Identification and characterisation of tissue- and ethnicity-independent hypervariable DNA methylation loci in humans

Maria Derakhshan

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of London

August 2023

Department of Population Health

Faculty of Epidemiology & Population Health

London School of Hygiene & Tropical Medicine

Funded by the Biotechnology and Biological Sciences Research Council

Research group affiliations: MRC International Nutrition Group,

MRC Unit The Gambia at LSHTM

Statement of own work

I, Maria Derakhshan, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

[redacted]

15th August 2023

Acknowledgements

I would like to thank my primary supervisor at LSHTM, Dr Matt Silver, to whom I am immensely grateful for his unwavering support throughout the PhD. Matt's expertise, kindness and generosity with his time have been inspirational and I feel very lucky to have been supervised by him. I have learned so much from Matt, and his mentoring and unfailing pastoral support has meant a great deal to me.

I would also like to thank my second supervisor Dr Garrett Hellenthal for his expertise, support and unflappable approach to research and supervision, which again have greatly shaped my development as a researcher. From the outset Garrett welcomed me as a full member of his group at UCL and I am very grateful for my experiences at the UCL Genetics Institute.

I would like to thank Dr Michael Imbeault for being a fantastic collaborator. Michael's enthusiasm for KRAB zinc finger proteins is infectious and our regular meetings at the University of Cambridge have been a highlight of my PhD.

Many thanks go to several colleagues who have been particularly engaged with my research and have helped shape it: Dr Prachand Issarapu, Noah Kessler, and Professor Andrew Prentice.

I am also grateful to all those who generated data that was used in this thesis. Much of the data is publicly available which I believe is a testament to the openness of the research community.

My heartfelt thanks to my colleagues and friends at LSHTM, the LIDo DTP and the UCL Genetics Institute, who have each enriched my PhD journey: Kevin, Julius, Sophie, Greg, Aine, Hanadi, Merlin, Nancy, Olga, Stephanie, Diana and Georgina.

Finally, my biggest thanks go to my parents, Alexandra and Mir, for their belief in me throughout the highs and lows of this journey. Thank you for teaching me resilience, and for encouraging me to be brave and to pursue my goals.

In memory of my beloved grandfather, Alexander Poulovassilis, a wonderful scientist, educator and mentor.

Abstract

Interindividual variation in DNA methylation (DNAm) has been linked to both environmental exposures and phenotypic variation. However, DNAm is generally tissue-specific, making the identification of environmentally sensitive variable DNAm states in disease-relevant tissues a challenge.

First discovered in isogenic mice, metastable epialleles (MEs) are DNAm loci that vary between individuals but are correlated across tissues derived from different germ layers within a given individual. This property, termed systemic interindividual variation or 'SIV', is attributed to stochastic methylation establishment in the early embryo. In humans, the ME definition can be broadened to include SIV loci that are influenced but not determined by genotype. Human MEs show sensitivity to the periconceptional environment and have been associated with metabolism, cancer, immunity, and neurocognitive function, making them useful for exploring how the environment experienced in early development can influence health outcomes in later life via epigenetic mechanisms using easily accessible tissues.

We analysed DNA methylation data from 30 datasets comprising 3,474 individuals, 19 tissues and 8 ethnicities at CpGs covered by the Illumina450K array. We identified 4,143 hypervariable CpGs ('hvCpGs') with methylation in the top 5% most variable sites across multiple tissues and ethnicities. hvCpG methylation was influenced but not determined by genetic variation, and was not linked to probe reliability, epigenetic drift, age, sex or cell heterogeneity effects. Instead, hvCpGs overlapped or colocalised with many previously identified MEs, showed unique methylation patterns in monozygotic (MZ) twins, and were enriched for loci previously associated with the periconceptional environment, suggesting that DNAm establishment at many hvCpGs occurs in the early embryo. hvCpGs also showed SIV in foetal tissues at previously unreported MEs, suggesting that the identification of hvCpGs may be a powerful ME detection method compared to existing SIV screens that rely on rare multi-tissue methylation datasets with small sample sizes.

Variable methylation at murine MEs has been recently linked to Krüppel-associated box (KRAB) domain proteins (KZFPs), that epigenetically silence transposable elements (TEs) in embryonic stem cells (ESCs) by recruiting KAP1 (KRAB-associated protein-1). By examining

hvCpG proximity to retrotransposons, KAP1 binding sites, and associations with genetic variation at KZFP genes *in trans*, we found evidence that KZFP-mediated silencing of TEs may influence methylation at some hvCpGs. We identified an example of how this mechanism may affect phenotype in later life: a differentially methylated region comprising hvCpGs in the promoter of *C8orf31* which is influenced by ZNF808-mediated silencing of the retrotransposon MER11C and shows evidence of a causal effect on pancreatic beta cell function.

In summary, our findings: i) position hvCpGs as age-stable biomarkers of early stochastic and/or environmental effects on DNA methylation, useful for exploring the developmental origins of health and disease; and ii) highlight KZFP-mediated silencing of TEs as a novel mechanism by which variable DNAm loci established in the early human embryo can influence phenotype in later life.

Table of Contents

Statement of own work 2
Acknowledgements 3
Abstract 5
Table of Contents 7
List of Figures 11
Abbreviations 12
Glossary of key terms 14

1. Introduction 16
1.1 Background to DNA methylation 17
1.1.1 DNA methylation as an epigenetic mechanism 17
1.1.2 Reprogramming of DNAm in early development 19
1.1.3 DNAm in epigenetic epidemiology 21
1.2 DNAm and the Developmental Origins of Health and Disease (DOHaD) 22
1.2.1 The DOHaD hypothesis 22
1.2.2 Influence of DNAm patterning in early development on later health 22
1.2.3 Metastable epialleles 24
1.3 Tissue- and ethnicity- independent hypervariable DNAm loci 26
1.4 PhD aim and research questions 28
1.5 PhD structure 29
1.6 PhD outputs 30
1.7 PhD timeframe and funding 31
1.8 References 32

2. Metastable Epialleles in Humans 43
2.1 Introduction 45
2.2 Identifying putative human metastable epialleles 46
2.2.1 Systemic Interindividual Variation (SIV) 46
2.2.2 The influence of genotype on SIV 48
2.2.3 A definition for 'putative human ME' 50
2.3 Sensitivity to the environment in early development 51

2.3.1 A natural experiment in rural Gambia: a role for 1C metabolism 51
2.3.2 Artificial Reproductive Technologies (ART) 53
2.3.3 Beyond periconception 54

2.4 Genomic context 56
2.4.1 Transposable elements 56
2.4.2 Parent-of-origin specific methylation (PofOm) and ZFP57 binding sites 57

2.5 Putative MEs and the developmental origins of health and disease 58
2.5.1 Putative ME methylation can predict disease 58
2.5.2 Putative MEs are amongst top hit loci in EWAS 59
2.5.3 Putative MEs as mediators of environmental risk on phenotype 59
2.5.4 Developmental plasticity as an adaptive mechanism 61

2.6 Conclusion 64
2.7 References 65

3. Tissue- and ethnicity- hypervariable DNA methylation states show evidence of establishment in the early human embryo 75

4. KZFP-mediated silencing of transposable elements as a mechanism influencing hypervariable methylation in early development 98
4.1 Background 100
4.2 Methods 104
4.2.1 CpG sets 104
4.2.2 Retrotransposon nomenclature and genomic coordinates 104
4.2.3 Retrotransposon evolutionary age 105
4.2.4 Retrotransposon CpG density 105
4.2.5 KRAB-zinc finger protein (KZFP) and KAP1 binding sites 105
4.2.6 Regions of parent-of-origin specific methylation (PofOm) 106
4.2.7 hvCpG proximity to features of interest 106
4.2.8 Histone modification analysis 106
4.2.9 Association between hvCpG methylation and genetic variation at KZFP genes 107
4.2.10 Characterising KZFP mQTLs 107

4.3 Results 108
4.3.1 hvCpGs are enriched for proximity to certain retrotransposon subfamilies 108
4.3.2 Coverage of retrotransposons on the Illumina450K array 109
4.3.3 mQTLs of hvCpGs are enriched in KZFP genes 110
4.3.4 Linking KZFP *trans* mQTLs to KZFP expression and activity 112

4.3.5 hvCpGs are enriched for proximal KAP1 binding sites and KAP1-bound retrotransposons 113
4.3.6 Histone modifications in embryonic stem cells 114
4.3.7 hvCpGs are enriched for proximal binding sites of KZFPs linked PofOm 116
4.3.8 Investigating hvCpG subsets based on notable features 116

4.4 Discussion 117

4.5 References 121

5. A hypervariably methylated region upstream of *C8orf31* is linked to ZNF808-mediated silencing of MER11C and has a causal effect on insulin dynamics 128

5.1 Background 130

5.2 Methods 132

5.2.1 Study cohorts 132
5.2.2 Methylation data 133
5.2.3 Genotype data 133
5.2.4 CpG sets 134
5.2.5 ZNF808 binding sites and MER11C locations 135
5.2.6 Methylation quantitative trait loci (mQTL) associated with hvCpGs 135
5.2.7 Association between the *C8orf31* DMR and glucose metabolism outcomes 136
5.2.8 Two-sample Mendelian randomisation (2SMR) 136
5.2.9 Phenome-wide Two-Sample Mendelian randomisation 138
5.2.10 Association between the *C8orf31* DMR and thyroid function measurements 138
5.2.11 Association between the *C8orf31* DMR and 1C-metabolism maternal biomarkers 139
5.2.12 *C8orf31* gene expression and methylation in thyroid tissue 139
5.2.13 BLAST analysis 139
5.2.14 Characterisation of mQTLs 140

5.3 Results 141

5.3.1 A DMR upstream of *C8orf31* shows high methylation variation and is associated with fasting insulin levels and insulin secretion in Gambian children 141
5.3.2 The *C8orf31* DMR is associated with ZNF808 genetic variation in Gambian children 143
5.3.3 Characterising ZNF808 mQTLs and the ZNF808 binding site mQTL 147
5.3.4 The *C8orf31* DMR shows evidence of a causal effect on beta cell function 148
5.3.5 Phenome-wide Two-Sample MR with *C8orf31* DMR CpGs 152
5.3.6 Association between offspring *C8orf31* DMR methylation and periconceptional levels of maternal 1C metabolism biomarkers 152
5.3.7 *C8orf31* expression is highest in thyroid and influenced by methylation 154
5.3.8 BLASTn analysis of *C8orf31* 155

5.4 Discussion 157

5.5 References **162**

6. Discussion **168**

6.1 Summary of findings **169**

6.2 Significance of findings **172**

6.2.1 Relevance to epigenetic epidemiology 172

6.2.2 Insights into fundamental biological mechanisms driving methylation variation 174

6.2.3 The role of DNAm in human adaptation and evolution 175

6.3 Limitations **177**

6.4 Future research **181**

6.4.1 Functional characterisation of the DMR upstream of *C8orf31* 181

6.4.2 Further probe KZFP-mediated silencing of TEs as a mechanism influencing methylation variation in early development 182

6.4.3 *C8orf31*: the first of many examples? 182

6.4.4 Beyond KZFP-mediated silencing of TEs? 183

6.4.5 hvCpGs on chromosome X 183

6.4.6 Screen for methylation variance QTLs 'mvQTLs' 184

6.5 Conclusion **184**

6.6 References **185**

Appendix A **191**

Supplementary Figures for Chapter 3 ..... **191**

Appendix B **205**

Supplementary Figures for Chapter 4 ..... **205**

Supplementary Tables for Chapter 4 ..... **217**

Appendix C **226**

Supplementary Figures for Chapter 5 ..... **226**

Supplementary Tables for Chapter 5 ..... **235**

List of Figures

Figure 1.1. Fundamental epigenetic mechanisms. 18
Figure 1.2. Reprogramming of DNA methylation in early embryogenesis and gametogenesis. 20
Figure 1.3. DNAm in the Developmental Origins of Health and Disease (DOHaD). 23
Figure 1.4. The Agouti mouse model showing how stochastic establishment of DNAm in early development results in variable expressivity of fur colour between isogenic A^{vy}/a mice. 25
Figure 1.5. Schematic of hypothesis that tissue- and ethnicity- independent hypervariable CpGs ('hvCpGs') may have been established in the early human embryo. 27
Figure 2.1. Model linking systemic interindividual variation (SIV) to the pre-gastrulation embryo. 46
Figure 2.2. Model of developmental origins of epigenetic supersimilarity (ESS). 49
Figure 2.3. A schematic of 1C metabolism. 53
Figure 4.1. KZFP-mediated epigenetic silencing via the induction of heterochromatin formation and DNA methylation. 100
Figure 4.2. Model hypothesising that variable binding of KZFPs to IAP elements results in stochastic setting of methylation states at the cellular level, driving interindividual methylation variation at metastable epialleles. 102
Figure 4.3. Enrichment for proximal retrotransposon subfamilies at hvCpGs. 109
Figure 4.4. hvCpGs are enriched for *trans* mQTLs located in KZFP genes relative to the array background. 111
Figure 4.5. hvCpGs are enriched for proximal KAP1 binding sites and H3K9me3 modifications in H1-ESCs relative to the array background. 115
Figure 5.1. Overview of data available for ENID and EMPHASIS cohorts of Gambian children. 132
Figure 5.2. Schematic of *C8orf31* gene region and key features. 142
Figure 5.3. The DMR upstream of *C8orf31* is associated with fasting insulin and insulin secretion in the EMPHASIS cohort. 143
Figure 5.4. Rationale for testing the association between the *C8orf31* DMR and genetic variation at *ZNF808 in trans* and at the ZNF808 binding site *in cis* 144
Figure 5.5. Association between *C8orf31* DMR and genotype at ZNF808 gene and at ZNF808 binding ChIP-exo peak in MER11C. 147
Figure 5.6. Two-sample Mendelian randomisation (2SMR) framework exploring the causal link between DMR methylation and phenotypes linked to glucose metabolism. 149
Figure 5.8. Testing the association between the *C8orf31* DMR and maternal 1C metabolism biomarkers in the ENID cohort. 153
Figure 5.9. Expression and methylation of *C8orf31* in thyroid. 155
Figure 5.10. Schematic of *C8orf31* and *Ly6S* genes. 156
Figure 5.11. Summary of findings. 161

