:4395–408.
69. Maschietto M, Bastos LC, Tahira AC, Bastos EP, Euclydes VLV, Brentani A, et al. Sex differences in DNA methylation of the cord blood are related to sex-bias psychiatric diseases. Sci Rep. 2017;7:1–11.
70. McCarthy NS, Melton PE, Cadby G, Yazar S, Franchina M, Moses EK, et al. Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns. BMC Genom. 2014;15:981. https://doi.org/10.1186/1471-2164-15-981.
71. Berkel S, Eltokhi A, Fröhlich H, Porras-Gonzalez D, Rafiullah R, Sprengel R, et al. Sex hormones regulate SHANK expression. Front Mol Neurosci. 2018;11:337.
72. Mottron L, Duret P, Mueller S, Moore RD, Forgeot D'Arc B, Jacquemont S, et al. Sex differences in brain plasticity: a new hypothesis for sex ratio bias in autism understanding the links between sex/gender and autism Dr Meng-Chuan Lai. Mol Autism. 2015. https://doi.org/10.1186/s13229-015-0024-1.
73. Yokomori N, Kobayashi R, Moore R, Sueyoshi T, Negishi M. A DNA methylation site in the male-specific P450 (Cyp 2d–9) promoter and binding of the heteromeric transcription factor GABP. Mol Cell Biol. 1995;15:5355–62.
74. Lu T, Mar JC. Investigating transcriptome-wide sex dimorphism by multi-level analysis of single-cell RNA sequencing data in ten mouse cell types. Biol Sex Differ. 2020;11:1–20.
75. Perdomo-Sabogal A, Nowick K, Piccini I, Sudbrak R, Lehrach H, Yaspo M-L, et al. Human lineage-specific transcriptional regulation through GA-binding protein transcription factor alpha (GABPa). Mol Biol Evol. 2016;33:1231–44.
76. Ahmed EA, van der Vaart A, Barten A, Kal HB, Chen J, Lou Z, Minter-Dykhouse K, Bartkova J, Bartek J, de Boer P, de Rooij DG. Differences in DNA double strand breaks repair in male germ cell types: lessons learned from a differential expression of Mdc1 and 53BP1. DNA Repair. 2017;6:1243–54.
77. Perumal N, Funke S, Pfeiffer N, Grus FH. Proteomics analysis of human tears from aqueous-deficient and evaporative dry eye patients. Sci Rep. 2016;6:1–12.
78. Bergstedt J, Ait S, Azzou K, Tsuo K, Jaquaniello A, Urrutia A, et al. Factors driving DNA methylation variation in human blood. https://doi.org/10.1101/2021.06.23.449602.
79. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17:1–17. https://doi.org/10.1186/s13059-016-1066-1.
80. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. 2012. www.nature.com/reviews/genetics.
81. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2019;41:178–86.
82. Ehrlich M, Lacey M. DNA methylation and differentiation: silencing, upregulation and modulation of gene expression. Epigenomics. 2013;5:553–68.
83. Newell-Price J, Clark AJ, King P. DNA methylation and silencing of gene expression. Trends Endocrinol Metab. 2000;11:142–8.
84. Vaissière T, Sawan C, Herceg Z. Epigenetic interplay between histone modifications and DNA methylation in gene silencing. Mutat Res Rev Mutat Res. 2008;659:40–8.
85. Song Y, Liu T, Wang Y, Deng J, Chen M, Yuan L, et al. Mutation of the Sp1 binding site in the 5′ flanking region of SRY causes sex reversal in rabbits. Oncotarget. 2017;8:38176–83.
86. Matthews BJ, Waxman DJ. Impact of 3D genome organization, guided by cohesin and CTCF looping, on sex-biased chromatin interactions and gene expression in mouse liver. Epigenetics Chromatin. 2020;13:1–25. https://doi.org/10.1186/s13072-020-00350-y.
87. Rocks D, Shukla M, Finnemann SC, Kalluchi A, Jordan Rowley M, Kundakovic M. Sex-specific multi-level 3D genome dynamics in the mouse brain. bioRxiv. 2021. https://doi.org/10.1101/2021.05.03.442383.
88. Hughes A, Smart M, Gorrie-Stone T, Hannon E, Mill J, Bao Y, et al. Socioeconomic position and DNA methylation age acceleration across the life course. Am J Epidemiol. 2018;187:2346–54.
89. Gorrie-Stone TJ, Smart MC, Saffari A, Malki K, Hannon E, Burrage J, et al. Bigmelon: tools for analysing large DNA methylation datasets. Bioinformatics. 2019;35:981–6.
90. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genom. 2013;14:293.
91. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genom. 2013;14:1–10. https://doi.org/10.1186/1471-2164-14-293.
92. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47.
93. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, et al. ChAMP: 450k chip analysis methylation pipeline. Bioinformatics. 2014;30:428–30.
94. van Iterson M, van Zwet EW, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. Genome Biol. 2017;18:1–13.
95. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.
96. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics (Oxford, England). 2016;32:286–8.
97. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.
98. Stojnic R DD. PWMEnrich: PWM enrichment analysis. R package version 4260. 2020.
99. Shannon P, Richards M. MotifDb: an annotated collection of protein-DNA binding sequence motifs. R package version 1340. 2021.

100. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
101. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3:95–8. https://doi.org/10.1016/j.cels.2016.07.002.
102. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
103. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
104. Chin C-H, Chen S-H, Wu H-H, Ho C-W, Ko M-T, Lin C-Y. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol. 2014;8:1–7. https://doi.org/10.1186/1752-0509-8-S4-S11.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Bibliography

[Che,] 33

[Alegría-Torres et al., 2011] Alegría-Torres, J. A., Baccarelli, A., and Bollati, V. (2011). Epigenetics and lifestyle. *Epigenomics*, 3(3):267–277. 17, 44, 50, 59

[Andrews et al., 2022] Andrews, S. V., Yang, I. J., Froehlich, K., Oskotsky, T., and Sirota, M. (2022). Large-scale placenta dna methylation integrated analysis reveals fetal sex-specific differentially methylated cpg sites and regions. *Scientific Reports*, 12(1):1–15. 107

[Arain et al., 2007] Arain, M. A., Blair, R., Finkelstein, N., Brook, J. R., Sahsuvaroglu, T., Beckerman, B., Zhang, L., and Jerrett, M. (2007). The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. *Atmospheric Environment*, 41:3453–3464. 44, 66

[Argeson et al., 1996] Argeson, A. C., Nelson, K. K., and Siracusa, L. D. (1996). Molecular basis of the pleiotropic phenotype of mice carrying the hypervariable yellow (ahvy) mutation at the agouti locus. *Genetics*, 142(2):557–567. 137

[Aryee et al., 2014] Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014). Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30:1363–1369. 36

[Asmar et al., 2015] Asmar, F., Søgaard, A., and Grønbæk, K. (2015). Dna methylation and hydroxymethylation in cancer. 19

[B et al., 2016] B, P., J, M., and A, O. (2016). missmethyl: an r package for analyzing data from illumina's humanmethylation450 platform. *Bioinformatics (Oxford, England)*, 32:286–288. 37

[Bai et al., 2016] Bai, J., Zhang, X., Hu, K., Liu, B., Wang, H., Li, A., Lin, F., Zhang, L., Sun, X., Du, Z., and Song, J. (2016). Silencing dna methyltransferase 1 (dnmt1) inhibits proliferation, metastasis and invasion in escc by suppressing methylation of rassf1a and dapk. *Oncotarget*, 7:44129–44141. 20

[Bains, 2021] Bains, A. (2021). "does air pollution play a role in infertility?: a systematic review"(2017), by julie carré, nicolas gatimel, jessika moreau, jean parinaud and roger léandri. *Embryo Project Encyclopedia*. 47

[Bao et al., 2022] Bao, Y., Gorrie-Stone, T., Hannon, E., Hughes, A., Andrayas, A., Neilson, G., Burrage, J., Mill, J., Schalkwyk, L., and Kumari, M. (2022). Social mobility across the lifecourse and dna methylation age acceleration in adults in the uk. *Scientific Reports*, 12(1):1–12. 30, 108, 139

[Barbosa et al., 2018] Barbosa, M., Joshi, R. S., Garg, P., Martin-Trujillo, A., Patel, N., Jadhav, B., Watson, C. T., Gibson, W., Chetnik, K., Tessereau, C., et al. (2018). Identification of rare de novo epigenetic variations in congenital disorders. *Nature communications*, 9(1):1–11. 98

[Baubec et al., 2015] Baubec, T., Colombo, D. F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A. R., Akalin, A., and Schübeler, D. (2015). Genomic profiling of dna methyltransferases reveals a role for dnmt3b in genic methylation. *Nature*, 520:243–247. 21

[Beery and Zucker, 2011] Beery, A. K. and Zucker, I. (2011). Sex bias in neuroscience and biomedical research. 106

[Bellavia et al., 2013] Bellavia, A., Urch, B., Speck, M., Brook, R. D., Scott, J. A., Albetti, B., Behbod, B., North, M., Valeri, L., Bertazzi, P. A., Silverman, F., Gold, D., and Baccarelli, A. A. (2013). Dna hypomethylation, ambient particulate matter, and increased blood pressure: Findings from controlled human exposure experiments. *Journal of the American Heart Association*, 2. 58

[Bergstedt et al., 2022] Bergstedt, J., Azzou, S. A. K., Tsuo, K., Jaquaniello, A., Urrutia, A., Rotival, M., Lin, D. T., MacIsaac, J. L., Kobor, M. S., Albert, M. L., et al. (2022). The immune factors driving dna methylation variation in human blood. *Nature communications*, 13(1):1–20. 131

[Berkel et al., 2018] Berkel, S., Eltokhi, A., Fröhlich, H., Porras-Gonzalez, D., Rafiullah, R., Sprengel, R., and Rappold, G. A. (2018). Sex hormones regulate shank expression. *Frontiers in Molecular Neuroscience*, 11:337. 130

[Bibikova et al., 2011] Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., Fan, J. B., and Shen, R. (2011). High density dna methylation array with single cpg site resolution. *Genomics*, 98:288–295. 104

[Bibikova et al., 2009] Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K. L. (2009). Genome-wide dna methylation profiling using infinium® assay. *Epigenomics*, 1:177–200. 24

[Bind et al., 2014] Bind, M. A., Lepeule, J., Zanobetti, A., Gasparrini, A., Baccarelli, A., Coull, B. A., Tarantini, L., Vokonas, P. S., Koutrakis, P., and Schwartz, J. (2014). Air pollution and gene-specific methylation in the normative aging study:association, effect modification, and mediation analysis. *Epigenetics*, 9:448–458. 58

[Bird et al., 1985] Bird, A., Taggart, M., Frommer, M., Miller, O. J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. *Cell*, 40:91–99. 19

[Bird, 1987] Bird, A. P. (1987). Cpg islands as gene markers in the vertebrate nucleus. *Trends in Genetics*, 3:342–347. 19

[Bjornsson et al., 2008] Bjornsson, H. T., Sigurdsson, M. I., Fallin, M. D., Irizarry, R. A., Aspelund, T., Cui, H., Yu, W., Rongione, M. A., Ekström, T. J., Harris, T. B., et al. (2008). Intra-individual change over time in dna methylation with familial clustering. *Jama*, 299(24):2877–2883. 27

[Bochtler et al., 2017] Bochtler, M., Kolano, A., and Xu, G.-L. (2017). Dna demethylation pathways: Additional players and regulators. *BioEssays*, 39:e201600178. 23

[Bock et al., 2008] Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2008). Inter-individual variation of dna methylation and its implications for large-scale epigenome mapping. *Nucleic acids research*, 36(10):e55–e55. 136

[Bove et al., 2018] Bove, R. M., Patrick, E., Aubin, C. M., Srivastava, G., Schneider, J. A., Bennett, D. A., Jager, P. L. D., and Chibnik, L. B. (2018). Reproductive period and epigenetic modifications of the oxidative phosphorylation pathway in the human prefrontal cortex. *PLoS ONE*, 13. 124

[Brauer et al., 2008] Brauer, M., Lencar, C., Tamburic, L., Koehoorn, M., Demers, P., and Karr, C. (2008). A cohort study of traffic-related air pollution impacts on birth outcomes. *Environmental Health Perspectives*, 116:680–686. 46, 68, 72

[Breitling et al., 2011] Breitling, L. P., Yang, R., Korn, B., Burwinkel, B., and Brenner, H. (2011). Tobacco-smoking-related differential dna methylation: 27k discovery and replication. *The American Journal of Human Genetics*, 88(4):450–457. 25

[Breton et al., 2016] Breton, C. V., Yao, J., Millstein, J., Gao, L., Siegmund, K. D., Mack, W., Whitfield-Maxwell, L., Lurmann, F., Hodis, H., Avol, E., and Gilliland, F. D. (2016). Prenatal air pollution exposures, dna methyl transferase genotypes, and associations with newborn line1 and alu methylation and childhood blood pressure and carotid intima-media thickness in the children's health study. *Environmental Health Perspectives*, 124:1905–1912. 48, 50, 58

[Burgess et al., 2020] Burgess, S., Foley, C. N., Allara, E., Staley, J. R., and Howson, J. M. (2020). A robust and efficient method for mendelian randomization with hundreds of genetic variants. *Nature communications*, 11(1):1–11. 27