Abbreviations

CpG	Cytosine-phosphate-Guanine
ChIP-exo	Chromatin immunoprecipitation with exonuclease digestion
CI	Confidence Interval
DMR	Differentially Methylated Region
DNA	Deoxyribonucleic acid
DNAm	DNA methylation
DNMT	DNA methyltransferase
DOHaD	Developmental Origins of Health and Disease
DZ	Dizygotic
eQTL	Expression Quantitative Trait Locus
ERV	Endogenous Retrovirus
ESS	Epigenetic Supersimilarity
evCpG	equivalently variable CpG
EWAS	Epigenome Wide Association Study
FDR	False Discovery Rate
FET	Fishers Exact Test
GEO	Gene Expression Omnibus
GWAS	Genome Wide Association Study
HEK293T	Human Embryonic Kidney 293T cells
hvCpGs	hypervariable CpGs
IAP	Intracisternal A Particle
KAP1	KRAB-associated protein 1
KRAB	Krüppel-associated box
KZFP	KRAB Zinc Finger Protein

LD	Linkage Disequilibrium
LINE	Long Interspersed Repeat Element
LTR	Long Terminal Repeat
MAF	Minor Allele Frequency
ME	Metastable Epiallele
Methyl-Seq	Methylation Sequencing
mQTL	Methylation Quantitative Trait Locus
MR	Mendelian Randomisation
MZ	Monozygotic
OR	Odds Ratio
PC	Principal Component
PofOm	Parent-of-origin specific methylation
RNA	Ribonucleic Acid
SINE	Short Interspersed Repeat Element
SIV	Systemic Interindividual Variation
SNP	Single Nucleotide Polymorphism
SoC	Season of Conception
TCGA	The Cancer Genome Atlas
TE	Transposable Element
TET	Ten-Eleven Translocation
TF	Transcription Factor
TSS	Transcription Start Site
1C	1-carbon metabolism
1000G	1000 Genomes
5mC	5-methylcytosine
H3K9me3	Histone 3 Lysine 9 trimethylation

Glossary of key terms

Developmental Origins of Health and Disease:	The hypothesis proposed by David Barker and colleagues that environmental exposures experienced in early development can influence disease risk in later life.
DNA methylation:	The covalent bonding of a methyl group (CH_3) to the 5' carbon position of cytosine at a CpG dinucleotide.
Differentially Methylation Region:	Regions at which methylation levels are associated with an exposure or phenotype of interest.
Epigenetics:	The study of mitotically and/or meiotically heritable genome modifications that influence gene expression without altering underlying DNA sequence.
Epigenetic epidemiology:	The study of the associations between interindividual epigenetic variation and disease risk.
Epigenetic supersimilarity:	CpGs that show variability between unrelated individuals, but greater-than-expected similarity between monozygotic (MZ) co-twins than between dizygotic (DZ) co-twins, indicating that methylation states at these loci may have been established before MZ twinning events in the cleavage-stage embryo.
Systemic interindividual variation:	CpGs with this property show methylation variation between individuals but methylation correlation across tissues derived from different germ layers within a given individual, suggestive of their establishment in the pre-gastrulation embryo.

Metastable Epiallele (ME):	A SIV locus at which methylation varies between isogenic mice, indicative of stochastic methylation establishment in the early embryo.
Putative human ME:	A SIV locus at which methylation variability is not determined by genotype but that may be influenced by genetics effects.
Gastrulation:	The process by which the blastocyst forms the gastrula. The blastocyst is comprised of the trophoblast and the inner cell mass (ICM). After implantation, the ICM forms a two-layered structure called the bilaminar disk that comprises the epiblast and the hypoblast. At gastrulation, the epiblast cells are rearranged and undergo differentiation to form the three germ cell layers: the endoderm, mesoderm, and ectoderm.
KRAB zinc finger protein:	A large family of transcription factors characterised by an N-terminal KRAB domain and C-terminal DNA binding domain of zinc fingers. These proteins are involved in the epigenetic regulation of transposable elements via DNA methylation and H3K9me3.
Transposable element:	DNA sequences that occupy nearly half of the human genome and that have the potential to move location in the genome.

Chapter 1

1. Introduction

Summary of chapter

This chapter introduces the background to the thesis, covering the main biological concepts relating to DNA methylation and the motivation for identifying and characterising tissue- and ethnicity- independent hypervariable CpGs in the human genome. The chapter introduces our key research questions and the rationale behind these questions. This chapter also describes the structure of the thesis and gives details on PhD timeframe and funding.

1.1 Background to DNA methylation

1.1.1 DNA methylation as an epigenetic mechanism

The term 'epigenetics' can be defined as the study of mitotically (and potentially meiotically) heritable changes in gene expression potential that do not affect DNA at the sequence level[1]. Whilst this is a commonly accepted definition, other definitions do not include a requirement for heritability[2]. The main mechanisms of epigenetic gene regulation include DNA methylation (DNAm), post-translational histone modifications and non-coding regulatory RNAs[3] (**Figure 1.1**). Together, these mechanisms allow genetically identical cells to have a diverse range of gene expression profiles and are critical for cell fate determination and differentiation[4,5].

DNAm is the most well-studied epigenetic mechanism. DNA methyltransferase enzymes (DNMTs) recognise CpG dinucleotides within the genome, where they catalyse the covalent bonding of a methyl group ($-CH_3$) donated by S-adenosyl-l-methionine (SAM), to the 5th carbon of the cytosine pyrimidine ring. There are approximately 28 million CpG sites in the human genome[6], 20% of the expected dinucleotide frequency because 5-methylcytosine (5mC) is susceptible to spontaneous deamination to thymine[7,8]. CpGs are not distributed evenly and are enriched within 'CpG islands' (CGIs), defined as regions of length 500-2000 bp that have a GC content > 50%[9]. The majority of CGIs are found in or near promoters of genes[10] and are largely hypomethylated to promote an open chromatin state that is accessible to transcription factors[11,12]. Methylation at CGIs can mediate long-term silencing of gene expression by inhibiting transcription factor binding at promoter regions and enhancer regions, recruiting chromatin-condensing machinery via Methyl-CpG-binding domain proteins, and altering nucleosome dynamics[13–16].

DNAm-mediated gene silencing underpins genomic imprinting[17,18], X-chromosome inactivation[19,20], silencing of repeat sequences such as transposable elements[21] and pericentromeric satellite repeats[22,23] and cellular differentiation[4,24]. Aberrations in this silencing can lead to genomic imprinting disorders[25], cancers[26], autoimmune diseases[27], and metabolic disorders[28]. Although considered a silencing mechanism, DNAm within gene bodies has been shown to positively correlate with gene expression[29], alter mRNA alternative splicing patterns[30,31]

and influence transcriptional elongation[32]. DNAm therefore regulates gene expression in a context specific manner.

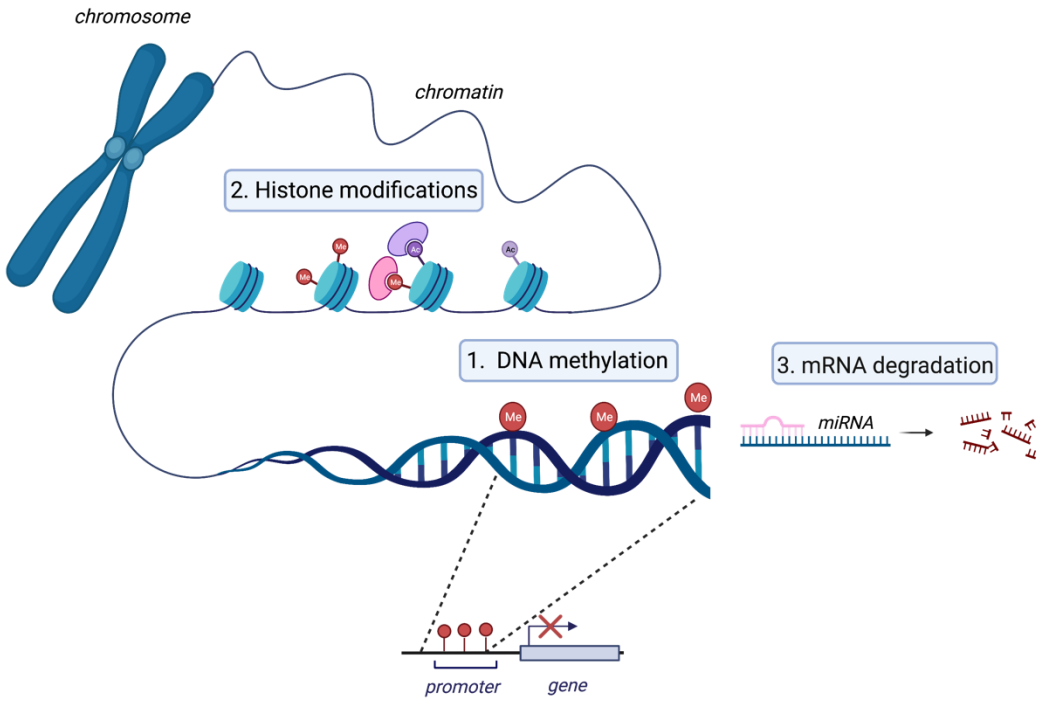


Figure 1.1. Fundamental epigenetic mechanisms. 1) DNA methylation involves the covalent bonding of a methyl group (red) to the cytosine in CpG dinucleotides. Methylated promoters are usually associated with gene silencing. **2)** Posttranslational modifications of the amino-acid tails of histones can influence chromatin structure. For example, methylation of histone 3 lysine 9 (H3K9me3) is associated with heterochromatin, whilst acetylation of histone 3 lysine 4 (H3K4) is associated with an open chromatin environment that is accessible to transcription factors. **3)** Non-coding RNAs such as microRNAs (miRNAs) can regulate gene expression by degrading mRNA transcripts or preventing their translation. Figure created with BioRender.com.

1.1.2 Reprogramming of DNAm in early development

DNAm is considered a 'stable' epigenetic mechanism because DNAm marks are faithfully copied between parent and daughter cells during mitotic cell division by DNMT1, which recognises hemi-methylated DNA and is recruited to the parental strand through its binding partner, Np95[33,34]. However, DNAm patterns are erased and re-established at each generation shortly after fertilisation and in developing gametes[35] (**Figure 1.2**).

The first reprogramming event occurs shortly after fertilisation. The paternal and maternal genomes are extensively de-methylated in order to establish a pluripotent state in the early embryo[35]. The paternal genome undergoes rapid active demethylation via the action of TET (ten-eleven translocation) proteins, that catalyse the stepwise oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)[36,37]. In contrast, the maternal genome is passively de-methylated through successive cell divisions due to the nuclear exclusion of DNMT1[38,39]. Imprinting control regions (ICRs), that are differentially methylated between the egg and sperm, escape this erasure so that parent-of-origin specific methylation (PofOm) is retained at imprinted genes in the developing embryo[40,41]. In mice, some retrotransposons such as Intracisternal A Particles (IAPs) and certain LINE-1 and ERV-1 elements can partially escape this erasure[42,43]. When the blastocyst implants into the uterine epithelium, methylation states are re-established in a cell-type specific manner by the *de novo* methyltransferases DNMT3A and DNMT3B whose activity is dependent on the catalytically inactive cofactor DNMT3L[44,45]. The exact mechanism of site-specific *de novo* methylation is not fully understood, but certain *cis*-acting sequence features can influence local DNAm patterning[46], and CpG islands are protected from DNAm as they are marked with H3K4 modifications recruited by transcription factors and RNA polymerase[47].

The second wave of re-programming occurs during gametogenesis, in primordial germ cells (PGCs), as they migrate to the genital ridge (the future gonads) of the post-implantation embryo[40]. During this wave, pre-existing methylation patterns including parental imprints in the PGCs are erased so that, following sex determination of the embryo, new methylation states can be established in either the developing spermatogonia or oocytes[40]. De-methylation of the PGCs is achieved through a combination of both active and passive mechanisms[35]. Only a handful of loci are resistant to this erasure, most of which reside in human specific transposable elements[48–

[50] and at IAP elements in mice[51]. *De novo* methylation in sperm is completed before birth, whilst in oocytes *de novo* methylation occurs after birth and is largely restricted to gene bodies in a transcription-dependent manner[40,52].

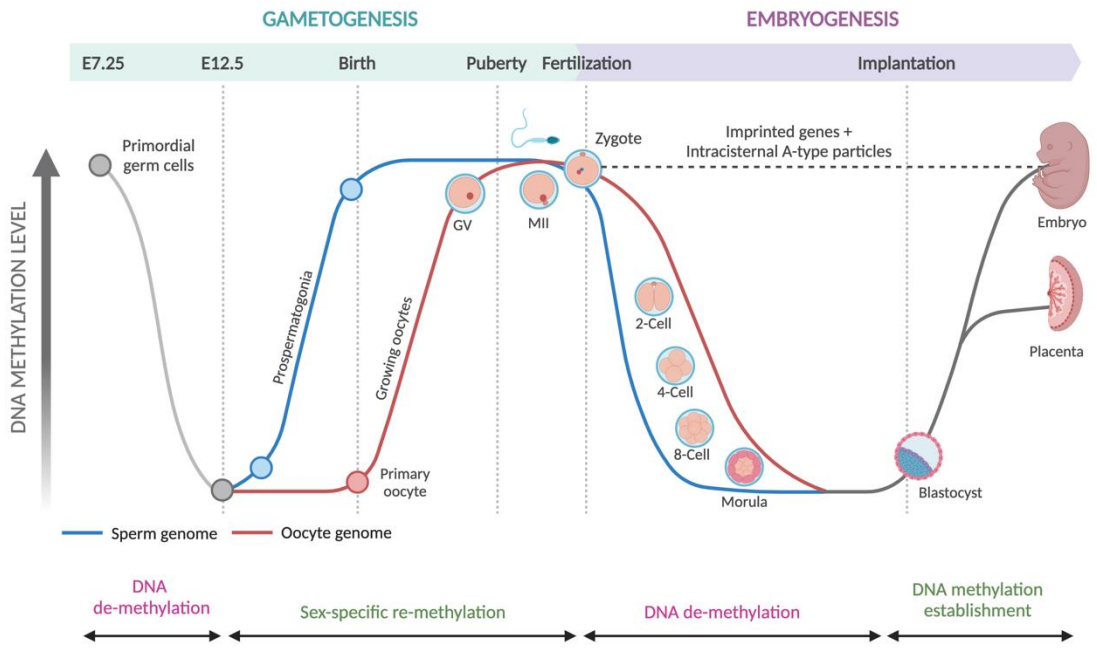


Figure 1.2. Reprogramming of DNA methylation in early embryogenesis and gametogenesis. Two periods of DNAm reprogramming occur in early development. During gametogenesis, methylation states in primordial germ cells are erased and re- established in a sex-specific manner. In males, methylation patterns in sperm are largely established before birth. In females, methylation re-establishment occurs after birth and is largely completed by the time that developing oocyte reaches the germinal vesicle stage (GV) at puberty. These states are maintained at the meiosis II (MII) stage during ovulation. Shortly after fertilisation, the maternal and paternal genomes are de-methylated, and DNAm levels reach their lowest in the preimplantation embryo. DNAm states are re-established upon implantation into the uterine epithelium. Adapted from "DNA Methylation Levels During Mammalian Development" by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

1.1.3 DNAm in epigenetic epidemiology

DNAm patterns are highly cell-type and tissue-specific[53,54]. DNAm has a key role in cell identity and differentiation by silencing pluripotency genes upon implantation and regulating the expression and repression of lineage-specific genes[4]. As a result, inter-tissue DNAm variation is more common than inter-individual variation[55]. However, CpGs showing inter-individual variation do exist and are being increasingly studied as potential drivers of phenotypic variation due to the ability of methylation to influence gene expression[56].

Sources of interindividual methylation variation include genetic variation[57], pre-and post-natal environmental exposures[58], and stochastic effects that arise in early development[59] and throughout the life-course[60,61]. Studies using monozygotic (MZ) and dizygotic (DZ) twin datasets to determine the proportion of methylation variance that is explained by additive genetic effects (A), common environment (C) and non-shared environment (E) have found that methylation variation is primarily driven by the non-shared environment[62–64]. DNAm is therefore particularly useful for studying the aetiology of non-heritable and complex traits.

The discovery of associations between DNAm and phenotype and environmental exposures has been accelerated by Epigenome-Wide Association Studies (EWAS) that use high-throughput array-based technologies to profile genome-wide DNAm states[65]. However, there are key drawbacks to these studies. Firstly, the cell-type and tissue-specificity of DNAm means that methylation states measured in easily accessible tissues such as blood and saliva are not necessarily reflective of those in pathologically relevant tissues[66]. Secondly, associations between DNAm and a given trait/exposure can be confounded by factors including differences in cell-type proportions between individuals when using bulk tissues[67], sex[68,69], age[70], and technical artefacts introduced by the array technology[71], which need to be properly accounted for in association models. Thirdly, the pathology of the disease can itself influence methylation, making the direction of causality difficult to interpret[66]. Fourthly, the arrays focus on CpGs located in promoter and enhancer regions, and therefore only cover a fraction of the CpG sites genome-wide[72,73] , giving an incomplete picture of methylation variability.

1.2 DNAm and the Developmental Origins of Health and Disease (DOHaD)

1.2.1 The DOHaD hypothesis

The Developmental Origins of Health and Disease (DOHaD) hypothesis posits that the environment experienced in early life has a causal effect on health outcomes in later life[74]. Barker and colleagues formulated this hypothesis based on observations that low birth weight is associated with adverse health outcomes such as high blood pressure and cardiovascular disease[75–78]. Many associations between adverse gestational environments and health outcomes have been subsequently discovered in humans. For example, in the Dutch Hunger Winter cohort, prenatal famine exposure (November 1944 – April 1945) is associated with elevated blood pressure, body mass index, and increased risk of schizophrenia in later life[79–82]. However, in some cases these associations may be driven by confounding effects (such as genetics or early environment) that are difficult to rule out when using observational data[83,84].

1.2.2 Influence of DNAm patterning in early development on later health

An increasing body of evidence suggests that epigenetic mechanisms may underpin some DOHaD observations[85,86]. DNAm is postulated to be particularly sensitive to environmental and stochastic effects during the periconceptional period, which spans gamete maturation, fertilisation, and early embryogenesis[87]. These effects may alter DNAm patterns by interfering with demethylation, re-methylation, and methylation maintenance mechanisms. Perturbed methylation states can be stably maintained between cell generations through DNMT1, and impact offspring phenotype in later life[87] (**Figure 1.3**). Various maternal and foetal exposures during gestation such as dietary folate[88–91], gestational famine[92–97], alcohol[98–100], aflatoxins[101] and smoking[102–105] have been associated with offspring DNAm. In several cases, DNAm changes induced by such exposures are associated with offspring phenotype[94,96,99]. Studies using isogenic mouse models have also shown that DNAm patterns can be established stochastically (that is, independently of genotype and environmental effects) during this time-window and have the potential to drive phenotypic variation in later life[106,107].

Developmental plasticity may act as an adaptive mechanism, whereby sensitivity to the early developmental environment could prime the offspring to the anticipated future environment[85]. Where there is mismatch between these pre-and post-natal environments', this adaptive mechanism may instead manifest as disease[85,108]. Alternatively, instead of being an adaptive mechanism, associations with disease status may be reflective of epigenetic errors in the establishment and maintenance of DNAm[109]. More recently, associations with adverse environments have been argued to reflect the selection of stochastically established epigenomic profiles that favour embryo survival *in utero*[110].

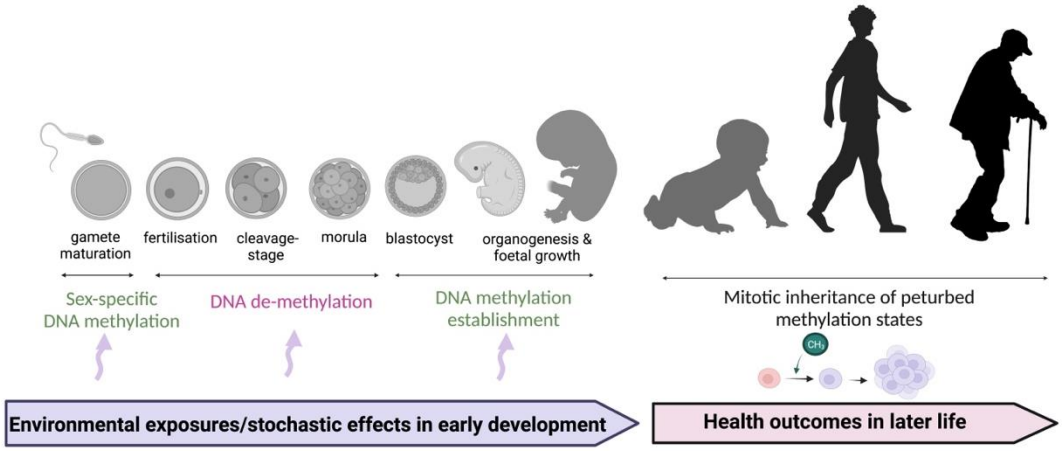


Figure 1.3. DNAm in the Developmental Origins of Health and Disease (DOHaD). Environmental exposures and stochastic effects in early development may influence DNAm patterns by interfering with DNA de-methylation and re-methylation events in the gametes and early embryo. These methylation states can be mitotically inherited from daughter cells to progeny, with potential consequences on gene expression and disease risk in later life.

1.2.3 Metastable epialleles

Metastable epialleles (MEs) are promising candidate loci for studying the epigenetic mechanisms underpinning DOHaD. First discovered in isogenic mice, MEs are defined as DNAm states that vary between individuals *independently of genetic variation*, but that are correlated across tissues derived from different germ layers within a given individual. This property is termed 'systemic interindividual variation' (SIV) and is attributed to stochastic methylation establishment in the early embryo before gastrulation[106]. MEs have also been defined as alleles that show variable expression between genetically identical individuals as a result of variable methylation profiles[111], although the consensus definition does not include a requirement for an influence on gene expression[112,107].

Murine MEs have been identified at Intracisternal A Particle (IAP) elements, a type of murine-specific retrotransposon. In the paradigm Agouti viable yellow (A^{vy}) mouse model, the degree of methylation at an IAP element upstream of the Agouti gene is associated with gene expression, resulting in isogenic littermates with a spectrum of fur colours and metabolic phenotypes linked to obesity and diabetes[106] (**Figure 1.4**). Offspring methylation states at several characterised murine MEs can be 'shifted' by supplementing dams with methyl donors during early gestation[113,114], and exposing dams to endocrine disruptors[115] and toxins[116]. However, environmental sensitivity at murine MEs has been shown to be the exception rather than the rule[117].

The identification of MEs in humans is challenging because human populations are genetically heterogeneous. A growing body of evidence supports the existence of *putative human MEs*, defined as SIV loci that are *influenced but not determined* by genotype[118]. Like their murine counterparts, putative human MEs are enriched for proximity to transposable elements[118,119] and their methylation status has been associated with various phenotypes, including traits linked to metabolism[28,120–122], cancer[123,124], immunity[125], and neurocognitive disorders[126,127]. Putative human MEs have also been associated with environmental exposures including season-of-conception[128–130], gestational famine[131], and maternal gestational folate supplementation[89], indicating that in some instances they could mediate the effects of the early developmental environment on phenotype in later life.

Whilst the interplay between genetic variation, stochastic effects, and environmental effects on methylation variability at putative human MEs is unknown, these loci have two key properties that position them as attractive candidates for studying epigenetic mechanisms underpinning DOHaD. Firstly, the timing of methylation establishment is constrained to the pre-gastrulation embryo, a period that is postulated to be highly sensitive to stochastic and environmental perturbation[87]. Secondly, unlike the majority of DNAm states that are cell-type and tissue specific, their methylation states are correlated across tissues, meaning that associations between putative human MEs and a given phenotype/exposure identified in accessible tissues such as blood are reflective of those in inaccessible but potentially pathologically more relevant ones.

Despite their utility, methods to identify putative human MEs to date have relied on datasets comprising methylation data from multiple tissues for each individual in order to detect SIV (**Figure 1.5**). These datasets are difficult-to-acquire and have therefore had small sample sizes, rendering putative ME screens underpowered to detect interindividual variation. For example, in the largest SIV screen to date, with 10 individuals, the well-established putative human ME *VTRNA2-1* (*Vault RNA2-1*) did not show interindividual variation[132].