[Burris and Baccarelli, 2014] Burris, H. H. and Baccarelli, A. A. (2014). Environmental epigenetics: from novelty to scientific discipline. *Journal of Applied Toxicology*, 34:113–116. 48

[C et al., 2007] C, O., E, P., J, T., M, C., E, D., T, N., A, C., A, F., and D, S. (2007). Loss of wnt4 and foxl2 leads to female-to-male sex reversal extending to germ cells. *Human molecular genetics*, 16:2795–2804. 109

[Cai et al., 2017] Cai, J., Zhao, Y., Liu, P., Xia, B., Zhu, Q., Wang, X., Song, Q., Kan, H., and Zhang, Y. (2017). Exposure to particulate air pollution during early pregnancy is associated with placental dna methylation. *Science of the Total Environment*, 607-608:1103–1108. 48

[Campagna et al., 2021] Campagna, M. P., Xavier, A., Lechner-Scott, J., Maltby, V., Scott, R. J., Butzkueven, H., Jokubaitis, V. G., and Lea, R. A. (2021). Epigenome-wide association studies: current knowledge, strategies and recommendations. *Clinical epigenetics*, 13(1):1–24. 24

[Cantone et al., 2017] Cantone, L., Iodice, S., Tarantini, L., Albetti, B., Restelli, I., Vigna, L., Bonzini, M., Pesatori, A. C., and Bollati, V. (2017). Particulate matter exposure is associated with inflammatory gene methylation in obese subjects. *Environmental Research*, 152:478–484. 58

[Carlsten et al., 2016] Carlsten, C., Blomberg, A., Pui, M., Sandstrom, T., Wong, S. W., Alexis, N., and Hirota, J. (2016). Diesel exhaust augments allergen-induced lower airway inflammation in allergic individuals: A controlled human exposure study. *Thorax*, 71:35–44. 57

[Carlsten et al., 2011] Carlsten, C., Dybuncio, A., Becker, A., Chan-Yeung, M., and Brauer, M. (2011). Traffic-related air pollution and incident asthma in a high-risk birth cohort. *Occupational and Environmental Medicine*, 68:291–295. 57

[Catoni et al., 2018] Catoni, M., Tsang, J. M., Greco, A. P., and Zabet, N. R. (2018). Dmrcaller: A versatile r/bioconductor package for detection and visualization of differentially methylated regions in cpg and non-cpg contexts. *Nucleic Acids Research*, 46. 18

[Chathoth et al., 2022] Chathoth, K. T., Mikheeva, L. A., Crevel, G., Wolfe, J. C., Hunter, I., Beckett-Doyle, S., Cotterill, S., Dai, H., Harrison, A., and Zabet, N. R. (2022). The role of insulators and transcription in 3d chromatin organization of flies. *Genome research*, 32(4):682–698. 163

[Chathoth and Zabet, 2019] Chathoth, K. T. and Zabet, N. R. (2019). Chromatin architecture reorganization during neuronal cell differentiation in drosophila genome. *Genome Research*, 29:613–625. 115

[Chedin et al., 2002] Chedin, F., Lieber, M. R., and Hsieh, C.-L. (2002). The dna methyltransferase-like protein dnmt3l stimulates de novo methylation by dnmt3a. *Proceedings of the National Academy of Sciences*, 99(26):16916–16921. 21

[Chen and Schwartz, 2008] Chen, J. C. and Schwartz, J. (2008). Metabolic syndrome and inflammatory responses to long-term particulate air pollutants. *Environmental Health Perspectives*, 116:612–617. 43

[Chen et al., 2005] Chen, Z. X., Mann, J. R., Hsieh, C. L., Riggs, A. D., and Chédin, F. (2005). Physical and functional interactions between the human dnmt3l protein and members of the de novo methyltransferase family. *Journal of Cellular Biochemistry*, 95:902–917. 21

[Cho et al., 2014] Cho, H., Lee, K., Hwang, Y., Richardson, P., Bratset, H., Teeters, E., Record, R., Riker, C., and Hahn, E. J. (2014). Outdoor tobacco smoke exposure at the perimeter of a tobacco-free university. *Journal of the Air Waste Management Association*, 64:863–866. 59

[Cho et al., 2017] Cho, Y., Lim, J.-h., Song, M.-K., Jeong, S.-C., Lee, K., Heo, Y., Kim, T. S., and Ryu, J.-C. (2017). Toxicogenomic analysis of the pulmonary toxic effects of hexanal in f344 rat. *Environmental toxicology*, 32(2):382–396. 104

[Clarke et al., 2009] Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore dna sequencing. *Nature Nanotechnology*, 4:265–270. 24

[Clifford et al., 2017] Clifford, R. L., Jones, M. J., MacIsaac, J. L., McEwen, L. M., Goodman, S. J., Mostafavi, S., Kobor, M. S., and Carlsten, C. (2017). Inhalation of diesel exhaust and allergen alters human bronchial epithelium dna methylation. *Journal of Allergy and Clinical Immunology*, 139:112–121. 45, 57, 67

[Collins et al., 2011] Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C. D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Sitch, S., Totterdell, I., Wiltshire, A., and Woodward, S. (2011). Development and evaluation of an earth-system model - hadgem2. *Geoscientific Model Development*, 4:1051–1075. 43

[Conforti et al., 2018] Conforti, A., Mascia, M., Cioffi, G., Angelis, C. D., Coppola, G., Rosa, P. D., Pivonello, R., Alviggi, C., and Placido, G. D. (2018). Air pollution and female fertility: A systematic review of literature. 47

[Costello et al., 2021] Costello, K. R., Leung, A., Trac, C., Lee, M., Basam, M., Pospisilik, J. A., and Schones, D. E. (2021). Sequence features of retrotransposons allow for epigenetic variability. *Elife*, 10. 136

[Credendino et al., 2020] Credendino, S. C., Neumayer, C., and Cantone, I. (2020). Genetics and epigenetics of sex bias: insights from human cancer and autoimmunity. *Trends in Genetics*, 36(9):650–663. 106

[Curtis et al., 2020] Curtis, S. W., Gerkowicz, S. A., Cobb, D. O., Kilaru, V., Terrell, M. L., Marder, M. E., Barr, D. B., Marsit, C. J., Marcus, M., Conneely, K. N., and Smith, A. K. (2020). Sex-specific dna methylation differences in people exposed to polybrominated biphenyl. *Epigenomics*, 12:757–770. 125

[Davegårdh et al., 2019] Davegårdh, C., Hall Wedin, E., Broholm, C., Henriksen, T. I., Pedersen, M., Pedersen, B. K., Scheele, C., and Ling, C. (2019). Sex influences dna methylation and gene expression in human skeletal muscle myoblasts and myotubes. *Stem cell research & therapy*, 10(1):1–17. 106, 125

[de Mendoza et al., 2022] de Mendoza, A., Nguyen, T. V., Ford, E., Poppe, D., Buckberry, S., Pflueger, J., Grimmer, M. R., Stolzenburg, S., Bogdanovic, O., Oshlack, A., et al. (2022). Large-scale manipulation of promoter dna methylation reveals context-specific transcriptional responses and stability. *Genome Biology*, 23(1):1–31. 138, 162, 172

[Derakhshan et al., 2022] Derakhshan, M., Kessler, N. J., Ishida, M., Demetriou, C., Brucato, N., Moore, G. E., Fall, C. H., Chandak, G. R., Ricaut, F.-X., Prentice, A. M., et al. (2022). Tissue-and ethnicity-independent hypervariable dna methylation states show evidence of establishment in the early human embryo. *Nucleic Acids Research*, 50(12):6735–6752. 136, 156, 160, 161

[Díez-Villanueva et al., 2021] Díez-Villanueva, A., Jordà, M., Carreras-Torres, R., Alonso, H., Cordero, D., Guinó, E., Sanjuan, X., Santos, C., Salazar, R., Sanz-Pamplona, R., et al. (2021). Identifying causal models between genetically regulated methylation patterns and gene expression in healthy colon tissue. *Clinical Epigenetics*, 13(1):162. 148

[Dimas et al., 2012] Dimas, A. S., Nica, A. C., Montgomery, S. B., Stranger, B. E., Raj, T., Buil, A., Giger, T., Lappalainen, T., Gutierrez-Arcelus, M., Consortium, M., McCarthy, M. I., and Dermitzakis, E. T. (2012). Sex-biased genetic effects on gene regulation in humans. *Genome Research*, 22:2368–2375. 117

[Dolinoy et al., 2007] Dolinoy, D. C., Das, R., Weidman, J. R., and Jirtle, R. L. (2007). Metastable epialleles, imprinting, and the fetal origins of adult diseases. *Pediatric research*, 61(7):30–37. 137, 155

[Drong et al., 2013] Drong, A. W., Nicholson, G., Hedman, Å. K., Meduri, E., Grundberg, E., Small, K. S., Shin, S.-Y., Bell, J. T., Karpe, F., Soranzo, N., et al. (2013). The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of dna methylation in adipose tissue. *PloS one*, 8(2):e55923. 138

[Du and Pang, 2015] Du, X. and Pang, T. Y. (2015). Is dysregulation of the hpa-axis a core pathophysiology mediating co-morbid depression in neurodegenerative diseases? *Frontiers in Psychiatry*, 6. 60

[Durand et al., 2016] Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., and Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell Systems*, 3:95–98. 38

[D'Antona et al., 2022] D'Antona, S., Castiglioni, I., Porro, D., and Cava, C. (2022). Consequences of exposure to pollutants on respiratory health: From genetic correlations to causal relationships. *Plos one*, 17(11):e0277235. 89, 105

[EA et al., 2007] EA, A., van der Vaart A, A, B., HB, K., J, C., Z, L., K, M.-D., J, B., J, B., de Boer P, and de Rooij DG (2007). Differences in dna double strand breaks repair in male germ cell types: lessons learned from a differential expression of mdc1 and 53bp1. *DNA repair*, 6:1243–1254. 131

[Edwards et al., 2001] Edwards, R. D., Jurvelin, J., Saarela, K., and Jantunen, M. (2001). Voc concentrations measured in personal samples and residential indoor, outdoor and workplace microenvironments in expolis-helsinki, finland. *Atmospheric Environment*, 35:4531–4543. 64

[Egger et al., 2006] Egger, G., Jeong, S., Escobar, S. G., Cortez, C. C., Li, T. W., Saito, Y., Yoo, C. B., Jones, P. A., and Liang, G. (2006). Identification of dnmt1 (dna methyltransferase 1) hypomorphs in somatic knockouts suggests an essential role for dnmt1 in cell survival. *Proceedings of the National Academy of Sciences of the United States of America*, 103:14080–14085. 20

[Ehrlich and Lacey, 2013] Ehrlich, M. and Lacey, M. (2013). Dna methylation and differentiation: silencing, upregulation and modulation of gene expression. *Epigenomics*, 5(5):553–568. 18, 132

[El-Maarri et al., 2009] El-Maarri, O., Kareta, M. S., Mikeska, T., Becker, T., Diaz-Lacava, A., Junen, J., Nüsgen, N., Behne, F., Wienker, T., Waha, A., et al. (2009). A systematic search for dna methyltransferase polymorphisms reveals a rare dnmt3l variant associated with subtelomeric hypomethylation. *Human molecular genetics*, 18(10):1755–1768. 137

[Elliott et al., 2015] Elliott, G., Hong, C., Xing, X., Zhou, X., Li, D., Coarfa, C., Bell, R. J., Maire, C. L., Ligon, K. L., Sigaroudinia, M., et al. (2015). Intermediate dna methylation is a conserved signature of genome regulation. *Nature communications*, 6(1):1–10. 161

[Faria et al., 2019] Faria, C. C., Peixoto, M. S., Carvalho, D. P., and Fortunato, R. S. (2019). The emerging role of estrogens in thyroid redox homeostasis and carcinogenesis. *Oxidative Medicine and Cellular Longevity*, 2019. 48

[Feinberg and Irizarry, 2010] Feinberg, A. P. and Irizarry, R. A. (2010). Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences*, 107(suppl_1):1757–1764. 166

[Feinberg et al., 2010] Feinberg, A. P., Irizarry, R. A., Fradin, D., Aryee, M. J., Murakami, P., Aspelund, T., Eiriksdottir, G., Harris, T. B., Launer, L., Gudnason, V., et al. (2010). Personalized epigenomic signatures that are stable over time and covary with body mass index. *Science translational medicine*, 2(49):49ra67–49ra67. 155

[Fernandes et al., 2018] Fernandes, A. G. O., Souza-Machado, C. D., Pinheiro, G. P., Oliva, S. T. D., Mota, R. C. L., Lima, V. B. D., Cruz, C. S., Chatkin, J. M., and Álvaro A. Cruz (2018). Dual exposure to smoking and household air pollution is associated with an increased risk of severe asthma in adults in brazil. *Clinical and Translational Allergy*, 8. 60

[Finer et al., 2011] Finer, S., Holland, M. L., Nanty, L., and Rakyan, V. K. (2011). The hunt for the epiallele. *Environmental and molecular mutagenesis*, 52(1):1–11. 164, 165

[Fish, 2008] Fish, E. N. (2008). The x-files in immunity: sex-based differences predispose immune responses. *Nature Reviews Immunology*, 8(9):737–744. 107

[Flanagan et al., 2006] Flanagan, J. M., Popendikyte, V., Pozdniakovaite, N., Sobolev, M., Assadzadeh, A., Schumacher, A., Zangeneh, M., Lau, L., Virtanen, C., Wang, S.-C., et al. (2006). Intra-and interindividual epigenetic variation in human germ cells. *The American Journal of Human Genetics*, 79(1):67–84. 137