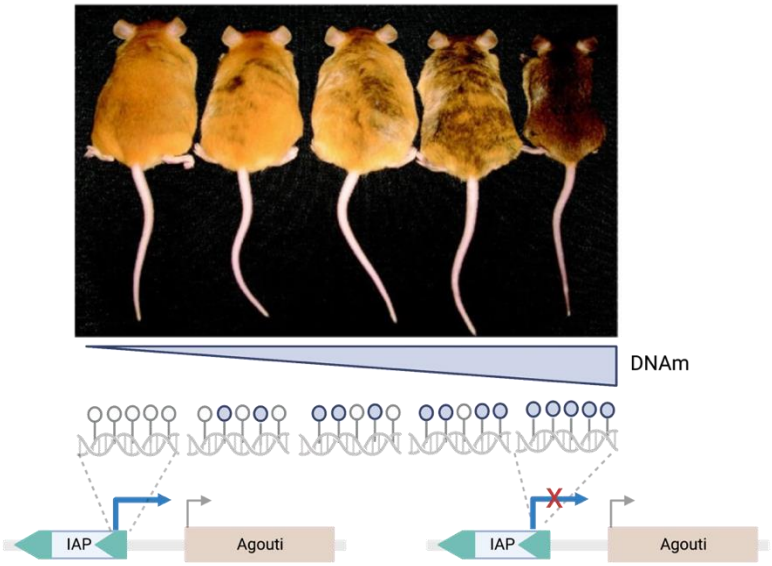


Figure 1.4. The Agouti mouse model showing how stochastic establishment of DNAm in early development results in variable expressivity of fur colour between isogenic A^{vy}/a mice. An IAP element (a class of murine-specific retrotransposon) upstream of the *Agouti* gene contains a long terminal repeat (LTR). Methylation states at this LTR are established stochastically before germ-layer differentiation. Hypomethylation of the IAP element leads to ectopic expression of the *Agouti* gene, resulting in yellow obese mice (far left). Hypermethylation of the IAP element silences ectopic expression, resulting in mice with a pseudo agouti phenotype (far right). Figure adapted from Dolinoy et al. (2006) and created with BioRender.com.

1.3 Tissue- and ethnicity- independent hypervariable DNAm loci

Growing evidence suggests that common drivers of DNAm such as genetic variation, pre-and post-natal environment, sex, and age can influence DNAm in a tissue-specific manner[61,133–136]. DNAm states also differ between human populations due to population-specific genetic and/or environmental effects[137–144]. In this thesis, we reason that CpGs showing consistently high inter-individual variation in a large number of datasets that vary by tissue-type and ethnicity would be robust to tissue-specific and ethnicity-specific drivers of methylation variation, thereby revealing insights into fundamental mechanisms driving methylation variation. Noting that tissue-independent methylation variation has been observed at putative human MEs (at which methylation is also correlated across tissues), we hypothesise that DNAm states at tissue- and ethnicity- independent hypervariable CpGs ('hvCpGs') may have been established in the early human embryo, and therefore share features with putative human MEs. If so, the identification of hvCpGs using large, publicly available datasets might be a high-powered method for detecting putative MEs compared to existing approaches (**Figure 1.5**).

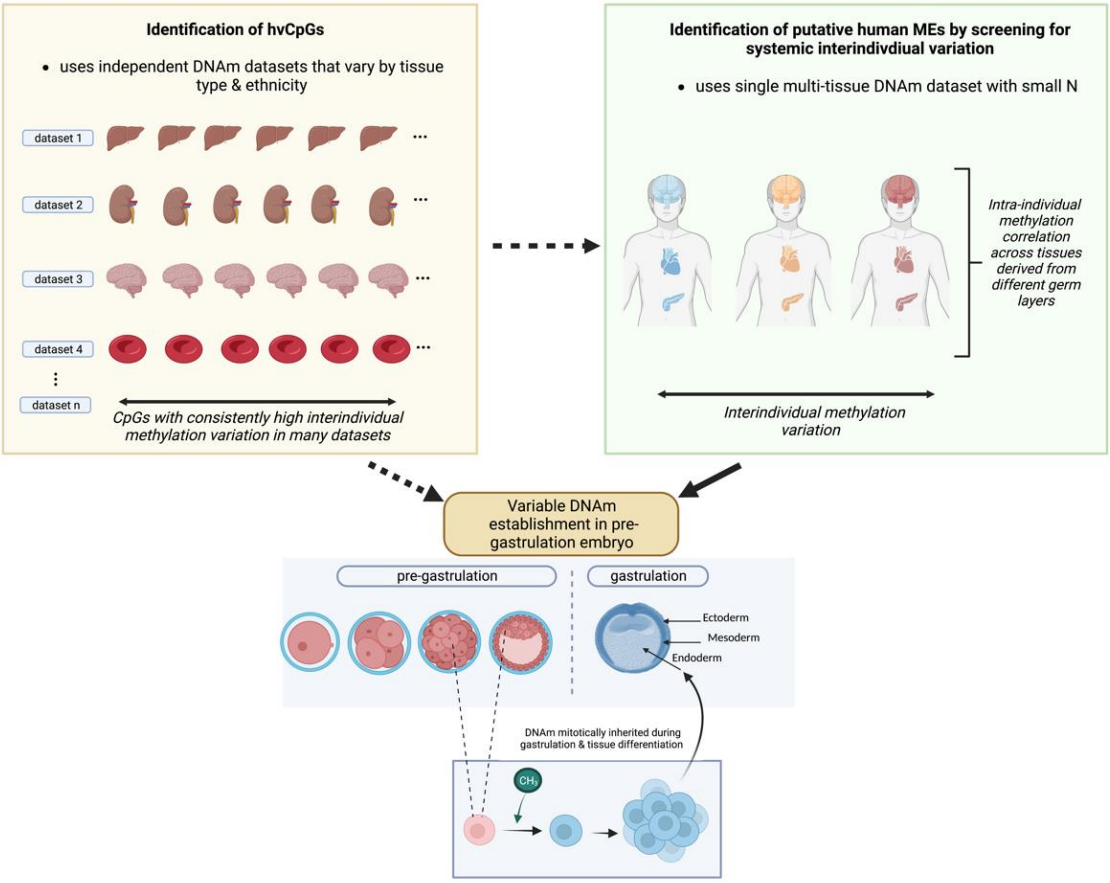


Figure 1.5. Schematic of hypothesis that tissue- and ethnicity- independent hypervariable CpGs ('hvCpGs') may have been established in the early human embryo. Green box: putative human MEs are identified by screening for *systemic interindividual variation (SIV)*, where DNAm varies between individuals but is correlated across tissues derived from different germ layers within a given individual. This approach requires methylation data from several tissues for each individual. To-date, SIV screens have therefore had small sample sizes. Yellow box: in this thesis, we aim to identify CpGs that show inter-individual methylation variation in many diverse datasets that vary by tissue type and ethnicity. If these exist, we hypothesise that they may have been established in the early human embryo. If so, the detection of hvCpGs may be a more high-powered approach to detect putative human MEs than SIV screens.

1.4 PhD aim and research questions

The overall aim of this thesis is to identify and characterise tissue- and ethnicity- independent hypervariable CpGs, termed 'hvCpGs', in the human genome. This research centres on three key research questions:

Q1. Do hypervariable CpGs ('hvCpGs') exist in the human genome?

Many studies have examined variable DNAm states using data from multiple tissues. The main aims of these studies have been to i) identify CpGs showing SIV in order to find putative human MEs[118,124]; ii) determine whether variable DNAm states in peripheral tissues reflect those in inaccessible tissues[145,146]; iii) identify CpGs that are differentially methylated between tissues[147,148]; or iv) examine the tissue-specificity of common drivers of methylation variation[61,133–136]. To our knowledge, no studies have examined whether DNAm states that are highly variable in a large number of diverse tissues and ethnicities exist.

Q2. Do hvCpGs show evidence of establishment in the early human embryo?

As discussed in **1.3**, we hypothesise that methylation variability at hvCpGs may have been established in the early human embryo. If so, this would position hvCpGs as useful loci for studying how stochastic and environmental effects in early development may influence phenotype in later life.

Q3. Can we identify molecular mechanisms underpinning the establishment of variable methylation states in the early embryo and can we link these to postnatal phenotype?

Murine ME methylation variation is linked to incomplete epigenetic silencing of retrotransposon elements by KRAB zinc finger proteins[149], a class of transcription factor that usually silences transposable elements by depositing chromatin condensing H3K9me3 histone modifications and DNAm[150]. Whilst putative human MEs are enriched for proximity to retrotransposons, reminiscent of murine MEs, specific molecular mechanisms influencing interindividual methylation variation in the early human embryo have not been directly explored. In this thesis, we explore the role of KZFP-mediated silencing of TEs in influencing methylation variability at hvCpGs and investigate potential links to phenotype in later life.

1.5 PhD structure

This is a book style thesis comprising six chapters. One chapter is a published paper, in accordance with thesis regulations at the London School of Hygiene and Tropical medicine. All chapters are preceded with a 'Chapter Summary' which gives an overview of the chapter content. Chapters 2-5 also have a 'Chapter Rationale', which describes the motivation for the research presented in that chapter and its context within the thesis. The Supplementary Figures and Tables for Chapters 3-5 are given in Appendices A-C respectively. The six chapters are as follows:

Chapter 1: A brief overview of DNA methylation, DOHaD, metastable epialleles, key research questions and rationale, thesis structure, and timeframe and funding for this PhD.

Chapter 2: A literature review entitled 'Metastable Epialleles in Humans' describing the methods used to identify putative human MEs to date, their key features, associations with disease, and their relevance to DOHaD. This review also introduces key concepts and terminologies that are used throughout the thesis.

Chapter 3: A published research paper entitled 'Tissue- and ethnicity- independent hypervariable CpGs show evidence of establishment in the early human embryo'. This paper details the method for identifying hvCpGs and explores evidence that these CpGs are established in the early embryo.

Chapter 4: Investigation into epigenetic silencing of transposable elements (TEs) by KRAB zinc finger proteins (KZFPs) in the early embryo as a mechanism influencing methylation at hvCpGs.

Chapter 5: Identification of a differentially methylated region in the promoter of *C8orf31* that is linked to ZNF808-mediated silencing of MER11C and shows evidence of a causal association with insulin secretion and T1 diabetes in humans.

Chapter 6: Summary of the overall PhD findings in relation to the three key research questions, and discussion of the significance of these findings, limitations, and avenues for future research.

1.6 PhD outputs

Published paper:

- **Derakhshan, M.**, Kessler, N.J., Ishida, M., Demetriou, C., Brucato, N., Moore, G.E., Fall, C.H.D, Chandak, G.R., Ricaut, F.X., Prentice, A.M, Hellenthal, G., and Silver, M.J., 2022. Tissue-and ethnicity-independent hypervariable DNA methylation states show evidence of establishment in the early human embryo. *Nucleic Acids Research*, *50*(12), pp.6735-6752. https://doi.org/10.1093/nar/gkac503.

Conference talks and poster presentations:

- 'High inter-individual variability at metastable epialleles is conserved across multiple human populations'
  - Wellcome Epigenomics of Common Disease, November 2019 (poster)

- 'Tissue and ethnicity independent hypervariable DNA methylation states show evidence of establishment in the early human embryo'
  - Epihub New Horizons in Epigenomics, September 2020 (15-min talk)
  - Wellcome Epigenomics of Common Diseases, November 2021 (flash talk and poster)
  - European Society of Human Genetics, Vienna, June 2022 (poster)
  - Developmental Origins of Health and Disease, Vancouver, August 2022 (poster)

Invited seminar presentation:

- 'KZFP-mediated silencing of TEs: linking DNAm states established in the early embryo to phenotype in later life'
  - 'Dark Matter of the Genome' club, Department of Genetics, Cambridge University, November 2022 (30-min talk)

Other contributions:

- Silver, M.J., Saffari, A., Kessler, N.J., Chandak, G.R. Fall, C.H.D, Issarapu, P., Dedaniya, A,. Betts, M., Moore, S.E., Routledge, M.N., Herceg, Z., Cuenin, C.**, Derakhshan, M.**, James, P.T., Monk, D., and Prentice, A.M., 2021, *Elife*. Environmentally sensitive hotspots in the methylome of the early human embryo.
- Dawson, L.F., Peltier, J., Hall, C.L., Harrison, M.A., **Derakhshan, M**., Shaw, H.A., Fairweather, N.F. and Wren, B.W., 2021, *Scientific reports*. Extracellular DNA, cell surface proteins and c-di-GMP promote biofilm formation in Clostridioides difficile.
- Jordans, E., **Derakhshan, M.**, Rutter, Z., de Villiers S., and Ng'weno, B., 2022, *International Leadership, Effecting Success Across Boarders*. Developing Global Leadership in Africa*:* Addressing Young Leaders' Ambitions for a People- and Community-Oriented Approach
- Issarapu, P., Arumalla, M., Elliott, H.R., Nongmaithem, S.S., Alagu, S., Betts, M., Sajjadi, S., Kessler, N.J., Bayyana, S., Mansuri, S.R., **Derakhshan, M**., Krishnaveni, G.V., Shrestha, S., Kumaran, K., Di Gravio, C., Sahariah, S.A., Sanderson, E., Relton, C.L., Ward, K.A, Moore, S.E., Prentice, A.M., Lillycrop, K.A., Fall., C.H.D, Silver, M.J., Chandak R.J., and EMPHASIS study group, 2022, *bioRxiv*. DNA methylation at the suppressor of cytokine signaling 3 (SOCS3) gene influences height in childhood.

1.7 PhD timeframe and funding

This PhD was undertaken full-time from September 2019 – December 2022 and was funded by the British Biotechnology Research Council (BBSRC) (BB/M009513/1) as part of the London Interdisciplinary Doctoral training programme (LIDo). 3 months (January 2020 – March 2020) were spent undertaking a mandatory Professional Internship for PhD Students (PIPS) unrelated to the PhD research.

1.8 References

1. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*. 2003;33(S3):245-254.
2. Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease. *Nature*. 2019;571(7766):489-499.
3. Gibney ER, Nolan CM. Epigenetics and gene expression. *Heredity (Edinb)*. 2010;105(1):4-13.
4. Suelves M, Carrió E, Núñez-Álvarez Y, Peinado MA. DNA methylation dynamics in cellular commitment and differentiation. *Brief Funct Genomics*. 2016;15(6):443-453.
5. Atlasi Y, Stunnenberg HG. The interplay of epigenetic marks during stem cell differentiation and development. *Nat Rev Genet*. 2017;18(11):643-658.
6. Lövkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Res*. 2016;44(11):5123-5132.
7. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921.
8. Hermann A, Gowher H, Jeltsch A. Biochemistry and biology of mammalian DNA methyltransferases. *Cellular and Molecular Life Sciences*. 2004;61(19):2571-2587.
9. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*. 1987;196(2):261-282.
10. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*. 2006;103(5):1412-1417.
11. Blackledge NP, Thomson JP, Skene PJ. CpG island chromatin is shaped by recruitment of ZF-CxxC proteins. *Cold Spring Harb Perspect Biol*. 2013;5(11):a018648.
12. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev*. 2011;25(10):1010-1022.
13. Harteis S, Schneider S. Making the Bend: DNA Tertiary Structure and Protein-DNA Interactions. *Int J Mol Sci*. 2014;15(7):12335.
14. Hendrich B, Bird A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol*. 1998;18(11):6538-6547.
15. Nan X, Ng HH, Johnson CA, et al. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*. 1998;393(6683):386-389.

16. Li S, Peng Y, Landsman D, Panchenko AR. DNA methylation cues in nucleosome geometry, stability and unwrapping. *Nucleic Acids Res*. 2022;50(4):1864-1874.

17. Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. *Nature*. 1993;366(6453):362-365.

18. Bartolomei MS. Genomic imprinting: employing and avoiding epigenetic processes. *Genes Dev*. 2009;23(18):2124-2133.

19. Lock LF, Takagi N, Martin GR. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell*. 1987;48(1):39-46.

20. Mohandas T, Sparkes RS, Shapiro LJ. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science (1979)*. 1981;211(4480):393-396.

21. Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet*. 1998;20(2):116-117.

22. Lehnertz B, Ueda Y, Derijck AAHA, et al. Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Current Biology*. 2003;13(14):1192-1200.

23. Chen T, Tsujimoto N, Li E. The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin. *Mol Cell Biol*. 2004;24(20):9048-9058.

24. Xie W, Schultz MD, Lister R, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*. 2013;153(5):1134-1148.

25. Elhamamsy AR. Role of DNA methylation in imprinting disorders: an updated review. *J Assist Reprod Genet*. 2017;34(5):549-562.

26. Koch A, Joosten SC, Feng Z, et al. Analysis of DNA methylation in cancer: location revisited. *Nat Rev Clin Oncol*. 2018;15(7):459-466.

27. Ballestar E, Sawalha AH, Lu Q. Clinical value of DNA methylation markers in autoimmune rheumatic diseases. *Nat Rev Rheumatol*. 2020;16(9):514-524.

28. van Dijk SJ, Tellam RL, Morrison JL, Muhlhausler BS, Molloy PL. Recent developments on the role of epigenetics in obesity and metabolic disease. *Clin Epigenetics*. 2015;7(1):1-13.

29. Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell*. 2014;26(4):577-590.

30. Maunakea AK, Nagarajan RP, Bilenky M, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*. 2010;466(7303):253-257.

31. Maunakea AK, Chepelev I, Cui K, Zhao K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res.* 2013;23(11):1256-1269.

32. Lorincz MC, Dickerson DR, Schmitt M, Groudine M. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol.* 2004;11(11):1068-1075.

33. Arand J, Spieler D, Karius T, et al. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet.* 2012;8(6):e1002750.

34. Sharif J, Muto M, Takebayashi S ichiro, et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature.* 2007;450(7171):908-912.

35. Messerschmidt DM, Knowles BB, Solter D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev.* 2014;28(8):812-828.

36. He YF, Li BZ, Li Z, et al. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science (1979).* 2011;333(6047):1303-1307.

37. Ito S, Shen L, Dai Q, et al. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science (1979).* 2011;333(6047):1300-1303.

38. Hirasawa R, Chiba H, Kaneda M, et al. Maternal and zygotic Dnmt1 are necessary and sufficient for the maintenance of DNA methylation imprints during preimplantation development. *Genes Dev.* 2008;22(12):1607-1616.

39. Branco MR, Oda M, Reik W. Safeguarding parental identity: Dnmt1 maintains imprints during epigenetic reprogramming in early embryogenesis. *Genes Dev.* 2008;22(12):1567-1571.

40. Smallwood SA, Kelsey G. De novo DNA methylation: A germ cell perspective. *Trends in Genetics.* 2012;28(1):33-42.

41. Barlow DP, Bartolomei MS. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol.* 2014;6(2):a018382.

42. Lane N, Dean W, Erhardt S, et al. Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *genesis.* 2003;35(2):88-93.

43. Guibert S, Forné T, Weber M. Global profiling of DNA methylation erasure in mouse primordial germ cells. *Genome Res.* 2012;22(4):633-641.

44. Suetake I, Shinozaki F, Miyagawa J, Takeshima H, Tajima S. DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. *Journal of Biological Chemistry.* 2004;279(26):27816-27823.

45. Jia D, Jurkowska RZ, Zhang X, Jeltsch A, Cheng X. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature*. 2007;449(7159):248-251.

46. Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schübeler D. Identification of genetic elements that autonomously determine DNA methylation states. *Nat Genet*. 2011;43(11):1091-1097.

47. Greenberg MVC, Glaser J, Borsos M, et al. Transient transcription in the early embryo sets an epigenetic state that programs postnatal growth. *Nat Genet*. 2017;49(1):110-118.

48. Tang WWC, Dietmann S, Irie N, et al. A unique gene regulatory network resets the human germline epigenome for development. *Cell*. 2015;161(6):1453-1467.

49. Dietmann S, Keogh MJ, Tang W, et al. Transposable elements resistant to epigenetic resetting in the human germline are epigenetic hotspots for development and disease. *bioRxiv*. Published online. doi:10.1101/2020.03.19.998930

50. Iouranova A, Grun D, Rossy T, et al. KRAB zinc finger protein ZNF676 controls the transcriptional influence of LTR12-related endogenous retrovirus sequences. *Mob DNA*. 2022;13(1):4.

51. Seisenberger S, Andrews S, Krueger F, et al. The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. *Mol Cell*. 2012;48(6):849-862.

52. Stewart KR, Veselovska L, Kelsey G. Establishment and functions of DNA methylation in the germline. *Epigenomics*. 2016;8(10):1399-1413.

53. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. Published online 2015. doi:10.1038/nature14248

54. Ziller MJ, Gu H, Müller F, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*. 2013;500(7463):477-481.

55. Byun HM, Siegmund KD, Pan F, et al. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum Mol Genet*. 2009;18(24):4808-4817.

56. Murphy TM, Mill J. Epigenetics in health and disease: heralding the EWAS era. *The Lancet*. 2014;383(9933):1952-1954.

57. Villicaña S, Bell JT. Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biol*. 2021;22(1).

58. Martin EM, Fry RC. Environmental influences on the epigenome: exposure-associated DNA methylation in human populations. *Annu Rev Public Health*. 2018;39(1):309-333.

59. Kazachenka A, Bertozzi TM, Sjoberg-Herrera MK, et al. Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell*. 2018;175(5):1259-1271.e13.

60. Fraga MF, Ballestar E, Paz MF, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences*. 2005;102(30):10604-10609.

61. Slieker RC, Relton CL, Gaunt TR, Slagboom PE, Heijmans BT. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics Chromatin*. 2018;11(1).

62. Hannon E, Knox O, Sugden K, et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. Greally JM, ed. *PLoS Genet*. 2018;14(8):e1007544.

63. van Dongen J, Nivard MG, Willemsen G, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat Commun*. 2016;7(1):1-13.

64. Busche S, Shao X, Caron M, et al. Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation The Multiple Tissue Human Expression Resource. *Genome Biol*. 2015;16(1):1-18.

65. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*. 2011;12(8):529-541.

66. Birney E, Smith GD, Greally JM. Epigenome-wide Association Studies and the Interpretation of Disease -Omics. Barsh GS, ed. *PLoS Genet*. 2016;12(6):e1006105-e1006105.

67. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*. 2014;15(2):R31.

68. Inoshita M, Numata S, Tajima A, et al. Sex differences of leukocytes DNA methylation adjusted for estimated cellular proportions. *Biol Sex Differ*. 2015;6(1):11.

69. Yousefi P, Huen K, Davé V, Barcellos L, Eskenazi B, Holland N. Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. *BMC Genomics*. 2015;16:911.

70. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2013;14(10):1-20.

71. Price EM, Robinson WP. Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned. *Front Genet*. 2018;0(MAR):83.

72. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288-295.

73. Mansell G, Gorrie-Stone TJ, Bao Y, et al. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics*. 2019;20(1):366.

74. Barker DJP. The Developmental Origins of Adult Disease. *J Am Coll Nutr*. 2004;23(6 Suppl):588S-595S.

75. Barker DJP, Osmond C. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. *The Lancet*. 1986;327(8489):1077-1081.

76. Barker DJ, Osmond C, Golding J, Kuh D, Wadsworth M. Growth in utero, blood pressure in childhood and adult life, and mortality from cardiovascular disease. *Br Med J*. 1989;298(6673):564-567.

77. Barker DJP, Godfrey KM, Gluckman PD, Harding JE, Owens JA, Robinson JS. Fetal nutrition and cardiovascular disease in adult life. *The Lancet*. 1993;341(8850):938-941.

78. Ravelli ACJ, van der Meulen JHP, Osmond C, Barker DJP, Bleker OP. Obesity at the age of 50 y in men and women exposed to famine prenatally. *Am J Clin Nutr*. 1999;70(5):811-816.

79. Susser E, Neugebauer R, Hoek HW, et al. Schizophrenia after prenatal famine further evidence. *Arch Gen Psychiatry*. 1996;53(1):25-31.

80. Stein AD, Kahn HS, Rundle A, Zybert PA, van der Pal-De Bruin K, Lumey LH. Anthropometric measures in middle age after exposure to famine during gestation: Evidence from the Dutch famine. *American Journal of Clinical Nutrition*. 2009;85(3):869-876.

81. Roseboom TJ, van der Meulen JHP, Ravelli ACJ, et al. Blood pressure in adults after prenatal exposure to famine. *J Hypertens*. 1999;17(3):325-330.

82. Roseboom T, de Rooij S, Painter R. The Dutch famine and its long-term consequences for adult health. *Early Hum Dev*. 2006;82(8):485-491.

83. Gage SH, Munafò MR, Davey Smith G. Causal inference in developmental origins of health and disease (DOHaD) research. *Annu Rev Psychol*. 2016;67:567-585.

84. Donofrio B, Class Q, Lahey B, Larsson H. Testing the developmental origins of health and disease hypothesis for psychopathology using family-based, quasi-experimental designs. *Child Dev Perspect*. 2014;8(3):151-157.

85. Gluckman PD, Hanson MA, Low FM. The role of developmental plasticity and epigenetics in human health. *Birth Defects Res C Embryo Today*. 2011;93(1):12-18.

86. Waterland RA, Michels KB. Epigenetic Epidemiology of the Developmental Origins Hypothesis. *Annu Rev Nutr*. 2007;27(1):363-388.

87. Fleming TP, Watkins AJ, Velazquez MA, et al. Origins of lifetime health around the time of conception: causes and consequences. *The Lancet*. 2018;391(10132):1842-1852.

88. Joubert BR, den Dekker HT, Felix JF, et al. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. *Nat Commun*. 2016;7:10577.

89. Amarasekera M, Martino D, Ashley S, et al. Genome-wide DNA methylation profiling identifies a folate-sensitive region of differential methylation upstream of ZFP57-imprinting regulator in humans. *The FASEB Journal*. 2014;28(9):4068-4076.

90. Steegers-Theunissen RP, Obermann-Borst SA, Kremer D, et al. Periconceptional maternal folic acid use of 400 μg per day is related to increased methylation of the IGF2 gene in the very young child. *PLoS One*. 2009;4(11).

91. Richmond RC, Sharp GC, Herbert G, et al. The long-term impact of folic acid in pregnancy on offspring DNA methylation: Follow-up of the Aberdeen folic acid supplementation trial (AFAST). *Int J Epidemiol*. 2018;47(3):928-937.

92. Heijmans BT, Tobi EW, Stein AD, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci*. 2008;105(44):17046-17049.

93. Tobi EW, Lumey LH, Talens RP, et al. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum Mol Genet*. 2009;18(21):4046-4053.

94. Tobi EW, Goeman JJ, Monajemi R, et al. DNA methylation signatures link prenatal famine exposure to growth and metabolism. *Nat Commun*. 2014;5(1):5592.

95. Tobi EW, Slieker RC, Luijk R, et al. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Sci Adv*. 2018;4(1):eaao4364.

96. Shen L, Li C, Wang Z, et al. Early-life exposure to severe famine is associated with higher methylation level in the IGF2 gene and higher total cholesterol in late adulthood: The Genomic Research of the Chinese Famine (GRECF) study. *Clin Epigenetics*. 2019;11(9):1-9.