[Ford et al., 2017] Ford, E., Grimmer, M. R., Stolzenburg, S., Bogdanovic, O., de Mendoza, A., Farnham, P. J., Blancafort, P., and Lister, R. (2017). Frequent lack of repressive capacity of promoter dna methylation identified through genome-wide epigenomic manipulation. *bioRxiv*, page 170506. 123

[Forman and Finch, 2018] Forman, H. J. and Finch, C. E. (2018). A critical review of assays for hazardous components of air pollution. 47

[Francis et al., 2018] Francis, P. T., Costello, H., and Hayes, G. M. (2018). Brains for dementia research: evolution in a longitudinal brain donation cohort to maximize current and future value. *Journal of Alzheimer's Disease*, 66(4):1635–1644. 31

[Freijer and Bloemen, 2000] Freijer, J. I. and Bloemen, H. J. (2000). Modeling relationships between indoor and outdoor air quality. *Journal of the Air and Waste Management Association*, 50:292–300. 44, 66

[Frommer et al., 1992] Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual dna strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89:1827–1831. 24

[Gao et al., 2011] Gao, Q., Steine, E. J., Barrasa, M. I., Hockemeyer, D., Pawlak, M., Fu, D., Reddy, S., Bell, G. W., and Jaenisch, R. (2011). Deletion of the de novo dna methyltransferase dnmt3a promotes lung tumor progression. *Proceedings of the National Academy of Sciences of the United States of America*, 108:18061–18066. 21

[Gao et al., 2019] Gao, X., Colicino, E., Shen, J., Kioumourtzoglou, M. A., Just, A. C., Nwanaji-Enwerem, J. C., Coull, B., Lin, X., Vokonas, P., Zheng, Y., Hou, L., Schwartz, J., and Baccarelli, A. A. (2019). Impacts of air pollution, temperature, and relative humidity on leukocyte distribution: An epigenetic perspective. *Environment International*, 126:395–405. 56

[Gao et al., 2017] Gao, X., Thomsen, H., Zhang, Y., Breitling, L. P., and Brenner, H. (2017). The impact of methylation quantitative trait loci (mqtls) on active smoking-related dna methylation changes. *Clinical epigenetics*, 9(1):1–13. 138

[García-Calzón et al., 2018] García-Calzón, S., Perfilyev, A., de Mello, V. D., Pihlajamäki, J., and Ling, C. (2018). Sex differences in the methylome and transcriptome of the human liver and circulating hdl-cholesterol levels. *The Journal of Clinical Endocrinology & Metabolism*, 103(12):4395–4408. 125

[Garg et al., 2018] Garg, P., Joshi, R. S., Watson, C., and Sharp, A. J. (2018). A survey of inter-individual variation in dna methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. *PLoS genetics*, 14(10):e1007707. 136, 137, 160, 161

[Gatev et al., 2021] Gatev, E., Inkster, A., Negri, G. L., Konwar, C., Lussier, A., Skakkebaek, A., Sokolowski, M., Gravholt, C., Dunn, E., Kobor, M., and Aristizabal, M. (2021). Autosomal sex-associated co-methylated regions predict biological sex from dna methylation. *Nucleic Acids Research*. 107, 122, 125, 129, 133, 170

[Gaunt et al., 2016] Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., McArdle, W. L., Ho, K., et al. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome biology*, 17(1):1–14. 138

[Gervin et al., 2012] Gervin, K., Vigeland, M. D., Mattingsdal, M., Hammerø, M., Nygård, H., Olsen, A. O., Brandt, I., Harris, J. R., Undlien, D. E., and Lyle, R. (2012). Dna methylation and gene expression changes in monozygotic twins discordant for psoriasis: identification of epigenetically dysregulated genes. *PLoS genetics*, 8(1):e1002454. 136

[Gibbs et al., 2010] Gibbs, J. R., Van Der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I. P., Troncoso, J., et al. (2010). Abundant quantitative trait loci exist for dna methylation and gene expression in human brain. *PLoS genetics*, 6(5):e1000952. 138

[Gilraine, 2020] Gilraine, M. (2020). Air filters, pollution and student achievement. 53

[Gondalia et al., 2019] Gondalia, R., Baldassari, A., Holliday, K. M., Justice, A. E., Méndez-Giráldez, R., Stewart, J. D., Liao, D., Yanosky, J. D., Brennan, K. J., Engel, S. M., Jordahl, K. M., Kennedy, E., Ward-Caviness, C. K., Wolf, K., Waldenberger, M., Cyrys, J., Peters, A., Bhatti, P., Horvath, S., Assimes, T. L., Pankow, J. S., Demerath, E. W., Guan, W., Fornage, M., Bressler, J., North,

K. E., Conneely, K. N., Li, Y., Hou, L., Baccarelli, A. A., and Whitsel, E. A. (2019). Methylome-wide association study provides evidence of particulate matter air pollution-associated dna methylation. *Environment International*, 132:104723. 45, 55, 67

[Goodrich et al., 2018] Goodrich, A. J., Volk, H. E., Tancredi, D. J., McConnell, R., Lurmann, F. W., Hansen, R. L., and Schmidt, R. J. (2018). Joint effects of prenatal air pollutant exposure and maternal folic acid supplementation on risk of autism spectrum disorder. *Autism Research*, 11:69–80. 62

[Gorrie-Stone, 2019] Gorrie-Stone, T. J. (2019). Dna methylation: Methods and analyses. 24, 32, 33

[Grant et al., 2022] Grant, O. A., Wang, Y., Kumari, M., Zabet, N. R., and Schalkwyk, L. (2022). Characterising sex differences of autosomal dna methylation in whole blood using the illumina epic array. *Clinical epigenetics*, 14(1):1–16. 162

[Greenberg and Bourc'his, 2019] Greenberg, M. V. and Bourc'his, D. (2019). The diverse roles of dna methylation in mammalian development and disease. *Nature reviews Molecular cell biology*, 20(10):590–607. 21, 22

[Greenblatt and Himes, 2019] Greenblatt, R. E. and Himes, B. E. (2019). Facilitating inclusion of geocoded pollution data into health studies. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2019:553–561. 45, 67, 73

[Gruzieva et al., 2018] Gruzieva, O., Breton, C. V., den Dekker, H. T., Ghantous, A., Just, A. C., Plusquin, M., Ruiz, J. L., Volk, H. E., Baccarelli, A., and Melén, E. (2018). Epigenome-wide meta-analysis of dna methylation in children related to prenatal particulate air pollution exposure. *ISEE Conference Abstracts*, 2017:172. 49

[Gruzieva et al., 2019] Gruzieva, O., Xu, C. J., Yousefi, P., Relton, C., Merid, S. K., Breton, C. V., Gao, L., Volk, H. E., Feinberg, J. I., Ladd-Acosta, C., Bakulski, K., Auffray, C., Lemonnier, N., Plusquin, M., Ghantous, A., Herceg, Z., Nawrot, T. S., Pizzi, C., Richiardi, L., Rusconi, F., Vineis, P., Kogevinas, M., Felix, J. F., Duijts, L., Dekker, H. T. D., Jaddoe, V. W., Ruiz, J. L., Bustamante, M., Antó, J. M., Sunyer, J., Vrijheid, M., Gutzkow, K. B., Grazuleviciene, R., Hernandez-Ferrer, C., Annesi-Maesano, I., Lepeule, J., Bousquet, J., Bergström, A., Kull, I., Söderhäll, C., Kere, J., Gehring, U., Brunekreef, B., Just, A. C., Wright, R. J., Peng, C., Gold, D. R., Kloog, I., Demeo, D. L., Pershagen, G., Koppelman, G. H., London, S. J., Baccarelli, A. A., and Melén, E. (2019). Prenatal particulate air pollution and dna methylation in newborns: An epigenome-wide meta-analysis. *Environmental Health Perspectives*, 127. 46, 48, 51

[Gulliver and Briggs, 2011] Gulliver, J. and Briggs, D. (2011). Stems-air: A simple gis-based air pollution dispersion model for city-wide exposure assessment. *Science of the Total Environment*, 409:2419–2429. 69

[Habermann et al., 2015] Habermann, M., Billger, M., and Haeger-Eugensson, M. (2015). Exposure to air pollution traffic-related in the urban area of gothenburg, sweden. *ISEE Conference Abstracts*, 2015:806. 64

[Hachiya et al., 2017] Hachiya, T., Furukawa, R., Shiwa, Y., Ohmomo, H., Ono, K., Katsuoka, F., Nagasaki, M., Yasuda, J., Fuse, N., Kinoshita, K., et al. (2017). Genome-wide identification of inter-individually variable dna methylation sites improves the efficacy of epigenetic association studies. *NPJ genomic medicine*, 2(1):1–14. 136, 160

[Hall et al., 2014] Hall, E., Volkov, P., Dayeh, T., Esguerra, J. L. S., Salö, S., Eliasson, L., Rönn, T., Bacos, K., and Ling, C. (2014). Sex differences in the genome-wide dna methylation pattern and impact on gene expression, microrna levels and insulin secretion in human pancreatic islets. *Genome biology*, 15:522. 107, 125

[Hannon et al., 2018] Hannon, E., Gorrie-Stone, T. J., Smart, M. C., Burrage, J., Hughes, A., Bao, Y., Kumari, M., Schalkwyk, L. C., and Mill, J. (2018).

Leveraging dna-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *The American Journal of Human Genetics*, 103(5):654–665. 39, 138, 163

[Hannon et al., 2016] Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., Troakes, C., Turecki, G., O'donovan, M. C., Schalkwyk, L. C., et al. (2016). Methylation qtls in the developing brain and their enrichment in schizophrenia risk loci. *Nature neuroscience*, 19(1):48–54. 138

[Hannum et al., 2013] Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S. V., Klotzle, B., Bibikova, M., Fan, J. B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., and Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49:359–367. 62

[Harley et al., 2003a] Harley, V. R., Clarkson, M. J., and Argentaro, A. (2003a). The molecular action and regulation of the testis-determining factors, sry (sex-determining region on the y chromosome) and sox9 [sry-related high-mobility group (hmg) box 9]. *Endocrine Reviews*, 24:466–487. 118

[Harley et al., 2003b] Harley, V. R., Clarkson, M. J., and Argentaro, A. (2003b). The molecular action and regulation of the testis-determining factors, sry (sex-determining region on the y chromosome) and sox9 [sry-related high-mobility group (hmg) box 9]. *Endocrine reviews*, 24(4):466–487. 122, 132

[Harris et al., 2013] Harris, R. A., Nagy-Szakal, D., and Kellermayer, R. (2013). Human metastable epiallele candidates link to common disorders. *Epigenetics*, 8(2):157–163. 164

[Harrod et al., 2003] Harrod, K. S., Jaramillo, R. J., Rosenberger, C. L., Wang, S. Z., Berger, J. A., McDonald, J. D., and Reed, M. D. (2003). Increased susceptibility to rsv infection by exposure to inhaled diesel engine emissions. *American Journal of Respiratory Cell and Molecular Biology*, 28:451–463. 57

[Hartigan and Hartigan, 1985] Hartigan, J. A. and Hartigan, P. M. (1985). The dip test of unimodality. *The annals of Statistics*, pages 70–84. 40

[Hartman et al., 2018] Hartman, R. J. G., Huisman, S. E., and den Ruijter, H. M. (2018). Sex differences in cardiovascular epigenetics—a systematic review. *Biology of Sex Differences*, 9:19. 106

[He et al., 2022] He, M., Wu, G., Wang, Z., Ren, K., Yang, Z., and Xue, Q. (2022). Development and validation of a trp-related gene signature for overall survival prediction in lung adenocarcinoma. *Frontiers in Genetics*, 13. 104

[Héberlé and Bardet, 2019] Héberlé, É. and Bardet, A. F. (2019). Sensitivity of transcription factors to dna methylation. *Essays in biochemistry*, 63(6):727–741. 162

[Henschel et al., 2012] Henschel, S., Atkinson, R., Zeka, A., Le Tertre, A., Analitis, A., Katsouyanni, K., Chanel, O., Pascal, M., Forsberg, B., Medina, S., et al. (2012). Air pollution interventions and their impact on public health. *International journal of public health*, 57(5):757–768. 75

[Hernando-Herraez et al., 2019] Hernando-Herraez, I., Evano, B., Stubbs, T., Commere, P.-H., Jan Bonder, M., Clark, S., Andrews, S., Tajbakhsh, S., and Reik, W. (2019). Ageing affects dna methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. *Nature communications*, 10(1):1–11. 122

[Hew et al., 2015] Hew, K. M., Walker, A. I., Kohli, A., Garcia, M., Syed, A., Mcdonald-Hyman, C., Noth, E. M., Mann, J. K., Pratt, B., Balmes, J., Hammond, S. K., Eisen, E. A., and Nadeau, K. C. (2015). Childhood exposure to ambient polycyclic aromatic hydrocarbons is linked to epigenetic modifications and impaired systemic immunity in t cells. *Clinical and Experimental Allergy*, 45:238–248. 53

[Hijmans et al., 2016] Hijmans, R., Williams, E., and Vennes, C. (2016). Geosphere: spherical trigonometry. r package. 76

[Holgate, 2017] Holgate, S. T. (2017). Every breath we take: The lifelong impact of air pollution' - a call for action. 42