97. Wang Z, Song J, Li Y, Dong B, Zou Z, Ma J. Early-Life Exposure to the Chinese Famine Is Associated with Higher Methylation Level in the INSR Gene in Later Adulthood. *Sci Rep*. 2019;9(1):1-9.

98. Loke YJ, Muggli E, Nguyen L, et al. Time-and sex-dependent associations between prenatal alcohol exposure and placental global DNA methylation. *Epigenomics*. 2018;10(7):981-991.

99. Frey S, Eichler A, Stonawski V, et al. Prenatal alcohol exposure is associated with adverse cognitive effects and distinct whole-genome DNA methylation patterns in primary school children. *Front Behav Neurosci*. 2018;12:125.

100. Mandal C, Halder D, Jung KH, Chai YG. Gestational alcohol exposure altered DNA methylation status in the developing fetus. *Int J Mol Sci*. 2017;18(7):1386.

101. Hernandez-Vargas H, Castelino J, Silver MJ, et al. Exposure to aflatoxin B1 in utero is associated with DNA methylation in white blood cells of infants in The Gambia. *Int J Epidemiol*. 2015;44(4):1238-1248.

102. Joubert BR, Håberg SE, Nilsen RM, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect*. 2012;120(10):1425-1431.

103. Markunas CA, Xu Z, Harlid S, et al. Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy. *Environ Health Perspect*. 2014;122(10):1147-1153.

104. Richmond RC, Simpkin AJ, Woodward G, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum Mol Genet*. 2015;24(8):2201-2217.

105. Lee K, Kunkeaw N, Jeon SH, et al. Precursor miR-886, a novel noncoding RNA repressed in cancer, associates with PKR and modulates its activity. *RNA*. 2011;17(6):1076-1089.

106. Rakyan VK, Blewitt ME, Druker R, Preis JI, Whitelaw E. Metastable epialleles in mammals. *Trends in Genetics*. 2002;18(7):348-351.

107. Kazachenka A, Bertozzi TM, Sjoberg-Herrera MK, et al. Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell*. 2018;175(5):1259-1271.e13.

108. Gluckman PD, Hanson MA, Beedle AS. Early life events and their consequences for later disease: a life history and evolutionary perspective. *American journal of human biology*. 2007;19(1):1-19.

109. Heijmans BT, Tobi EW, Lumey LH, Slagboom PE. The epigenome: Archive of the prenatal environment. *Epigenetics*. 2009;4(8):526-531.

110. Tobi EW, van den Heuvel J, Zwaan BJ, Lumey LH, Heijmans BT, Uller T. Selective Survival of Embryos Can Explain DNA Methylation Signatures of Adverse Prenatal Environments. *Cell Rep*. 2018;25(10):2660-2667.e4.

111. Dolinoy DC, Das R, Weidman JR, Jirtle RL. Metastable Epialleles, Imprinting, and the Fetal Origins of Adult Diseases. *Pediatr Res*. 2007;61(5 Part 2):30R-37R.

112. Bertozzi TM, Ferguson-Smith AC. Metastable epialleles and their contribution to epigenetic inheritance in mammals. *Semin Cell Dev Biol*. 2020;97:93-105.

113. Cooney CA, Dave AA, Wolff GL. Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. In: *Journal of Nutrition*. 2002;132(8):2393S-2400S

114. Wolff GL, Kodell RL, Moore SR, Cooney CA. Maternal Epigenetics and Methyl Supplements Affect Agouti Gene Expression in A vy /a Mice. *The FASEB Journal*. 1998;12(11):949-957.

115. Dolinoy DC, Huang D, Jirtle RL. Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. *Proceedings of the National Academy of Sciences*. 2007;104(32):13056-13061.

116. Kaminen-Ahola N, Ahola A, Maga M, et al. Maternal Ethanol Consumption Alters the Epigenotype and the Phenotype of Offspring in a Mouse Model. *PLoS Genet*. 2010;6(1):e1000811

117. Bertozzi TM, Becker JL, Blake GET, et al. Variably methylated retrotransposons are refractory to a range of environmental perturbations. *Nat Genet*. 2021;53(8):1233-1242.

118. Kessler NJ, Waterland RA, Prentice AM, Silver MJ. Establishment of environmentally sensitive DNA methylation states in the very early human embryo. *Sci Adv*. 2018;4(7):eaat2624.

119. Silver MJ, Kessler NJ, Hennig BJ, et al. Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome Biol*. 2015;16(1).

120. Kühnen P, Handke D, Waterland RA, et al. Interindividual Variation in DNA Methylation at a Putative POMC Metastable Epiallele Is Associated with Obesity. *Cell Metab*. 2016;24(3):502-509.

121. Marttila S, Viiri LE, Mishra PP, et al. Methylation status of nc886 epiallele reflects periconceptional conditions and is associated with glucose metabolism through nc886 RNAs. *Clin Epigenetics*. 2021;13(1):143.

122. Clark J, Martin E, Bulka CM, et al. Associations between placental CpG methylation of metastable epialleles and childhood body mass index across ages one, two and ten in the Extremely Low Gestational Age Newborns (ELGAN) cohort. *Epigenetics*. 2019;14(11):1102-1111.

123. van Baak TE, Coarfa C, Dugué PA, et al. Epigenetic supersimilarity of monozygotic twin pairs. *Genome Biol*. 2018;19(1):2.

124. Silver MJ, Kessler NJ, Hennig BJ, et al. Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome Biol*. Published online 2015.

125. Mok A, Rhead B, Holingue C, et al. Hypomethylation of CYP2E1 and DUSP22 Promoters Associated With Disease Activity and Erosive Disease Among Rheumatoid Arthritis Patients. *Arthritis and Rheumatology*. 2018;70(4):528-536.

126. Zhu Y, Mordaunt CE, Yasui DH, et al. Placental DNA methylation levels at CYP2E1 and IRS2 are associated with child outcome in a prospective autism study. *Hum Mol Genet*. 2019;28(16):2659-2674.

127. Boks MP, Houtepen LC, Xu Z, et al. Genetic vulnerability to DUSP22 promoter hypermethylation is involved in the relation between in utero famine exposure and schizophrenia. *NPJ Schizophr*. 2018;4(1):1-8.

128. Dominguez-Salas P, Moore SE, Baker MS, et al. Maternal nutrition at conception modulates DNA methylation of human metastable epialleles. *Nat Commun*. 2014;5(1):3746.

129. Waterland RA, Kellermayer R, Laritsky E, et al. Season of Conception in Rural Gambia Affects DNA Methylation at Putative Human Metastable Epialleles. Whitelaw E, ed. *PLoS Genet*. 2010;6(12):e1001252.

130. James PT, Dominguez-Salas P, Hennig BJ, Moore SE, Prentice AM, Silver MJ. Maternal One-Carbon Metabolism and Infant DNA Methylation between Contrasting Seasonal Environments: A Case Study from The Gambia. *Curr Dev Nutr*. 2019;3(1).

131. Finer S, Iqbal MS, Lowe R, et al. Is famine exposure during developmental life in rural Bangladesh associated with a metabolic and epigenetic signature in young adulthood? A historical cohort study. *BMJ Open*. 2016;6(11):e011768.

132. Gunasekara CJ, Scott CA, Laritsky E, et al. A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol*. 2019;20(1):105.

133. Busche S, Shao X, Caron M, et al. Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. *Genome Biol*. 2015;16(1):1-18.

134. Smith AK, Kilaru V, Kocak M, et al. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*. 2014;15(1):145.

135. Lin D, Chen J, Perrone-Bizzozero N, et al. Characterization of cross-tissue genetic-epigenetic effects and their patterns in schizophrenia. *Genome Med*. 2018;10(1):13.

136. Day K, Waite LL, Thalacker-Mercer A, et al. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol*. 2013;14(9).

137. Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. *Genome Biol*. 2012;13(2):R8.

138. Carja O, MacIsaac JL, Mah SM, et al. Worldwide patterns of human epigenetic variation. *Nat Ecol Evol*. 2017;1(10):1577-1583.

139. Moen EL, Zhang X, Mu W, et al. Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics*. 2013;194(4):987-996.

140. Galanter JM, Gignoux CR, Oh SS, et al. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. *Elife*. 2017;6:e20532.

141. Rahmani E, Shenhav L, Schweiger R, et al. Genome-wide methylation data mirror ancestry information. *Epigenetics Chromatin*. 2017;10(1):1.

142. Husquin LT, Rotival M, Fagny M, et al. Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation. *Genome Biol*. 2018;19(1):222.

143. Mozhui K, Smith AK, Tylavsky FA. Ancestry Dependent DNA Methylation and Influence of Maternal Nutrition. Zhang H, ed. *PLoS One*. 2015;10(3):e0118466.

144. Giuliani C, Sazzini M, Bacalini MG, et al. Epigenetic variability across human populations: a focus on DNA methylation profiles of the KRTCAP3, MAD1L1 and BRSK2 genes. *Genome Biol Evol*. 2016;8(9):2760-2773.

145. Hannon E, Lunnon K, Schalkwyk L, Mill J. Interindividual methylomic variation across blood, cortex, and cerebellum: Implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics*. 2015;10(11):1024-1032.

146. Braun PR, Han S, Hing B, et al. Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. *Transl Psychiatry*. 2019;9(1):1-10.

147. Slieker RC, Bos SD, Goeman JJ, et al. Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics Chromatin*. 2013;6(1).

148. Davies MN, Volta M, Pidsley R, et al. Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol*. 2012;13(6):R43-R43.

149. Bertozzi TM, Elmer JL, Macfarlan TS, Ferguson-Smith AC. KRAB zinc finger protein diversification drives mammalian interindividual methylation variability. *Proceedings of the National Academy of Sciences*. 2020;117(49):31290-31300.

150. Bruno M, Mahgoub M, Macfarlan TS. The Arms Race between KRAB-Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. *Annu Rev Genet*. 2019;53:393-416.

Chapter 2

2. Metastable Epialleles in Humans

Rationale of chapter

A main hypothesis of this thesis is that hvCpGs may be established in the early human embryo, and that they may therefore share features with putative human metastable epialleles (MEs). An understanding of how human MEs have been identified to date, their key characteristics, and their relevance to the Developmental Origins of Health and Disease (DOHaD), is therefore necessary. Many of the concepts introduced in this narrative review are referred to in later chapters. Furthermore, whilst literature reviews on murine MEs have been published, there has not yet been an overview of the studies focussing on putative human MEs.

Summary of chapter

Putative human MEs are identified by screening for systemic interindividual variation (SIV), indicative of variable DNAm establishment in the pre-gastrulation embryo. Many of these loci appear to be influenced but not determined by genetic variation and are associated with certain classes of retrotransposon, reminiscent of their murine counterparts. Putative human MEs show stability with age, and can predict health outcomes relating to cancer, glucose metabolism, and thyroid function in later life. Several putative human MEs have been linked to both environmental exposures in early development and phenotypes, positioning them as mediators of early environmental risk factors on health outcomes. Although the molecular mechanisms driving methylation variability at putative human MEs remain unclear, the studies discussed in this review highlight their utility for exploring how stochastic and environmental effects in early development can influence disease risk in later life. Their correlation across diverse tissue types allows associations with phenotypes and exposures to be studied in easily accessible tissues.

2.1 Introduction

First identified in isogenic mice, metastable epialleles (MEs) are DNA methylation states that show interindividual variation between genetically identical individuals but correlation across tissues derived from different germ layers (endoderm, mesoderm, and ectoderm) within a given individual[1]. Sometimes referred to as 'systemic interindividual variation' (SIV)[2,3], this property is attributed to stochastic methylation establishment in the early embryo before germ layer differentiation. In mice, MEs occur at Intracisternal A Particle (IAP) elements, a class of long terminal repeat (LTR) retrotransposon[4]. In the paradigm Agouti viable yellow (A^{vy}) and Axin Fused (A^{Fu}) models, the degree of IAP methylation is correlated with the expression of nearby genes, resulting in isogenic littermates that show a spectrum of fur colours and degrees of tail kinking respectively. In these models, the distribution of offspring phenotypes can be shifted by supplementing the maternal diet with methyl donors[5–7] and by exposing mothers to endocrine disruptors (such as genistein and bisphenol A)[8–10] and toxins (such as ethanol and phthalates)[11,12] during early gestation.

The term *metastable epiallele* was coined to refer an allele that can exist in multiple epigenetic states ("epialleles") and can probabilistically switch between these states ("metastable")[1].

Although MEs appear to be rare in mice[4,13], their ability to drive phenotypic variation in the absence of genetic variation, including susceptibility to metabolic diseases in later life[8,14], has motivated efforts to identify similar methylation states in humans. Several overviews on murine MEs have been published[15–17]. Here, we give an overview of the key studies investigating putative *human* MEs. Specifically, we discuss the methods used to identify MEs in humans, their genomic context, evidence for their sensitivity to the early developmental environment, and associations with disease. We use the term '*putative* human ME' to highlight the difficulty of ruling out the influence of genetic variation in studies of genetically heterogeneous human samples.