[Horvath, 2013] Horvath, S. (2013). Dna methylation age of human tissues and cell types. *Genome Biology*, 14. 56, 62

[Hotchkiss, 1948] Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *Journal of Biological Chemistry*, 175:315–32. 18

[Houseman et al., 2012] Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. (2012). Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13:1–16. 33

[Hughes et al., 2018] Hughes, A., Smart, M., Gorrie-Stone, T., Hannon, E., Mill, J., Bao, Y., Burrage, J., Schalkwyk, L., and Kumari, M. (2018). Socioeconomic position and dna methylation age acceleration across the life course. *American Journal of Epidemiology*, 187:2346–2354. 31, 42

[I et al., 2020] I, Y., MG, B., A, K., M, K., C, P., N, G., F, R., M, M., C, G., M, V., G, F., A, G., S, P., P, G., M, I., and C, F. (2020). Age-related dna methylation changes are sex-specific: a comprehensive assessment. *Aging*, 12:24057–24080. 113, 125

[Ichijima et al., 2011] Ichijima, Y., Ichijima, M., Lou, Z., Nussenzweig, A., Camerini-Otero, R. D., Chen, J., Andreassen, P. R., and Namekawa, S. H. (2011). Mdc1 directs chromosome-wide silencing of the sex chromosomes in male germ cells. *Genes and Development*, 25:959–971. 125, 131

[Inkster et al., 2021a] Inkster, A. M., Yuan, V., Konwar, C., Matthews, A. M., Brown, C. J., and Robinson, W. P. (2021a). A cross-cohort analysis of autosomal dna methylation sex differences in the term placenta. *Biology of sex Differences*, 12(1):1–14. 122

[Inkster et al., 2021b] Inkster, A. M., Yuan, V., Konwar, C., Matthews, A. M., Brown, C. J., and Robinson, W. P. (2021b). A cross-cohort analysis of autosomal dna methylation sex differences in the term placenta. *Biology of sex Differences*, 12(1):1–14. 127, 129

[Inoshita et al., 2015] Inoshita, M., Numata, S., Tajima, A., Kinoshita, M., Umehara, H., Yamamori, H., Hashimoto, R., Imoto, I., and Ohmori, T. (2015). Sex differences of leukocytes dna methylation adjusted for estimated cellular proportions. *Biology of Sex Differences*, 6. 107, 127

[Irizarry et al., 2009] Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J. B., Sabunciyan, S., and Feinberg, A. P. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific cpg island shores. *Nature Genetics*, 41:178–186. 104, 131

[Ito et al., 2010] Ito, S., Dalessio, A. C., Taranova, O. V., Hong, K., Sowers, L. C., and Zhang, Y. (2010). Role of tet proteins in 5mc to 5hmc conversion, es-cell self-renewal and inner cell mass specification. *Nature*, 466:1129–1133. 22

[J et al., 2000] J, N.-P., AJ, C., and P, K. (2000). Dna methylation and silencing of gene expression. *Trends in endocrinology and metabolism: TEM*, 11:142–148. 132

[Jackson et al., 2018] Jackson, V. E., Latourelle, J. C., Wain, L. V., Smith, A. V., Grove, M. L., Bartz, T. M., Obeidat, M., Province, M. A., Gao, W., Qaiser, B., et al. (2018). Meta-analysis of exome array data identifies six novel genetic loci for lung function. *Wellcome open research*, 3. 83

[Jaenisch and Bird, 2003] Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33(3):245–254. 17

[Jafarabadi, 2007] Jafarabadi, M. (2007). Episodic air pollution is associated with increased dna fragmentation in human sperm without other changes in semen quality. 47

[Jaffe and Irizarry, 2014] Jaffe, A. E. and Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15. 19

[Jiménez et al., 2021] Jiménez, R., Burgos, M., and Barrionuevo, F. J. (2021). Sex maintenance in mammals. *Genes*, 12(7):999. 109

[Jin et al., 2011] Jin, B., Li, Y., and Robertson, K. D. (2011). Dna methylation: superior or subordinate in the epigenetic hierarchy? *Genes & cancer*, 2(6):607–617. 21

[Jin and Liu, 2018] Jin, Z. and Liu, Y. (2018). Dna methylation in human diseases. *Genes & diseases*, 5(1):1–8. 22

[Joehanes et al., 2016] Joehanes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R., Guan, W., Xu, T., Elks, C. E., Aslibekyan, S., et al. (2016). Epigenetic signatures of cigarette smoking. *Circulation: cardiovascular genetics*, 9(5):436–447. 59

[Jones, 2012] Jones, P. A. (2012). Functions of dna methylation: islands, start sites, gene bodies and beyond. 131

[Kaminsky et al., 2009] Kaminsky, Z. A., Tang, T., Wang, S.-C., Ptak, C., Oh, G. H., Wong, A. H., Feldcamp, L. A., Virtanen, C., Halfvarson, J., Tysk, C., et al. (2009). Dna methylation profiles in monozygotic and dizygotic twins. *Nature genetics*, 41(2):240–245. 138

[Kasuga et al., 2009] Kasuga, K., Shimohata, T., Nishimura, A., Shiga, A., Mizuguchi, T., Tokunaga, J., Ohno, T., Miyashita, A., Kuwano, R., Matsumoto, N., Onodera, O., Nishizawa, M., and Ikeuchi, T. (2009). Identification of independent app locus duplication in japanese patients with early-onset alzheimer disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 80:1050–1052. 124

[Kelly, 2003] Kelly, F. J. (2003). Oxidative stress: Its role in air pollution and adverse health effects. 49

[Kelly and Fussell, 2015] Kelly, F. J. and Fussell, J. C. (2015). Air pollution and public health: emerging hazards and improved understanding of risk. *Environmental Geochemistry and Health*, 37:631–649. 42

[Kessler et al., 2018] Kessler, N. J., Waterland, R. A., Prentice, A. M., and Silver, M. J. (2018). Establishment of environmentally sensitive dna methylation states in the very early human embryo. *Science advances*, 4(7):eaat2624. 138, 164, 165, 166

[Kim et al., 2019] Kim, H. J., Min, J. Y., Seo, Y. S., and Min, K. B. (2019). Association of ambient air pollution with increased liver enzymes in korean adults. *International Journal of Environmental Research and Public Health*, 16. 61

[Kim et al., 2009] Kim, S. Y., Sheppard, L., and Kim, H. (2009). Health effects of long-term air pollution: Influence of exposure prediction methods. *Epidemiology*, 20:442–450. 72, 99

[Kingsley et al., 2016] Kingsley, S. L., Eliot, M. N., Whitsel, E. A., Huang, Y. T., Kelsey, K. T., Marsit, C. J., and Wellenius, G. A. (2016). Maternal residential proximity to major roadways, birth weight, and placental dna methylation. *Environment International*, 92-93:43–49. 48

[Kobayashi et al., 2015] Kobayashi, M., Nagashio, R., Jiang, S.-X., Saito, K., Tsuchiya, B., Ryuge, S., Katono, K., Nakashima, H., Fukuda, E., Goshima, N., et al. (2015). Calnexin is a novel sero-diagnostic marker for lung cancer. *Lung Cancer*, 90(2):342–345. 94

[Kochmanski et al., 2021] Kochmanski, J., Kuhn, N. C., and Bernstein, A. I. (2021). Parkinson's disease-associated, sex-specific changes in dna methylation at park7 (dj-1), atxn1, slc17a6, nr4a2, and ptprn2 in cortical neurons. *bioRxiv*, page 2021.09.08.459434. 113

[Kodavanti, 2016] Kodavanti, U. P. (2016). Stretching the stress boundary: Linking air pollution health effects to a neurohormonal stress response. *Biochimica et Biophysica Acta - General Subjects*, 1860:2880–2890. 60

[Koo et al., 2020] Koo, H. K., Morrow, J., Kachroo, P., Tantisira, K., Weiss, S. T., Hersh, C. P., Silverman, E. K., and DeMeo, D. L. (2020). Sex-specific associations with dna methylation in lung tissue demonstrate smoking interactions. *Epigenetics*, pages 1–12. 106, 125

[Krause et al., 2020] Krause, C., Geißler, C., Tackenberg, H., El Gammal, A. T., Wolter, S., Spranger, J., Mann, O., Lehnert, H., and Kirchner, H. (2020). Multilayered epigenetic regulation of irs2 expression in the liver of obese individuals with type 2 diabetes. *Diabetologia*, 63(10):2182–2193. 162

[Kulakova et al., 2016] Kulakova, O., Kabilov, M., Danilova, L., Popova, E., Baturina, O., Tsareva, E. Y., Baulina, N., Kiselev, I., Boyko, A., Favorov, A., et al. (2016). Whole-genome dna methylation analysis of peripheral blood mononuclear cells in multiple sclerosis patients with different disease courses. *Acta Naturae ( )*, 8(3 (30)):103–110. 27

[Künzli et al., 2009] Künzli, N., Bridevaux, P.-O., Liu, L. S., Garcia-Esteban, R., Schindler, C., Gerbase, M., Sunyer, J., Keidel, D., and Rochat, T. (2009). Traffic-related air pollution correlates with adult-onset asthma among never-smokers. *Thorax*, 64(8):664–670. 75

[Kurdyukov and Bullock, 2016] Kurdyukov, S. and Bullock, M. (2016). Dna methylation analysis: Choosing the right method. *Biology*, 5. 24

[Lam et al., 2012] Lam, L. L., Emberly, E., Fraser, H. B., Neumann, S. M., Chen, E., Miller, G. E., and Kobor, M. S. (2012). Factors underlying variable dna methylation in a human community cohort. *Proceedings of the National Academy of Sciences*, 109(supplement_2):17253–17260. 160

[Lamadema et al., 2019] Lamadema, N., Burr, S., and Brewer, A. C. (2019). Dynamic regulation of epigenetic demethylation by oxygen availability and cellular redox. *Free Radical Biology and Medicine*, 131:282–298. 18

[Lee et al., 2019] Lee, M. K., Xu, C. J., Carnes, M. U., Nichols, C. E., Ward, J. M., Kwon, S. O., Kim, S. Y., Kim, W. J., and London, S. J. (2019). Genome-wide dna methylation and long-term ambient air pollution exposure in korean adults. *Clinical Epigenetics*, 11:37. 140

[Legro et al., 2009] Legro, R., Sauer, M., Mottla, G., Richter, K., Dodson, W., and Liao, D. (2009). Effect of air quality on assisted human reproduction. *Fertility and Sterility*, 92:S42–S43. 47

[Li et al., 2018] Li, H., Chen, R., Cai, J., Cui, X., Huang, N., and Kan, H. (2018). Short-term exposure to fine particulate air pollution and genome-wide dna methylation: A randomized, double-blind, crossover trial. *Environment International*, 120:130–136. 68

[Li and Durbin, 2010] Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26:589–595. 38

[Li et al., 2014a] Li, Y., Zheng, M., and Lau, Y. F. C. (2014a). The sex-determining factors sry and sox9 regulate similar target genes and promote testis cord formation during testicular differentiation. *Cell Reports*, 8:723–733. 71, 72

[Li et al., 2014b] Li, Y., Zheng, M., and Lau, Y.-F. C. (2014b). The sex-determining factors sry and sox9 regulate similar target genes and promote testis cord formation during testicular differentiation. *Cell reports*, 8(3):723–733. 118, 132

[Liao et al., 2015] Liao, J., Karnik, R., Gu, H., Ziller, M. J., Clement, K., Tsankov, A. M., Akopian, V., Gifford, C. A., Donaghey, J., Galonska, C., Pop, R., Reyon, D., Tsai, S. Q., Mallard, W., Joung, J. K., Rinn, J. L., Gnirke, A., and Meissner, A. (2015). Targeted disruption of dnmt1, dnmt3a and dnmt3b in human embryonic stem cells. *Nature Genetics*, 47:469–478. 20

[Lim et al., 2019] Lim, S., Kierzek, M., O'Connor, A. E., Brenker, C., Merriner, D. J., Okuda, H., Volpert, M., Gaikwad, A., Bianco, D., Potter, D., Prabhakar, R., Strünker, T., and O'Bryan, M. K. (2019). Crisp2 is a regulator of multiple aspects of sperm function and male fertility. *Endocrinology*, 160:915–924. 113, 131

[Lim et al., 2021] Lim, Y., Arora, S., Schuster, S. L., Corey, L., Fitzgibbon, M., Wladyka, C. L., Wu, X., Coleman, I. M., Delrow, J. J., Corey, E., et al. (2021). Multiplexed functional genomic analysis of 5'untranslated region mutations across the spectrum of prostate cancer. *Nature Communications*, 12(1):1–18. 166

[Link et al., 2013] Link, J. C., Chen, X., Arnold, A. P., and Reue, K. (2013). Metabolic impact of sex chromosomes. *Adipocyte*, 2:74–79. 106

[Liu et al., 2010] Liu, J., Morgan, M., Hutchison, K., and Calhoun, V. D. (2010). A study of the influence of sex on genome wide methylation. *PLoS ONE*, 5. 107

[Liu et al., 2012] Liu, P., Morrison, C., Wang, L., Xiong, D., Vedell, P., Cui, P., Hua, X., Ding, F., Lu, Y., James, M., et al. (2012). Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis*, 33(7):1270–1276. 53, 88

[Liu et al., 2019] Liu, Y., Liu, T.-Y., Weinberg, D. E., White, B. W., Chris, J., Tan, C. L., Schmitt, A. D., Selvaraj, S., Tran, V., Laurent, L. C., et al. (2019).

Spatial co-fragmentation pattern of cell-free dna recapitulates in vivo chromatin organization and identifies tissues-of-origin. *BioRxiv*, page 564773. 32

[Liutkeviačiūte et al., 2011] Liutkeviačiūte, Z., Kriukiene, E., Grigaityte, I., Masevieius, V., and Klimašauskas, S. (2011). Methyltransferase-directed derivatization of 5-hydroxymethylcytosine in dna. *Angewandte Chemie - International Edition*, 50:2090–2093. 23

[Lodovici and Bigagli, 2011] Lodovici, M. and Bigagli, E. (2011). Oxidative stress and air pollution exposure. 49

[Lopes-Ramos et al., 2020a] Lopes-Ramos, C. M., Chen, C.-Y., Kuijjer, M. L., Paulson, J. N., Sonawane, A. R., Fagny, M., Platig, J., Glass, K., Quackenbush, J., and DeMeo, D. L. (2020a). Sex differences in gene expression and regulatory networks across 29 human tissues. *Cell reports*, 31(12):107795. 107, 111

[Lopes-Ramos et al., 2020b] Lopes-Ramos, C. M., Quackenbush, J., and DeMeo, D. L. (2020b). Genome-wide sex and gender differences in cancer. *Frontiers in oncology*, 10:597788. 111

[Love et al., 2014] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550. 37, 38

[Lovinsky-Desir et al., 2018] Lovinsky-Desir, S., Lawrence, J., Jung, K. H., Rundle, A. G., Hoepner, L. A., Yan, B., Perera, F., Perzanowski, M. S., Miller, R. L., and Chillrud, S. N. (2018). Assessment of exposure to air pollution in children: Determining whether wearing a personal monitor affects physical activity. *Environmental research*, 166:340–343. 52

[Lu and Mar, 2020] Lu, T. and Mar, J. C. (2020). Investigating transcriptome-wide sex dimorphism by multi-level analysis of single-cell rna sequencing data in ten mouse cell types. *Biology of sex Differences*, 11(1):1–20. 68, 130

[Lucock et al., 2017] Lucock, M., Jones, P., Veysey, M., and Beckett, E. (2017). B vitamins and pollution, an interesting, emerging, yet incomplete picture of folate and the exposome. *Proceedings of the National Academy of Sciences of the United States of America*, 114:E3878–E3879. 62

[Luo et al., 2021] Luo, X., Zhang, T., Zhai, Y., Wang, F., Zhang, S., and Wang, G. (2021). Effects of dna methylation on tfs in human embryonic stem cells. *Frontiers in genetics*, 12:639461. 162

[Madrigano et al., 2012] Madrigano, J., Baccarelli, A. A., Mittleman, M. A., Sparrow, D., Vokonas, P. S., Tarantini, L., and Schwartz, J. (2012). Aging and epigenetics: longitudinal changes in gene-specific dna methylation. *Epigenetics*, 7(1):63–70. 27

[Maghbooli et al., 2018] Maghbooli, Z., Hossein-nezhad, A., Adabi, E., Asadollahpour, E., Sadeghi, M., Mohammad-nabi, S., Rad, L. Z., Hosseini, A. A. M., Radmehr, M., Faghihi, F., Aghaei, A., Omidifar, A., Aghababei, Y., and Behzadi, H. (2018). Air pollution during pregnancy and placental adaptation in the levels of global dna methylation. *PLoS ONE*, 13. 46, 47

[Mahalingaiah et al., 2016] Mahalingaiah, S., Hart, J. E., Laden, F., Farland, L. V., Hewlett, M. M., Chavarro, J., Aschengrau, A., and Missmer, S. A. (2016). Adult air pollution exposure and risk of infertility in the nurses' health study ii. *Human Reproduction*, 31:638–647. 47

[Marques et al., 2015] Marques, C. R., Costa, R. S., de Oliveira Costa, G. N., da Silva, T. M., Teixeira, T. O., de Andrade, E. M. M., Galvão, A. A., Carneiro, V. L., and Figueiredo, C. A. (2015). Genetic and epigenetic studies of foxp3 in asthma and allergy. 52

[Martin et al., 2017] Martin, E., Smeester, L., Bommarito, P. A., Grace, M. R., Boggess, K., Kuban, K., Karagas, M. R., Marsit, C. J., O'Shea, T. M., and Fry,

R. C. (2017). Sexual epigenetic dimorphism in the human placenta: Implications for susceptibility during the prenatal period. *Epigenomics*, 9:267–278. 107

[Martino et al., 2011] Martino, D. J., Tulic, M. K., Gordon, L., Hodder, M., Richman, T. R., Metcalfe, J., Prescott, S. L., and Saffery, R. (2011). Evidence for age-related and individual-specific changes in dna methylation profile of mononuclear cells during early immune development in humans. *Epigenetics*, 6(9):1085–1094. 27

[Marttila et al., 2021] Marttila, S., Viiri, L. E., Mishra, P. P., Kühnel, B., Matias-Garcia, P. R., Lyytikäinen, L.-P., Ceder, T., Mononen, N., Rathmann, W., Winkelmann, J., et al. (2021). Methylation status of nc886 epiallele reflects periconceptional conditions and is associated with glucose metabolism through nc886 rnas. *Clinical epigenetics*, 13:1–18. 40

[Maschietto et al., 2017] Maschietto, M., Bastos, L. C., Tahira, A. C., Bastos, E. P., Euclydes, V. L. V., Brentani, A., Fink, G., Baumont, A. D., Felipe-Silva, A., Francisco, R. P. V., Gouveia, G., Grisi, S. J. F. E., Escobar, A. M. U., Moreira-Filho, C. A., Polanczyk, G. V., Miguel, E. C., and Brentani, H. (2017). Sex differences in dna methylation of the cord blood are related to sex-bias psychiatric diseases. *Scientific reports*, 7. 127

[Masser et al., 2018] Masser, D. R., Hadad, N., Porter, H., Stout, M. B., Unnikrishnan, A., Stanford, D. R., and Freeman, W. M. (2018). Analysis of dna modifications in aging research. *GeroScience*, 40:11–29. 24

[Matthews and Waxman, 2020] Matthews, B. J. and Waxman, D. J. (2020). Impact of 3d genome organization, guided by cohesin and ctcf looping, on sex-biased chromatin interactions and gene expression in mouse liver. *Epigenetics Chromatin 2020 13:1*, 13:1–25. 133

[McGeachie et al., 2016] McGeachie, M. J., Yates, K. P., Zhou, X., Guo, F., Sternberg, A. L., Van Natta, M. L., Wise, R. A., Szefler, S. J., Sharma, S., Kho,

A. T., et al. (2016). Genetics and genomics of longitudinal lung function patterns in individuals with asthma. *American journal of respiratory and critical care medicine*, 194(12):1465–1474. 88

[Michaud et al., 1994] Michaud, E. J., Van Vugt, M., Bultman, S. J., Sweet, H. O., Davisson, M. T., and Woychik, R. P. (1994). Differential expression of a new dominant agouti allele (aiapy) is correlated with methylation state and is influenced by parental lineage. *Genes & development*, 8(12):1463–1472. 137

[Miller et al., 2018] Miller, C. N., Dye, J. A., Schladweiler, M. C., Richards, J. H., Ledbetter, A. D., Stewart, E. J., and Kodavanti, U. P. (2018). Acute inhalation of ozone induces dna methylation of apelin in lungs of long-evans rats. *Inhalation Toxicology*, 30:178–186. 42

[Miller et al., 2007] Miller, K. A., Siscovick, D. S., Sheppard, L., Shepherd, K., Sullivan, J. H., Anderson, G. L., and Kaufman, J. D. (2007). Long-term exposure to air pollution and incidence of cardiovascular events in women. *New England Journal of Medicine*, 356:447–458. 56

[Miri et al., 2019] Miri, M., Nazarzadeh, M., Alahabadi, A., Ehrampoush, M. H., Rad, A., Lotfi, M. H., Sheikhha, M. H., Sakhvidi, M. J. Z., Nawrot, T. S., and Dadvand, P. (2019). Air pollution and telomere length in adults: A systematic review and meta-analysis of observational studies. 43

[Moen et al., 2015] Moen, E. L., Mariani, C. J., Zullow, H., Jeff-Eke, M., Litwin, E., Nikitas, J. N., and Godley, L. A. (2015). New themes in the biological functions of 5-methylcytosine and 5-hydroxymethylcytosine. *Immunological Reviews*, 263:36–49. 22

[Moore et al., 2013] Moore, L. D., Le, T., and Fan, G. (2013). Dna methylation and its basic function. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 38:23–38. 18

[Moran et al., 2016a] Moran, S., Arribas, C., and Esteller, M. (2016a). Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3):389–399. 25

[Moran et al., 2016b] Moran, S., Arribas, C., and Esteller, M. (2016b). Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8:389–399. 25

[Morris et al., 2014] Morris, T. J., Butcher, L. M., Feber, A., Teschendorff, A. E., Chakravarthy, A. R., Wojdacz, T. K., and Beck, S. (2014). Champ: 450k chip analysis methylation pipeline. *Bioinformatics*, 30:428–430. 34

[Mortusewicz et al., 2005] Mortusewicz, O., Schermelleh, L., Walter, J., Cardoso, M. C., and Leonhardt, H. (2005). Recruitment of dna methyltransferase i to dna repair sites. *Proceedings of the National Academy of Sciences of the United States of America*, 102:8905–8909. 20

[Mostafavi et al., 2018] Mostafavi, N., Vermeulen, R., Ghantous, A., Hoek, G., Probst-Hensch, N., Herceg, Z., Tarallo, S., Naccarati, A., Kleinjans, J. C., Imboden, M., Jeong, A., Morley, D., Amaral, A. F., van Nunen, E., Gulliver, J., Chadeau-Hyam, M., Vineis, P., and Vlaanderen, J. (2018). Acute changes in dna methylation in relation to 24h personal air pollution exposure measurements: A panel study in four european countries. *Environment International*, 120:11–21. 55

[MS et al., 2015] MS, G., S, Z., CJ, W., M, B., B, K., M, W., EA, O., JB, F., Z, F., and JL, S. (2015). Epigenomic profiling of dna methylation in paired prostate cancer versus adjacent benign tissue. *The Prostate*, 75:1941–1950. 123

[Mudway et al., 2019] Mudway, I. S., Dundas, I., Wood, H. E., Marlin, N., Jamaludin, J. B., Bremner, S. A., Cross, L., Grieve, A., Nanzer, A., Barratt, B. M., Beevers, S., Dajnak, D., Fuller, G. W., Font, A., Colligan, G., Sheikh, A., Walton, R., Grigg, J., Kelly, F. J., Lee, T. H., and Griffiths, C. J. (2019).

Impact of london's low emission zone on air quality and children's respiratory health: a sequential annual cross-sectional study. *The Lancet Public Health*, 4:e28–e40. 53, 54

[Mukhopadhyay and Sahu, 2018] Mukhopadhyay, S. and Sahu, S. K. (2018). A bayesian spatiotemporal model to estimate long-term exposure to outdoor air pollution at coarser administrative geographies in england and wales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181:465–486. 68, 69, 71, 76, 100

[N et al., 2016] N, P., S, F., N, P., and FH, G. (2016). Proteomics analysis of human tears from aqueous-deficient and evaporative dry eye patients. *Scientific reports*, 6. 131

[N et al., 1995] N, Y., R, K., R, M., T, S., and M, N. (1995). A dna methylation site in the male-specific p450 (cyp 2d-9) promoter and binding of the heteromeric transcription factor gabp. *Molecular and cellular biology*, 15:5355–5362. 130

[Nasser et al., 2021] Nasser, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., Lekschas, F., Mualim, K., Natri, H. M., Weeks, E. M., Munson, G., Kane, M., Kang, H. Y., Cui, A., Ray, J. P., Eisenhaure, T. M., Collins, R. L., Dey, K., Pfister, H., Price, A. L., Epstein, C. B., Kundaje, A., Xavier, R. J., Daly, M. J., Huang, H., Finucane, H. K., Hacohen, N., Lander, E. S., and Engreitz, J. M. (2021). Genome-wide enhancer maps link risk variants to disease genes. *Nature 2021 593:7858*, 593:238–243. 115

[Nef and Vassalli, 2009] Nef, S. and Vassalli, J.-D. (2009). Complementary pathways in mammalian female sex determination. *Journal of Biology 2009 8:8*, 8:1–3. 109

[Neven et al., 2018] Neven, K. Y., Saenen, N. D., Tarantini, L., Janssen, B. G., Lefebvre, W., Vanpoucke, C., Bollati, V., and Nawrot, T. S. (2018). Placental promoter methylation of dna repair genes and prenatal exposure to particulate air pollution: an environage cohort study. *The Lancet Planetary Health*, 2:e174–e183. 49, 68

[Nikolaienko et al., 2022] Nikolaienko, O., Lønning, P. E., and Knappskog, S. (2022). epialleler: an r/bioc package for sensitive allele-specific methylation analysis in ngs data. *bioRxiv*, pages 2022–06. 40

[Numata et al., 2012] Numata, S., Ye, T., Hyde, T. M., Guitart-Navarro, X., Tao, R., Wininger, M., Colantuoni, C., Weinberger, D. R., Kleinman, J. E., and Lipska, B. K. (2012). Dna methylation signatures in development and aging of the human prefrontal cortex. *American Journal of Human Genetics*, 90:260–272. 107