2.2 Identifying putative human metastable epialleles

2.2.1 Systemic Interindividual Variation (SIV)

The identification of putative human MEs requires methylation data from tissues derived from different germ layers per individual in order to screen for CpGs that show systemic (cross-tissue) interindividual variation ('SIV'), indicative of variable methylation establishment before gastrulation. SIV loci therefore contrast the majority of CpGs which show low variability between individuals but high inter-tissue variation due to the cell-type and tissue-specificity of DNAm[18,19]. To date, six independent SIV screens have been carried out, each varying in CpG coverage, sample size, number of tissues and methylation profiling platform (summarised in **Table 2.1**). SIV loci have been identified at both the regional level[2,3,20,21] (since regions of contiguous correlated methylation are likely to be of greater functional relevance)[22] and at the single-CpG level[23,24] (**Table 2.1**). Whilst most SIV loci have been identified in adult samples, several loci have been validated in foetal tissues[25], confirming that methylation variation at these loci arises in early life.

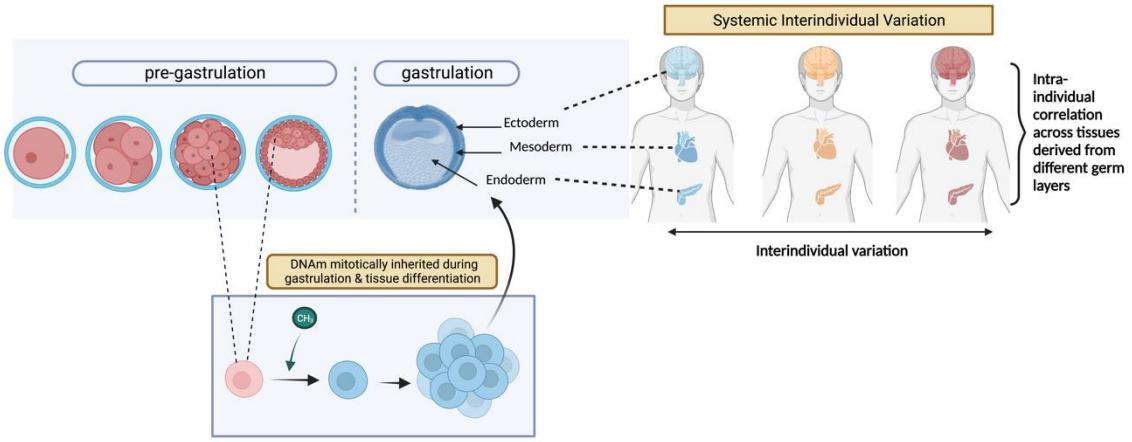


Figure 2.1. Model linking systemic interindividual variation (SIV) to the pre-gastrulation embryo. SIV loci show methylation variation between individuals but correlation across tissues derived from different germ layers within a given individual. Tissues are coloured by the methylation status of the SIV CpG. This suggests that variable DNAm states at these loci have been established before germ-layer differentiation (gastrulation), and mitotically inherited during gastrulation and tissue differentiation. Figure generated with BioRender.com.

The term 'systemic' is intended to imply that methylation states are correlated across a diverse range of cell types and tissues soma-wide. However, some SIV loci show cell-type and lineage-specific effects. For example, the well-characterised SIV regions *VTRNA2-1 (Vault RNA 2-1)* and *PAX8 (Paired box gene 8)* are not variable in the cerebellum and adult thyroid tissue respectively[26,27]. This phenomenon has been observed at the Axin^{Fu} murine ME, at which methylation in tail differs to that in other tissues[7].

Table 2.1. Summary of screens to detect systemic interindividual variation (SIV) in humans.

SIV Screen	Tissues	N	platform	Regional/ CpG-level	Method
Waterland *et al.* (2010)	Peripheral blood leukocytes (M), Hair follicles (Ec)	8	MSAM	Regional	Parallel two-tissue co-hybridisation design to detect SIV. Identified 40 *SmaI/XmaI* intervals showing SIV. Chose 13 loci for validation using pyrosequencing and confirmed interindividual variation and significant inter-tissue correlations at 5 of these.
Harris *et al.* (2014)	Peripheral blood leukocytes (M), Colonic mucosa (En)	10	Illumina450K	CpG	1776 SIV-CpGs identified by selecting CpGs for which the highest and lowest methylation beta values in each tissue appeared in the same two individuals. CpGs were then required to have an average difference in methylation (across the PBLs and colonic mucosa) between these individuals of at least 0.1 and at most 0.6 and an absolute Pearson correlation between tissues of at least 0.63.
Silver *et al.* (2015)*	Peripheral blood leukocytes (M), Hair follicles (Ec)	2	WGBS	Regional	109 SIV 'bins' were identified (where bins were 200bp in length, with at least 6 CpGs) using a 'systemic interindividual variation index' (SIVI). The SIVI was formulated to identify regions at which differences in methylation between the two individuals were concordant between PBL and HF. The threshold for SIV bins was set at SIVI $\geq$ 20.
van Baak *et al.* (2018)	Gall bladder (En), Abdominal aorta (M), Sciatic nerve (Ec)	4	Illumina450K	CpG	1042 SIV-CpGs with interindividual variation at least 3-fold greater than inter-tissue variation and with interindividual methylation range $\geq$ 0.2.
Kessler *et al.* (2018)*	Small bowel (En), Peripheral blood leukocytes (M), Hair Follicles (Ec)	5	WGBS	Regional	687 regions, each containing at least 4 SIV-CpGs and at least twice as many SIV-CpGs as non-SIV CpGs. SIV-CpGs were defined as CpGs with interindividual variation at least 3-fold greater than inter-tissue variation and with interindividual methylation range $\geq$ 0.15.
Gunasekara *et al.* (2019)	Thyroid (En), Heart (M), Brain (Ec)	10	WGBS	Regional	9926 correlated regions of SIV (coRSIVs). First identified 100 bp WGBS bins that were correlated. Then identified which of these correlated regions had inter-tissue correlation $\geq$ 0.71 and interindividual methylation range $\geq$ 0.2.

*Note, the samples used in the screens by Silver et al. (2015) and Kessler et al. (2018) partially overlap.
En = endoderm, M = mesoderm, Ec = ectoderm, MSAM = methylation-specific amplification microarray, WGBS = whole genome bisulphite sequencing.

2.2.2 The influence of genotype on SIV

A key driver of methylation variation is genetic variation[28,29], raising the possibility that SIV is driven by cross-tissue genetic effects instead of stochastic (alternatively described as random or probabilistic) and/or environmental effects in the pre-implantation embryo. Genetic effects may act at the SNP level (so-called methylation quantitative trait loci, 'mQTLs'), at the haplotype level[28,30], or may be attributed to broader genetic architecture such as topological domains. Murine MEs show variability between isogenic mice, confirming that methylation at these loci is established independently of genotype. Analysing the influence of genetics at SIV loci in heterogenous human populations is more challenging due to extensive genetic variation and has relied on the use of monozygotic (MZ) twin datasets and on examining the association between SIV loci and mQTLs.

2.2.2.1 Monozygotic twin discordance and supersimilarity

Early studies leveraged monozygotic (MZ) twin datasets to determine whether SIV loci show methylation discordance between MZ co-twins, indicative of methylation variation that is not linked to genetic variation[2,23,31]. A key advantage of this approach is that it accounts for genetic effects that may influence DNA methylation over large genomic distances[32,33]. Amongst the SIV-CpGs reported by Harris et al. (2013), 14% showed methylation differences in peripheral blood leukocyte (PBL) samples between MZ co-twins[23]. However, these twin datasets comprised adult samples from a single tissue, raising the possibility that MZ discordance may have arisen due to post-natal epigenetic drift effects[34,35]. In a recent study, Planterose-Jimenez et al.[36] identified CpG sites with strong evidence of stochastic methylation establishment in both whole blood and adipose tissue by finding loci at which methylation variation between MZ-co-twin pairs was equivalent to that between unrelated individuals. These 'equivalently variable' ('ev') CpGs were robust to epigenetic drift effects from ages 6 to 15 and were highly variable in 2-year-olds, indicating that stochastic methylation at these loci arises in early life[36]. Although methylation variability at these loci has not been definitively linked to the early embryo, SIV CpGs identified by Harris *et al.* were 10-fold enriched for evCpGs[36].

A major consideration when applying this approach is that MZ twins are not analogous to isogenic mice as they are derived from the same zygote. This means that examining MZ discordance is unsuitable for identifying variable loci that are established *before* MZ twinning events. Based on this premise, van Baak et al. identified 'epigenetic supersimilarity' ('ESS') CpGs[37] that are amongst the top 10% variable CpGs in adipose tissue and which show greater than expected similarity between MZ twins. Specifically, ESS CpGs show methylation concordance between MZ co-twins (that share 100% genetic identity) that is more than double that between dizygotic (DZ) co-twins (that share 50% genetic identity). This indicates that supersimilarity between MZ twins is not attributed to the shared genetic identity of MZ twins. Instead, because MZ twins are derived from the same zygote, van Baak et al. proposed that ESS is attributed to the timing of methylation establishment occurring before MZ twinning events in the very early embryo (**Figure 2.2**). A large proportion of SIV loci show ESS[24], indicating that methylation states at SIV loci may be established in the cleavage-stage embryo. ESS is considered necessary but not sufficient for SIV, as methylation states at ESS loci may not necessarily persist into diverse post-natal tissues[24].

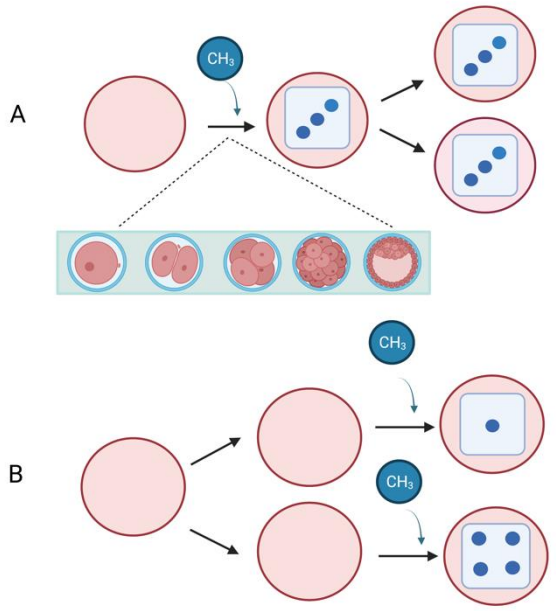


Figure 2.2. Model of developmental origins of epigenetic supersimilarity (ESS). A) If methylation establishment occurs before MZ twining, both twins will inherit the same methylation pattern, regardless of their shared genetic identity**. B)** If methylation establishment occurs after MZ twinning, this supersimilarity would be lost. Numbers on each die represent the methylation state at a given locus. Figure is based on that from van Baak et al. (2018). Created with BioRender.com.

2.2.2.2 Methylation quantitative trait loci (mQTLs)

An additional approach to investigating genetic effects at SIV CpGs in non-twin datasets is to screen for mQTLs. Because SIV screens have had small sample sizes, rendering them underpowered to detect methylation-genotype associations, most studies have used mQTL data from unrelated cohorts with larger sample sizes. SIV CpGs are enriched for mQTLs[3,20,21,24], in agreement with findings that CpGs that covary between blood and brain (indicative of SIV) are also enriched for mQTLs[38,39].

However, Kessler et al. (2018)[3] found that mQTLs explain only a moderate proportion of methylation variance at SIV loci, with a median percentage variance explained by mQTLs < 25%[3]. Furthermore, van Baak et al.[24] found that even SIV-CpGs with substantial mQTL effects (defined as explaining more than 33% of methylation variance), show variability between individuals matched for the same local sequence context[24]. Contradicting this, the majority of *correlated regions of SIV* (CoRSIVs) identified by Gunasekara et al. (2019)[21] (**Table 2.1**), were associated with large *cis* mQTL effects that explained a median 76% of methylation variance[40]. It is possible that this method for detecting SIV regions, which first groups correlated regions of the genome (**Table 2.1**), may enrich for *cis* genetic effects compared to other approaches to detect SIV loci.

2.2.3 A definition for 'putative human ME'

Together, these findings suggest that although many SIV loci are under genetic control, methylation variation at some of these loci may not be fully explained by genetic effects alone[3], raising the possibility that stochastic and/or environmental effects are at play. The definition of a metastable epiallele in genetically heterogenous human populations has therefore been extended to include SIV loci at which methylation is *influenced but not determined* by genotype[3]. However, the relative influences of genetic, environmental, and stochastic effects at SIV loci remain unclear, as the examination of long-range and multi-locus genetic effects on methylation variability is hindered by limited SNP coverage and lack of power. We use the term *putative human metastable epiallele,* to distinguish these loci from murine MEs where the influence of genotype can be ruled out.

2.3 Sensitivity to the environment in early development

Examination of environmental lability of putative human MEs has involved observational studies of cohorts subjected to naturally occurring exposures[2,3,24,31,41,42], retrospective studies of cohorts exposed to gestational famine[43,44], randomized supplementation trials[45–47], and *in vitro* approaches[43,48,49]. Although environmental sensitivity is not a hallmark feature of an ME[50], these studies support the notion that methylation variability at putative human MEs is established in early development and is not explained by genetic variation alone. Furthermore, if putative MEs are also associated with phenotype, they are promising candidates for studying the mechanisms mediating the effect of early developmental environment on life-long health outcomes (discussed in **2.5**).