[Nwanaji-Enwerem et al., 2016a] Nwanaji-Enwerem, J. C., Colicino, E., Trevisi, L., Kloog, I., Just, A. C., Shen, J., Brennan, K., Dereix, A., Hou, L., Vokonas, P., et al. (2016a). Long-term ambient particle exposures and blood dna methylation age: findings from the va normative aging study. *Environmental epigenetics*, 2(2):dvw006. 94

[Nwanaji-Enwerem et al., 2016b] Nwanaji-Enwerem, J. C., Colicino, E., Trevisi, L., Kloog, I., Just, A. C., Shen, J., Brennan, K., Dereix, A., Hou, L., Vokonas, P., Schwartz, J., and Baccarelli, A. A. (2016b). Long-term ambient particle exposures and blood dna methylation age: findings from the va normative aging study. *Environmental Epigenetics*, 2:dvw006. 62

[Okano et al., 1999] Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99:247–257. 20

[Olea, 1999] Olea, R. A. (1999). *Geostatistics for Engineers and Earth Scientists*. Springer US. 101

[Olsson et al., 2014] Olsson, A. H., Volkov, P., Bacos, K., Dayeh, T., Hall, E., Nilsson, E. A., Ladenvall, C., Rönn, T., and Ling, C. (2014). Genome-wide associations between genetic and epigenetic variation influence mrna expression and insulin secretion in human pancreatic islets. *PLoS genetics*, 10(11):e1004735. 138

[on the Health Effects of Traffic-Related Air Pollution, 2010] on the Health Effects of Traffic-Related Air Pollution, H. E. I. P. (2010). Traffic-related air pollution: a critical review of the literature on emissions, exposure, and health effects. 75

[Oudin et al., 2016] Oudin, A., Forsberg, B., Adolfsson, A. N., Lind, N., Modig, L., Nordin, M., Nordin, S., Adolfsson, R., and Nilsson, L.-G. (2016). Traffic-related air pollution and dementia incidence in northern sweden: a longitudinal study. *Environmental health perspectives*, 124(3):306–312. 75

[P and M, 2021] P, S. and M, R. (2021). Motifdb: An annotated collection of protein-dna binding sequence motifs. *R package version 1.34.0.* 36

[Palumbo et al., 2018] Palumbo, D., Affinito, O., Monticelli, A., and Cocozza, S. (2018). Dna methylation variability among individuals is related to cpgs cluster density and evolutionary signatures. *BMC genomics*, 19(1):1–9. 136

[Pebesma and Bivand, 2005a] Pebesma, E. and Bivand, R. S. (2005a). Classes and methods for spatial data: the sp package. 75

[Pebesma, 2004] Pebesma, E. J. (2004). Multivariable geostatistics in s: The gstat package. *Computers and Geosciences*, 30:683–691. 75

[Pebesma and Bivand, 2005b] Pebesma, E. J. and Bivand, R. S. (2005b). Classes and methods for spatial data in R. *R News*, 5(2):9–13. 75

[Perdomo-Sabogal et al., 2016] Perdomo-Sabogal, A., Nowick, K., Piccini, I., Sudbrak, R., Lehrach, H., Yaspo, M.-L., Warnatz, H.-J., and Querfurth, R. (2016).

Human lineage-specific transcriptional regulation through ga-binding protein transcription factor alpha (gabpa). *Molecular biology and evolution*, 33(5):1231–1244. 130

[Perera et al., 2009] Perera, F., yee Tang, W., Herbstman, J., Tang, D., Levin, L., Miller, R., and mei Ho, S. (2009). Correction: Relation of dna methylation of 5-cpg island of acsl3 to transplacental exposure to airborne polycyclic aromatic hydrocarbons and childhood asthma. *PLoS ONE*, 4. 46, 50, 53, 67

[Perry et al., 1994] Perry, W. L., Copeland, N. G., and Jenkins, N. A. (1994). The molecular basis for dominant yellow agouti coat color mutations. *Bioessays*, 16(10):705–707. 137

[Pidsley et al., 2013] Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C. (2013). A data-driven approach to preprocessing illumina 450k methylation array data. *BMC genomics*, 14(1):1–10. 32, 33

[Pidsley et al., 2016] Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Djik, S. V., Muhlhausler, B., Stirzaker, C., and Clark, S. J. (2016). Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling. *Genome Biology*, 17:1–17. 24, 25, 131

[Plusquin et al., 2018] Plusquin, M., Chadeau-Hyam, M., Ghantous, A., Alfano, R., Bustamante, M., Chatzi, L., Cuenin, C., Gulliver, J., Herceg, Z., Kogevinas, M., Nawrot, T. S., Pizzi, C., Porta, D., Relton, C. L., Richiardi, L., Robinson, O., Sunyer, J., Vermeulen, R., Vriens, A., Vrijheid, M., Henderson, J., and Vineis, P. (2018). Dna methylome marks of exposure to particulate matter at three time points in early life. *Environmental Science and Technology*, 52:5427–5437. 66

[Plusquin et al., 2017] Plusquin, M., Guida, F., Polidoro, S., Vermeulen, R., Raaschou-Nielsen, O., Campanella, G., Hoek, G., Kyrtopoulos, S. A., Georgiadis, P., Naccarati, A., Sacerdote, C., Krogh, V., de Mesquita, H. B. B., Verschuren,

W. M. M., Sayols-Baixeras, S., Panni, T., Peters, A., Hebels, D. G., Kleinjans, J., Vineis, P., and Chadeau-Hyam, M. (2017). Dna methylation and exposure to ambient air pollution in two prospective cohorts. *Environment International*, 108:127–136. 43, 44, 54, 66

[Popp et al., 2010] Popp, C., Dean, W., Feng, S., Cokus, S. J., Andrews, S., Pellegrini, M., Jacobsen, S. E., and Reik, W. (2010). Genome-wide erasure of dna methylation in mouse primordial germ cells is affected by aid deficiency. *Nature*, 463:1101–1105. 22

[Price et al., 2013] Price, M. E., Cotton, A. M., Lam, L. L., Farré, P., Emberly, E., Brown, C. J., Robinson, W. P., and Kobor, M. S. (2013). Additional annotation enhances potential for biologically-relevant analysis of the illumina infinium humanmethylation450 beadchip array. *Epigenetics and Chromatin*, 6:4. 107

[Prior and Walter, 1996] Prior, H. M. and Walter, M. A. (1996). Sox genes: Architects of development. 113

[Probst et al., 2009] Probst, A. V., Dunleavy, E., and Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. *Nature reviews Molecular cell biology*, 10(3):192–206. 20

[Qin et al., 2019] Qin, X., Li, J., Wu, T., Wu, Y., Tang, X., Gao, P., Li, L., Wang, M., Wu, Y., Wang, X., Chen, D., and Hu, Y. (2019). Overall and sex-specific associations between methylation of the abcg1 and apoe genes and ischemic stroke or other atherosclerosis-related traits in a sibling study of chinese population. *Clinical Epigenetics*, 11:189. 106

[Quach et al., 2017] Quach, A., Levine, M. E., Tanaka, T., Lu, A. T., Chen, B. H., Ferrucci, L., Ritz, B., Bandinelli, S., Neuhouser, M. L., Beasley, J. M., Snetselaar, L., Wallace, R. B., Tsao, P. S., Absher, D., Assimes, T. L., Stewart, J. D., Li, Y., Hou, L., Baccarelli, A. A., Whitsel, E. A., and Horvath, S. (2017). Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging*, 9:419–446. 59

[R, 2020] R, D. D. S. (2020). Pwmenrich: Pwm enrichment analysis. *R package verion 4.26.0.* 36

[Rakyan et al., 2004] Rakyan, V. K., Hildmann, T., Novik, K. L., Lewin, J., Tost, J., Cox, A. V., Andrews, T. D., Howe, K. L., Otto, T., Olek, A., et al. (2004). Dna methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS biology*, 2(12):e405. 136

[Ramírez et al., 2018] Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (2018). High-resolution tads reveal dna sequences underlying genome organization in flies. *Nature communications*, 9(1):1–15. 40

[Rasmussen and Helin, 2016] Rasmussen, K. D. and Helin, K. (2016). Role of tet enzymes in dna methylation, development, and cancer. *Genes & development*, 30(7):733–750. 22

[Rauluseviciute et al., 2020] Rauluseviciute, I., Drabløs, F., and Rye, M. B. (2020). Dna hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. *BMC Medical Genomics 2020 13:1*, 13:1–15. 123, 132, 166

[Razin and Riggs, 1980] Razin, A. and Riggs, A. D. (1980). Dna methylation and gene function. *Science*, 210:604–610. 18

[Rice et al., 2016] Rice, M. B., Rifas-Shiman, S. L., Litonjua, A. A., Oken, E., Gillman, M. W., Kloog, I., Luttmann-Gibson, H., Zanobetti, A., Coull, B. A., Schwartz, J., Koutrakis, P., Mittleman, M. A., and Gold, D. R. (2016). Lifetime exposure to ambient pollution and lung function in children. *American Journal of Respiratory and Critical Care Medicine*, 193:881–888. 57

[Rider and Carlsten, 2019] Rider, C. F. and Carlsten, C. (2019). Air pollution and dna methylation: effects of exposure in humans. *Clinical Epigenetics*, 11:131. 41, 47, 64

[Ritchie et al., 2015] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43:e47. 34, 35, 38

[Rivera-González et al., 2015a] Rivera-González, L. O., Zhang, Z., Sánchez, B. N., Zhang, K., Brown, D. G., Rojas-Bracho, L., Osornio-Vargas, A., Vadillo-Ortega, F., and O'Neill, M. S. (2015a). An assessment of air pollutant exposure methods in mexico city, mexico. *Journal of the Air and Waste Management Association*, 65:581–591. 61

[Rivera-González et al., 2015b] Rivera-González, L. O., Zhang, Z., Sánchez, B. N., Zhang, K., Brown, D. G., Rojas-Bracho, L., Osornio-Vargas, A., Vadillo-Ortega, F., and O'Neill, M. S. (2015b). An assessment of air pollutant exposure methods in mexico city, mexico. *Journal of the Air and Waste Management Association*, 65:581–591. 99, 100

[Rocks et al., 2021] Rocks, D., Shukla, M., Finnemann, S. C., Kalluchi, A., Rowley, M. J., and Kundakovic, M. (2021). Sex-specific multi-level 3d genome dynamics in the mouse brain. *bioRxiv*, page 2021.05.03.442383. 133

[Rubin, 2022] Rubin, J. B. (2022). The spectrum of sex differences in cancer. *Trends in Cancer*, 0. 111

[Rubin et al., 2020] Rubin, J. B., Lagas, J. S., Broestl, L., Sponagel, J., Rockwell, N., Rhee, G., Rosen, S. F., Chen, S., Klein, R. S., Imoukhuede, P., and Luo, J. (2020). Sex differences in cancer mechanisms. *Biology of sex differences*, 11. 111

[Rubtsova et al., 2015] Rubtsova, K., Marrack, P., and Rubtsov, A. V. (2015). Sexual dimorphism in autoimmunity. *Journal of Clinical Investigation*, 125:2187–2193. 107

[Ruprecht et al., 2016] Ruprecht, A. A., Marco, C. D., Pozzi, P., Mazza, R., Munarini, E., Paco, A. D., Paredi, P., Invernizzi, G., and Boffi, R. (2016). Outdoor second-hand cigarette smoke significantly affects air quality. *European Respiratory Journal*, 48:918–920. 60

[Ryan et al., 2005] Ryan, P. H., LeMasters, G., Biagini, J., Bernstein, D., Grinshpun, S. A., Shukla, R., Wilson, K., Villareal, M., Burkle, J., and Lockey, J. (2005). Is it traffic type, volume, or distance? wheezing in infants living near truck and bus traffic. *Journal of Allergy and Clinical Immunology*, 116:279–284. 44, 66, 99

[Rytel et al., 2021] Rytel, M. R., Butler, R., Eliot, M., Braun, J. M., Houseman, E. A., and Kelsey, K. T. (2021). Dna methylation in the adipose tissue and whole blood of agent orange-exposed operation ranch hand veterans: a pilot study. *Environmental Health*, 20(1):1–13. 89, 105

[Sadler et al., 2021] Sadler, M. C., Auwerx, C., Porcu, E., and Kutalik, Z. (2021). Quantifying mediation between omics layers and complex traits. *bioRxiv*, page 2021.09.29.462396. 123, 132

[Salam et al., 2008] Salam, M. T., Islam, T., and Gilliland, F. D. (2008). Recent evidence for adverse effects of residential proximity to traffic sources on asthma. *Current opinion in pulmonary medicine*, 14(1):3–8. 75

[Samoli et al., 2008] Samoli, E., Peng, R., Ramsay, T., Pipikou, M., Touloumi, G., Dominici, F., Burnett, R., Cohen, A., Krewski, D., Samet, J., and Katsouyanni, K. (2008). Acute effects of ambient particulate matter on mortality in europe and north america: Results from the aphena study. *Environmental Health Perspectives*, 116:1480–1486. 56

[Savage et al., 2013] Savage, N. H., Agnew, P., Davis, L. S., Ordóñez, C., Thorpe, R., Johnson, C. E., O'Connor, F. M., and Dalvi, M. (2013). Air quality modelling using the met office unified model (aqum os24-26): model description and initial evaluation. *Geoscientific Model Development*, 6:353–372. 100

[Sayols-Baixeras et al., 2019] Sayols-Baixeras, S., Fernández-Sanlés, A., Prats-Uribe, A., Subirana, I., Plusquin, M., Künzli, N., Marrugat, J., Basagaña, X., and Elosua, R. (2019). Association between long-term air pollution exposure and dna methylation: the regicor study. *Environmental research*, 176:108550. 56, 83, 93

[Schneider and Garrett, 2009] Schneider, K. and Garrett, L. (2009). The end of the era of generosity? global health amid economic crisis. *Philosophy, Ethics, and Humanities in Medicine*, 4:1. 100

[Schröder et al., 2017] Schröder, C., Leitão, E., Wallner, S., Schmitz, G., Klein-Hitpass, L., Sinha, A., Jöckel, K.-H., Heilmann-Heimbach, S., Hoffmann, P., Nöthen, M. M., et al. (2017). Regions of common inter-individual dna methylation differences in human monocytes: genetic basis and potential function. *Epigenetics & chromatin*, 10(1):1–18. 136

[Schuller and Montrose, 2020] Schuller, A. and Montrose, L. (2020). Influence of woodsmoke exposure on molecular mechanisms underlying alzheimer's disease: Existing literature and gaps in our understanding. *Epigenetics Insights*, 13. 59

[Schultz et al., 2017] Schultz, E. S., Litonjua, A. A., and Melén, E. (2017). Effects of long-term exposure to traffic-related air pollution on lung function in children. *Current Allergy and Asthma Reports*, 17. 57

[Schurmann et al., 2012] Schurmann, C., Heim, K., Schillert, A., Blankenberg, S., Carstensen, M., Dörr, M., Endlich, K., Felix, S. B., Gieger, C., Grallert, H., et al. (2012). Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PloS one*, 7(12):e50938. 32

[Sekine et al., 2004] Sekine, K., Shima, M., Nitta, Y., and Adachi, M. (2004). Long term effects of exposure to automobile exhaust on the pulmonary function of female adults in tokyo, japan. *Occupational and Environmental Medicine*, 61:350–357. 99

[Selevan et al., 2000] Selevan, S. G., Kimmel, C. A., and Mendola, P. (2000). Identifying critical windows of exposure for children's health. *Environmental Health Perspectives*, 108:451–455. 50

[Shabalin, 2012] Shabalin, A. A. (2012). Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358. 39

[Shaddick et al., 2008] Shaddick, G., Lee, D., Zidek, J. V., and Salway, R. (2008). Estimating exposure response functions using ambient pollution concentrations. *Annals of Applied Statistics*, 2:1249–1270. 101

[Shi et al., 2020] Shi, L., Wu, X., Yazdi, M. D., Braun, D., Awad, Y. A., Wei, Y., Liu, P., Di, Q., Wang, Y., Schwartz, J., Dominici, F., Kioumourtzoglou, M. A., and Zanobetti, A. (2020). Long-term effects of pm2·5 on neurological disorders in the american medicare population: a longitudinal cohort study. *The Lancet Planetary Health*, 0. 59

[Shi et al., 2022] Shi, Y., Zhang, H., Huang, S., Yin, L., Wang, F., Luo, P., and Huang, H. (2022). Epigenetic regulation in cardiovascular disease: Mechanisms and advances in clinical trials. *Signal Transduction and Targeted Therapy*, 7(1):1–28. 98

[Shireby et al., 2022] Shireby, G., Dempster, E., Policicchio, S., Smith, R., Pishva, E., Chioza, B., Davies, J., Burrage, J., Lunnon, K., Seiler-Vellame, D., et al. (2022). Dna methylation signatures of alzheimer's disease neuropathology in the cortex are primarily driven by variation in non-neuronal cell-types. *bioRxiv*. 31

[Singmann et al., 2015] Singmann, P., Shem-Tov, D., Wahl, S., Grallert, H., Fiorito, G., Shin, S.-Y., Schramm, K., Wolf, P., Kunze, S., Baran, Y., Guarrera, S., Vineis, P., Krogh, V., Panico, S., Tumino, R., Kretschmer, A., Gieger, C., Peters, A., Prokisch, H., Relton, C. L., Matullo, G., Illig, T., Waldenberger, M., and Halperin, E. (2015). Characterization of whole-genome autosomal differences of dna methylation between men and women. *Epigenetics Chromatin*. 107

[Smallwood et al., 2011] Smallwood, S. A., Tomizawa, S. I., Krueger, F., Ruf, N., Carli, N., Segonds-Pichon, A., Sato, S., Hata, K., Andrews, S. R., and Kelsey, G. (2011). Dynamic cpg island methylation landscape in oocytes and preimplantation embryos. *Nature Genetics*, 43:811–814. 19

[Smith-Bouvier et al., 2008] Smith-Bouvier, D. L., Divekar, A. A., Sasidhar, M., Du, S., Tiwari-Woodruff, S. K., King, J. K., Arnold, A. P., Singh, R. R., and Voskuhl, R. R. (2008). A role for sex chromosome complement in the female bias in autoimmune disease. *The Journal of experimental medicine*, 205(5):1099–1108. 106

[Snir et al., 2019] Snir, S., Farrell, C., and Pellegrini, M. (2019). Human epigenetic ageing is logarithmic with time across the entire lifespan. *Epigenetics*, 14(9):912–926. 27

[Somineni et al., 2016] Somineni, H. K., Zhang, X., Myers, J. M. B., Kovacic, M. B., Ulm, A., Jurcak, N., Ryan, P. H., Hershey, G. K. K., and Ji, H. (2016). Ten-eleven translocation 1 (tet1) methylation is associated with childhood asthma and traffic-related air pollution. *Journal of Allergy and Clinical Immunology*, 137:797–805.e5. 51, 52, 124

[Son et al., 2010] Son, J. Y., Bell, M. L., and Lee, J. T. (2010). Individual exposure to air pollution and lung function in korea: Spatial analysis using multiple exposure approaches. *Environmental Research*, 110:739–749. 68, 73

[Song and He, 2013] Song, C. X. and He, C. (2013). Potential functional roles of dna demethylation intermediates. *Trends in Biochemical Sciences*, 38:480–484. 24

[Song et al., 2017] Song, Y., Liu, T., Wang, Y., Deng, J., Chen, M., Yuan, L., Lu, Y., Xu, Y., Yao, H., Li, Z., and Lai, L. (2017). Mutation of the sp1 binding site in the 5' flanking region of sry causes sex reversal in rabbits. *Oncotarget*, 8:38176–38183. 132

[SP and D, 2015] SP, W. and D, W. (2015). Signaling pathways involved in mammalian sex determination and gonad development. *Sexual development : genetics, molecular biology, evolution, endocrinology, embryology, and pathology of sex determination and differentiation*, 9:297–315. 109

[Steinke, 2016] Steinke, J. W. (2016). Can genes control asthmatic lung function patterns? *American Journal of Respiratory and Critical Care Medicine*, 194(12):1439–1440. 89

[Steinvil et al., 2008] Steinvil, A., Kordova-Biezuner, L., Shapira, I., Berliner, S., and Rogowski, O. (2008). Short-term exposure to air pollution and inflammation-sensitive biomarkers. *Environmental Research*, 106:51–61. 43

[Stieb et al., 2012] Stieb, D. M., Chen, L., Eshoul, M., and Judek, S. (2012). Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. 48

[Suderman et al., 2017] Suderman, M., Simpkin, A., Sharp, G., Gaunt, T., Lyttleton, O., McArdle, W., Ring, S., Smith, G. D., and Relton, C. (2017). Sex-associated autosomal dna methylation differences are wide-spread and stable throughout childhood. *Biorxiv*, page 118265. 107, 122

[Sugathan and Waxman, 2013] Sugathan, A. and Waxman, D. J. (2013). Genome-wide analysis of chromatin states reveals distinct mechanisms of sex-dependent gene regulation in male and female mouse liver. *Molecular and cellular biology*, 33(18):3594–3610. 107

[Sun et al., 2014] Sun, L., Lin, J., Du, H., Hu, C., Huang, Z., Lv, Z., Zheng, C., Shi, X., Zhang, Y., and Yang, Z. (2014). Gender-specific dna methylome analysis of a han chinese longevity population. *BioMed Research International*, 2014:1–9. 107

[Sunny et al., 2021] Sunny, S. K., Zhang, H., Relton, C. L., Ring, S., Kadalayil, L., Mzayek, F., Ewart, S., Holloway, J. W., and Arshad, S. H. (2021). Sex-specific longitudinal association of dna methylation with lung function. *ERJ Open Research*, 7:00127–2021. 125

[Swan, 1996] Swan, A. (1996). H. wackernagel, 1995. multivariate geostatistics. an introduction with applications. xiv + 256 pp. berlin, heidelberg, new york, barcelona, budapest, hong kong, london, milan, paris, tokyo: Springer-verlag. price dm 74.00, Ös 540.20, sfr 71.50 (hard covers). isbn 3 540 60127 9. *Geological Magazine*, 133:628–628. 73

[Szyf et al., 2004] Szyf, M., Pakneshan, P., and Rabbani, S. A. (2004). Dna methylation and breast cancer. *Biochemical pharmacology*, 68(6):1187–1197. 22

[T and JC, 2020] T, L. and JC, M. (2020). Investigating transcriptome-wide sex dimorphism by multi-level analysis of single-cell rna sequencing data in ten mouse cell types. *Biology of sex differences*, 11. 130

[Teschendorff et al., 2009] Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Gayther, S. A., Apostolidou, S., Jones, A., Lechner, M., Beck, S., Jacobs, I. J., et al. (2009). An epigenetic signature in peripheral blood predicts active ovarian cancer. *PloS one*, 4(12):e8274. 25

[Tsukasaki et al., 2001] Tsukasaki, K., Miller, C. W., Greenspun, E., Eshaghian, S., Kawabata, H., Fujimoto, T., Tomonaga, M., Sawyers, C., Said, J. W., and

Koeffler, H. P. (2001). Mutations in the mitotic check point gene, mad1l1, in human cancers. *Oncogene*, 20(25):3301–3305. 143

[Vaissière et al., 2008] Vaissière, T., Sawan, C., and Herceg, Z. (2008). Epigenetic interplay between histone modifications and dna methylation in gene silencing. *Mutation Research/Reviews in Mutation Research*, 659(1-2):40–48. 132

[Van Baak et al., 2018] Van Baak, T. E., Coarfa, C., Dugué, P.-A., Fiorito, G., Laritsky, E., Baker, M. S., Kessler, N. J., Dong, J., Duryea, J. D., Silver, M. J., et al. (2018). Epigenetic supersimilarity of monozygotic twin pairs. *Genome biology*, 19(1):1–20. 138

[Van Der Plaat et al., 2019] Van Der Plaat, D. A., Vonk, J. M., Terzikhan, N., de Jong, K., De Vries, M., La Bastide-van Gemert, S., Van Diemen, C. C., Lahousse, L., Brusselle, G. G., Nedeljkovic, I., et al. (2019). Occupational exposure to gases/fumes and mineral dust affect dna methylation levels of genes regulating expression. *Human molecular genetics*, 28(15):2477–2485. 89

[van Rossem et al., 2015] van Rossem, L., Rifas-Shiman, S. L., Melly, S. J., Kloog, I., Luttmann-Gibson, H., Zanobetti, A., Coull, B. A., Schwartz, J. D., Mittleman, M. A., Oken, E., et al. (2015). Prenatal air pollution exposure and newborn blood pressure. *Environmental health perspectives*, 123(4):353–359. 50

[Villicaña and Bell, 2021] Villicaña, S. and Bell, J. T. (2021). Genetic impacts on dna methylation: research findings and future perspectives. *Genome biology*, 22(1):127. 148

[Vinson and Chatterjee, 2012] Vinson, C. and Chatterjee, R. (2012). Cg methylation. *Epigenomics*, 4(6):655–663. 19

[Vogt et al., 2011] Vogt, J., Kohlhase, J., Morlot, S., Kluwe, L., Mautner, V.-F., Cooper, D. N., and Kehrer-Sawatzki, H. (2011). Monozygotic twins discordant for neurofibromatosis type 1 due to a postzygotic nf1 gene mutation. *Human mutation*, 32(6):E2134–E2147. 136

[Voisin et al., 2020] Voisin, S., Harvey, N. R., Haupt, L. M., Griffiths, L. R., Ashton, K. J., Coffey, V. G., Doering, T. M., Thompson, J.-L. M., Benedict, C., Cedernaes, J., et al. (2020). An epigenetic clock for human skeletal muscle. *Journal of cachexia, sarcopenia and muscle*, 11(4):887–898. 31

[Wagner et al., 2014] Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. (2014). The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology*, 15:R37. 18

[Walton et al., 2011] Walton, E. L., Francastel, C., and Velasco, G. (2011). Maintenance of dna methylation: Dnmt3b joins the dance. *Epigenetics*, 6:1373–1377. 21

[Wang et al., 2016] Wang, C., Chen, R., Cai, J., Shi, J., Yang, C., Tse, L. A., Li, H., Lin, Z., Meng, X., Liu, C., Niu, Y., Xia, Y., Zhao, Z., and Kan, H. (2016). Personal exposure to fine particulate matter and blood pressure: A role of angiotensin converting enzyme and its dna methylation. *Environment International*, 94:661–666. 58

[Wang et al., 2012] Wang, J., Zhuang, J., Iyer, S., Lin, X. Y., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M., and Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22:1798–1812. 19

[Wang et al., 2020a] Wang, Q., Chen, Y., Readhead, B., Chen, K., Su, Y., Reiman, E. M., and Dudley, J. T. (2020a). Longitudinal data in peripheral blood confirm that pm20d1 is a quantitative trait locus (qtl) for alzheimer's disease and implicate its dynamic role in disease progression. *Clinical Epigenetics*, 12:1–18. 49

[Wang et al., 2020b] Wang, Q., Chen, Y., Readhead, B., Chen, K., Su, Y., Reiman, E. M., and Dudley, J. T. (2020b). Longitudinal data in peripheral blood confirm that pm20d1 is a quantitative trait locus (qtl) for alzheimer's disease and implicate its dynamic role in disease progression. *Clinical epigenetics*, 12(1):1–18. 155

[Wang et al., 2022] Wang, Y., Gorrie-Stone, T. J., Grant, O. A., Andrayas, A. D., Zhai, X., McDonald-Maier, K. D., and Schalkwyk, L. C. (2022). Interpolatedxy: a two-step strategy to normalize dna methylation microarray data avoiding sex bias. *Bioinformatics*, 38(16):3950–3957. 108, 129

[Wang et al., 2011] Wang, Y., Gulliver, J., and McHugh, C. (2011). Modeling the health impacts of air pollution exposures in london within the genesis system. volume 3, pages 2341–2344. 69

[Wang et al., 2021] Wang, Y., Hannon, E., Grant, O. A., Gorrie-Stone, T. J., Kumari, M., Mill, J., Zhai, X., McDonald-Maier, K. D., and Schalkwyk, L. C. (2021). Dna methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy. *BMC genomics*, 22(1):1–11. 33, 108, 129

[Wang et al., 2018] Wang, Z., Wu, X., and Wang, Y. (2018). A framework for analyzing dna methylation data from illumina infinium humanmethylation450 beadchip. *BMC bioinformatics*, 19(5):15–22. 25

[Ward et al., 2020] Ward, N. P., Kang, Y. P., Falzone, A., Boyle, T. A., and DeNicola, G. M. (2020). Nicotinamide nucleotide transhydrogenase regulates mitochondrial metabolism in nsclc through maintenance of fe-s protein function. *Journal of Experimental Medicine*, 217(6). 94

[Ward-Caviness et al., 2016] Ward-Caviness, C. K., Nwanaji-Enwerem, J. C., Wolf, K., Wahl, S., Colicino, E., Trevisi, L., Kloog, I., Just, A. C., Vokonas, P., Cyrys, J., Gieger, C., Schwartz, J., Baccarelli, A. A., Schneider, A., and Peters, A.

(2016). Long-term exposure to air pollution is associated with biological aging. *Oncotarget*, 7:74510–74525. 62

[Waters, 2017] Waters, N. (2017). Tobler's first law of geography. *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, pages 1–13. 72

[Werner et al., 2017] Werner, R. J., Schultz, B. M., Huhn, J. M., Jelinek, J., Madzo, J., and Engel, N. (2017). Sex chromosomes drive gene expression and regulatory dimorphisms in mouse embryonic stem cells. *Biology of Sex Differences*, 8:1–18. 106

[WHO, 2016] WHO (2016). Ambient air pollution: A global assessment of exposure and burden of disease. 41, 99

[Wickham, 2011] Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3:180–185. 75

[Wiehle et al., 2019] Wiehle, L., Thorn, G. J., Raddatz, G., Clarkson, C. T., Rippe, K., Lyko, F., Breiling, A., and Teif, V. B. (2019). Dna (de)methylation in embryonic stem cells controls ctcf-dependent chromatin boundaries. *Genome Research*, 29:750–761. 23

[Wijchers and Festenstein, 2011] Wijchers, P. J. and Festenstein, R. J. (2011). Epigenetic regulation of autosomal gene expression by sex chromosomes. *Trends in genetics*, 27(4):132–140. 106

[Wikenius et al., 2019] Wikenius, E., Moe, V., Smith, L., Heiervang, E. R., and Berglund, A. (2019). Dna methylation changes in infants between 6 and 52 weeks. *Scientific reports*, 9(1):1–12. 27

[Wiseman et al., 2015] Wiseman, F. K., Al-Janabi, T., Hardy, J., Karmiloff-Smith, A., Nizetic, D., Tybulewicz, V. L., Fisher, E., and Strydom, A. (2015). A genetic cause of alzheimer disease: mechanistic insights from down syndrome. *Nature Reviews Neuroscience*, 16(9):564–574. 124

[Wolff et al., 2022] Wolff, J., Backofen, R., and Grüning, B. (2022). Loop detection using hi-c data with hicexplorer. *GigaScience*, 11. 40

[Wong et al., 2005] Wong, A. H., Gottesman, I. I., and Petronis, A. (2005). Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Human molecular genetics*, 14(suppl_1):R11–R18. 135

[Wood et al., 2020] Wood, N. M., Trebilco, T., and Cohen-Woods, S. (2020). Scars of childhood socioeconomic stress: A systematic review. 42

[Wu et al., 2016] Wu, Q., Fukuda, K., Kato, Y., Zhou, Z., Deng, C.-X., and Saga, Y. (2016). Sexual fate change of xx germ cells caused by the deletion of smad4 and stra8 independent of somatic sex reprogramming. *PLoS Biology*, 14. 122

[Xavier et al., 2019] Xavier, M. J., Roman, S. D., Aitken, R. J., and Nixon, B. (2019). Transgenerational inheritance: how impacts to the epigenetic and genetic information of parents affect offspring health. *Human reproduction update*, 25(5):519–541. 138

[Xia et al., 2019] Xia, X., Zhou, X., Quan, Y., Hu, Y., Xing, F., Li, Z., Xu, B., Xu, C., and Zhang, A. (2019). Germline deletion of cdyl causes teratozoospermia and progressive infertility in male mice. *Cell Death Disease 2019 10:3*, 10:1–13. 117

[Xia et al., 2021] Xia, Y., Dai, R., Wang, K., Jiao, C., Zhang, C., Xu, Y., Li, H., Jing, X., Chen, Y., Jiang, Y., Kopp, R. F., Giase, G., Chen, C., and Liu, C. (2021). Sex-differential dna methylation and associated regulation networks in human brain implicated in the sex-biased risks of psychiatric disorders. *Molecular Psychiatry*, 26:835–848. 106

[Xie et al., 1999] Xie, S., Wang, Z., Okano, M., Nogami, M., Li, Y., He, W. W., Okumura, K., and Li, E. (1999). Cloning, expression and chromosome locations of the human dnmt3 gene family. *Gene*, 236:87–95. 18, 20

[Xu et al., 2014] Xu, H., Wang, F., Liu, Y., Yu, Y., Gelernter, J., and Zhang, H. (2014). Sex-biased methylome and transcriptome in human prefrontal cortex. *Human Molecular Genetics*, 23:1260–1270. 125, 127

[Xu et al., 2013] Xu, Z., Bolick, S. C., DeRoo, L. A., Weinberg, C. R., Sandler, D. P., and Taylor, J. A. (2013). Epigenome-wide association study of breast cancer using prospectively collected sister study samples. *Journal of the National Cancer Institute*, 105(10):694–700. 25

[Yan et al., 2022] Yan, N., Li, Y., Xing, Y., Wu, J., Li, J., Liang, Y., Tang, Y., Wang, Z., Song, H., Wang, H., et al. (2022). Developmental arsenic exposure impairs cognition, directly targets dnmt3a, and reduces dna methylation. *EMBO reports*, 23(6):e54147. 94

[Yang et al., 2022] Yang, R., Hu, Y., Lee, C. H., Liu, Y., Diaz-Canestro, C., Fong, C. H. Y., Lin, H., Cheng, K. K., Pravelil, A. P., Song, E., et al. (2022). Pm20d1 is a circulating biomarker closely associated with obesity, insulin resistance and metabolic syndrome. *European Journal of Endocrinology*, 186(2):151–161. 155

[Yaqinuddin et al., 2008] Yaqinuddin, A., Qureshi, S. A., Qazi, R., and Abbas, F. (2008). Down-regulation of dnmt3b in pc3 cells effects locus-specific dna methylation, and represses cellular growth and migration. *Cancer Cell International*, 8:13. 21

[Yokomori et al., 1995] Yokomori, N., Kobayashi, R., Moore, R., Sueyoshi, T., and Negishi, M. (1995). A dna methylation site in the male-specific p450 (cyp 2d-9) promoter and binding of the heteromeric transcription factor gabp. *Molecular and Cellular Biology*, 15(10):5355–5362. 130

[Young et al., 2017] Young, P. E., Kum Jew, S., Buckland, M. E., Pamphlett, R., and Suter, C. M. (2017). Epigenetic differences between monozygotic twins discordant for amyotrophic lateral sclerosis (als) provide clues to disease pathogenesis. *PLoS One*, 12(8):e0182638. 136

[Yousefi et al., 2011] Yousefi, P., Huen, K., Davé, V., Barcellos, L., Eskenazi, B., and Holland, N. (2011). Sex differences in dna methylation assessed by 450 k beadchip in newborns. 107, 124, 125, 127, 130, 131

[Yu et al., 2012] Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). Clusterprofiler: An r package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, 16:284–287. 37

[Yu et al., 2015] Yu, X.-J., Yang, M.-J., Zhou, B., Wang, G.-Z., Huang, Y.-C., Wu, L.-C., Cheng, X., Wen, Z.-S., Huang, J.-Y., Zhang, Y.-D., et al. (2015). Characterization of somatic mutations in air pollution-related lung cancer. *EBioMedicine*, 2(6):583–590. 88, 104

[Zandbergen, 2007] Zandbergen, P. A. (2007). Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health*, 7. 45, 67

[Zeng et al., 2019] Zeng, Y., Amador, C., Xia, C., Marioni, R., Sproul, D., Walker, R. M., Morris, S. W., Bretherick, A., Canela-Xandri, O., Boutin, T. S., et al. (2019). Parent of origin genetic effects on methylation in humans are common and influence complex trait variation. *Nature communications*, 10(1):1–13. 161

[Zhang et al., 2011] Zhang, F. F., Cardarelli, R., Carroll, J., Fulda, K. G., Kaur, M., Gonzalez, K., Vishwanatha, J. K., Santella, R. M., and Morabia, A. (2011). Significant differences in global genomic dna methylation by gender and race/ethnicity in peripheral blood. *Epigenetics*, 6:623–629. 125

[Zhang et al., 2021] Zhang, L., Young, J. I., Gomez, L., Silva, T. C., Schmidt, M. A., Cai, J., Chen, X., Martin, E. R., and Wang, L. (2021). Sex-specific dna methylation differences in alzheimer's disease pathology. *Acta Neuropathologica Communications 2021 9:1*, 9:1–19. 125

[Zhang and Sirard, 2021] Zhang, Y. and Sirard, M.-A. (2021). Epigenetic inheritance of acquired traits through dna methylation. *Animal Frontiers*, 11(6):19–27. 138

[Zhang et al., 2018] Zhang, Z. M., Lu, R., Wang, P., Yu, Y., Chen, D., Gao, L., Liu, S., Ji, D., Rothbart, S. B., Wang, Y., Wang, G. G., and Song, J. (2018). Structural basis for dnmt3a-mediated de novo dna methylation. *Nature*, 554:387–391. 52

[Zhao et al., 2018] Zhao, B., Vo, H. Q., Johnston, F. H., and Negishi, K. (2018). Air pollution and telomere length: A systematic review of 12,058 subjects. *Cardiovascular Diagnosis and Therapy*, 8:480–492. 43, 63

[Zhong et al., 2017] Zhong, J., Karlsson, O., Wang, G., Li, J., Guo, Y., Lin, X., Zemplenyi, M., Sanchez-Guerra, M., Trevisi, L., Urch, B., Speck, M., Liang, L., Coull, B. A., Koutrakis, P., Silverman, F., Gold, D. R., Wu, T., and Baccarelli, A. A. (2017). B vitamins attenuate the epigenetic effects of ambient fine particles in a pilot human intervention trial. *Proceedings of the National Academy of Sciences of the United States of America*, 114:3503–3508. 61, 65

[Zhou et al., 2019] Zhou, G., He, T., Huang, H., Feng, F., Liu, X., Li, Z., Zhang, Y., and Ba, Y. (2019). Prenatal ambient air pollution exposure and sod2 promoter methylation in maternal and cord blood. *Ecotoxicology and Environmental Safety*, 181:428–434. 49

[Zhou and Levy, 2007] Zhou, Y. and Levy, J. I. (2007). Factors influencing the spatial extent of mobile source air pollution impacts: A meta-analysis. *BMC Public Health*, 7:89. 98

[Zhu and Boutros, 2021] Zhu, C. and Boutros, P. C. (2021). Sex differences in cancer genomes: Much learned, more unknown. *Endocrinology*, 162. 111

[Zhu et al., 2017] Zhu, Z., Meng, W., Liu, P., Zhu, X., Liu, Y., and Zou, H. (2017). Dna hypomethylation of a transcription factor binding site within the promoter of a gout risk gene nrbp1 upregulates its expression by inhibition of tfap2a binding. *Clinical epigenetics*, 9(1):1–9. 162