2.3.1 A natural experiment in rural Gambia: a role for 1C metabolism

The periconceptional period (defined as spanning the 14 weeks before conception and the 10 weeks after)[51] is postulated to show heightened sensitivity to environmental perturbance due to the extensive reprogramming of the epigenome and rapid rates of DNA synthesis that occur after fertilisation and gametogenesis[52]. A natural experiment in rural Gambia provides an interesting setting to investigate the effects of periconceptional environment on methylation at putative human MEs. There, a community of subsistence farmers experiences differences in environmental exposures, including diet and infection levels, between the rainy ('hungry') season (July to September) and dry ('harvest') season (February to April)[53]. Methylation levels at putative MEs are elevated within individuals conceived during the rainy season compared to the dry season, a finding that has been robustly replicated across several studies[2,20,31,41]. These effects been observed in both blood and hair follicle DNA[31], confirming the notion that environmentally perturbed DNAm states in the early embryo can be maintained in different cell types.

Putative MEs identified in genome-wide screens are also enriched for CpGs that are differentially methylated between individuals conceived during the peaks of the rainy or the dry season (season of conception or 'SoC' -associated CpGs)[3,20,24]. More recently, an analysis of individuals with year-round conception dates confirmed that CpGs showing a significant association with the date of conception had methylation maxima and minima falling in the peaks

of the rainy and dry season respectively, and that these CpGs were highly enriched for putative MEs identified in screens in non-Africans[21,42]. Importantly, these CpGs were replicated across two Gambian cohorts with different ages and sample collection methods, and their association with SoC was not confounded by season of sample collection or blood cell-type proportions[20,42].

Several lines of evidence suggest that 1-carbon (1C) metabolism may underpin SoC associations at putative human MEs. 1C metabolic pathways use diet-derived methyl donors (such as folate, choline, and betaine) and cofactors (such as vitamins B2, B6 and B12) to convert S-adenosyl methionine (SAM) into S-adenosyl homocysteine (SAH). DNA methyltransferases (DNMTs) use methyl groups derived from this reaction to establish and maintain DNAm (**Figure 2.3**)[51,54]. Maternal blood plasma concentrations of key 1C metabolism biomarkers measured in Gambian women during pregnancy can be back extrapolated to the time of conception to give an estimate of periconceptional 1C metabolism biomarkers[31]. These concentrations are elevated in the rainy season and can predict DNA methylation at putative MEs[31,41,55]. Maternal periconceptional levels of amino acids valine and arginine in Gambian women that are linked to levels of 1C metabolites[56,57] have also been associated with offspring methylation status at the putative ME *PAX8*[27]. Methylation status at this locus is also altered in offspring born to mothers periconceptionally supplemented with micronutrients that provide methyl groups and cofactors involved in 1C pathways[45].

The link to 1C metabolism aligns with similar findings in *Agouti*^{py} mouse models, in which offspring born to dams supplemented with B12, folate and choline during pregnancy show altered methylation profiles at *Agouti IAP* element[58,59]. However, the pathways through which 1C metabolism influences DNA methylation at putative MEs are complex and likely to be context-specific, as evidenced by the finding of an interaction between SoC and maternal 1C biomarkers that positively influence offspring ME methylation[41].

Whilst the effect of the periconceptional environment on putative ME methylation at the cellular level has not been directly examined, a read-level analysis of putative human MEs in a single-conceptus (genetically identical) embryonic liver sample showed that putative MEs are highly enriched for intermediate methylation states, indicative of stochastic intra-tissue cell variegation effects[3]. This raises the possibility that environmental exposures may 'nudge' the proportion of

methylated cells in early development, resulting in a change in the aggregate sample methylation level within a given individual.

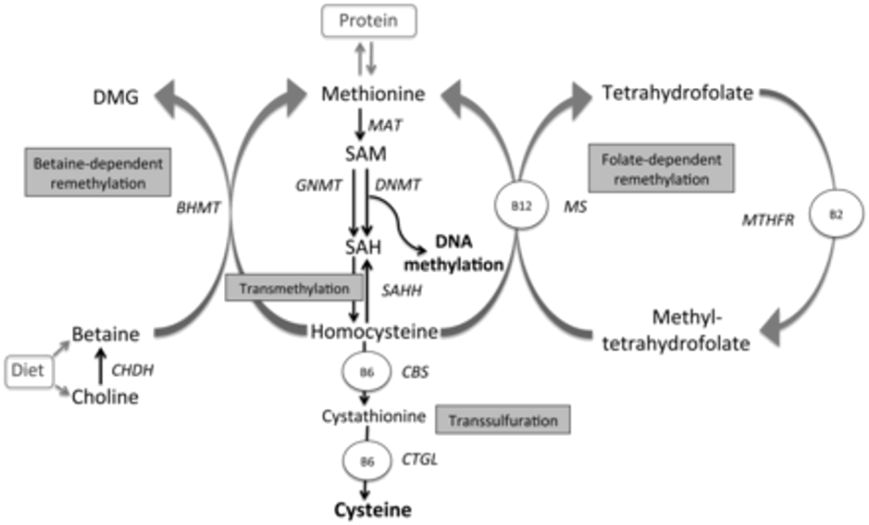


Figure 2.3. A schematic of 1C metabolism. Methyl group supply is dependent on the S-adenosyl methionine (SAM). SAM is converted to S-adenosyl homocysteine (SAH), which removes a methyl group that is used by DNA methyltransferase (DNMT) enzymes. SAH Is then converted into homocysteine. BHMT = betaine-homocysteine methyltransferase, CBS = cystathionine-β-synthase, CHDH = choline dehydrogenase, CTGL = cystathionine-γlyase, DMG = dimethylglycine, GNMT = glycine-N-methyltransferase, MAT = methionine adenosyltransferase, MTHFR = methylenetetrahydrofolate reductase, MS = methionine synthase, SAHH = S adenosylhomocysteine hydrolase. Image from **Dominguez-Salas et al. (2013).**

2.3.2 Artificial Reproductive Technologies (ART)

Use of assisted reproductive technologies (ART)[48] may expose the developing embryo to an atypical nutritional background compared to that experienced during natural conception[48,60]. Estill *et al.* (2016)[48] examined methylation states at 22 putative MEs overlapping the Illumina450K array in bloodspots of new-borns conceived naturally and through different ART strategies, including intrauterine insemination (IUI) and intracytoplasmic sperm injection (ICSI) with fresh or frozen embryos. The majority of putative MEs (19/22) were differentially methylated between individuals conceived in at least one conception-type comparison and several were consistently hypomethylated in naturally conceived individuals[48]. Whilst this supports the notion that putative human MEs are sensitive to the periconceptional environment, robust evidence for an association with ART is lacking and putative MEs were not amongst the differentially methylated regions (DMRs) identified in a genome-wide scan in IVF twins[61], and were similarly not associated with ART in a recent EWAS using the IlluminaEPIC array[62].

Furthermore, the ART-sensitive loci identified by Estill et al. (2016) were not amongst the top regions associated with SoC by Silver et al. (2015)[48].

2.3.3 Beyond periconception

Whilst SoC and ART focus on the periconceptional period, environmental sensitivity may not be limited to this time window[63].

2.3.3.1 Later gestational effects

Several putative human MEs are amongst the top hits in epigenome-wide association studies (EWAS) exploring the association between DNAm and gestational exposures: i) maternal folic acid supplementation at a gestational age of 16 weeks was associated with hypomethylation at the putative MEs *PAX-8* and *VTRNA2-1*[47], ii) maternal gestational diabetes was associated with methylation at the putative ME *CYP2E1 (Cytochrome P450 2E1)*[64], and iii) maternal tobacco smoke exposure *in utero* was associated with methylation at three putative MEs: *BOLA3 (BolA Family Member 3), PAX8,* and *ZFYVE28 (Zinc Finger FYVE-Type Containing 28)*[25].

Methylation states at putative human MEs have also been studied in individuals exposed to maternal gestational famine lasting at least 7 months in rural Bangladesh [44]. In that study, 6 out of 16 putative MEs that had been previously associated with SoC in Gambian children[44] were also associated with gestational famine exposure. However, conception during the rainy ('hungry') season and exposure to gestational famine have opposite effects on DNAm at putative human MEs: the former is associated with *hyper*methylation whilst the latter is associated with *hypo*methylation. Therefore, whilst SoC is linked to 1C metabolism, it is not an analogous exposure to gestational famine.

Together, these studies suggest that putative human MEs may be environmentally labile during mid- and late-gestation. However, in several of the studies discussed above [25,44,64], the exact timing of the exposure effect on methylation has not been pinpointed and may therefore still include periconceptional effects. Nutritional supplementation trials that target specific developmental windows' may in future give better insight into timing-specific exposure effects during gestation.

2.3.3.2 Pre-conceptual effects

The pre-conceptual environment may also influence methylation patterns in offspring if environmentally sensitive methylation states established in gametes are able to evade the global wave of epigenome reprogramming that occurs after fertilisation[65–67].

Evidence for this has been found at the putative ME *VTRNA2-1,* a polymorphically imprinted gene that is usually hypomethylated on the paternal allele but variably methylated on the maternal allele (with ~75% of individuals methylated on the maternal allele and 25% of individuals unmethylated)[68–70]. These methylation patterns are established in the oocyte [68], and evidence suggests that maternal age and alcohol consumption during the oocyte maturation period can increase *VTRNA2-1* methylation at the maternal allele in offspring[68]. This suggests that the methylation states established at this locus in the oocyte can be maintained in the zygote and persist throughout gastrulation.

Methylation states at the putative metastable epiallele *POMC (proopiomelanocortin)* measured in offspring are also correlated with those in paternal somatic cells[71]. However, putative human MEs, including *POMC*, show marked hypomethylation in sperm[3,72], suggesting that other factors such as sperm RNAs[73] or histone modifications may mediate the paternal inheritance of methylation status at *POMC*[71]. In regard to the latter, whilst most histones are replaced by protamines in sperm, putative MEs retain histone H3 lysine 4 trimethylation (H3K4me3)[72] modifications, which have been found to be sensitive to diet changes in males and can be retained during fertilisation in mice[72,74].

2.3.3.3 Postnatal effects

Although putative MEs are established in early development, their sensitivity to the post-natal environment cannot be ruled out. For example, *VTRNA2-1* and *POMC* have been associated with smoking status[75,76] and *VTRNA2-1* has additionally been linked to pesticide exposure and air pollution[76,77]. Similarly, *CYP2E1* has been associated with psychosocial deprivation in childhood[78] and with exposure to toluene and benzenes[79,80]. These studies have each been carried out in a single tissue type, meaning that it the extent to which these effects are tissue-specific is unclear.

2.4 Genomic context

Characterising the genomic features that distinguish putative human MEs from other loci may give insights into the molecular mechanisms underpinning variable methylation establishment at these loci in early development. Notable features associated with putative human MEs include transposable elements (TEs) and regions of parent-of-origin specific methylation (PofOm), raising the possibility that methylation variation at putative human MEs is linked to the epigenetic regulation of these features.

2.4.1 Transposable elements

Murine MEs have been identified at Intracisternal A particles (IAPs)[4], an evolutionarily young class of rodent-specific long terminal (LTR) repeat retrotransposon. In the paradigm Agouti A^{vy} model, the LTR acts as a heterologous promoter which, when hypomethylated, can drive exogenous expression of the Agouti gene[1]. Although IAPs are not present in humans, putative human MEs are enriched for proximity to LTR-containing endogenous retroviruses of the classes ERV1 and ERVK[3], and to non-LTR-containing long interspersed repeat elements (LINEs)[20]. They have also been linked to primate-specific short interspersed repeat (SINE) Alu elements[71,81]. A recent analysis found that SIV loci are enriched amongst evolutionarily young retrotransposon subfamilies with high CpG density such as the LTR12C (a member of the ERV1 family)[82]. Putative MEs are also enriched within sub-telomeric regions[2,20], which are known to contain many repeat sequences and TEs. It is possible that the environmental lability of putative human MEs may be linked to their association with TEs: methylation at several human TEs has been directly linked to environmental exposures[83–85], and loci that are associated with SoC in Gambian children (indicative of lability to the periconceptional environment) are also enriched for proximal ERV1 retrotransposons[42].

Growing evidence in mouse models indicates that methylation variability at murine MEs is driven by variable epigenetic silencing of IAP elements by trans-acting Krüppel associated box (KRAB) zinc finger proteins (KZFPs) in early development[82,86]. KZFPs usually silence TEs (including retrotransposons) by recruiting the co-factor KAP1 (KRAB-associated protein) which acts as a scaffold for proteins that deposit DNAm and H3K9me3 marks[87,88]. Stochastic methylation establishment at murine MEs is postulated to arise due to competition between

KZFP-KAP1 that promote DNAm, and CTCF and ZF-CxxC domain-containing proteins that prevent DNAm[82,89]. Notably, putative human MEs are enriched for proximal KAP1 and CTCF binding sites [3], suggesting that the determinants of methylation variability at putative human MEs may in some cases be similar to those at murine MEs. The association with CTCF is also noteworthy because it has been linked to stochastic switching between epigenetic states in humans[90].

2.4.2 Parent-of-origin specific methylation (PofOm) and ZFP57 binding sites

Putative human MEs are also enriched amongst regions of parent-of-origin specific methylation (PofOm), at which methylation differs between the maternal and paternal allele[42,91]. PofOm can arise from differential gametic methylation that is maintained during the epigenetic reprogramming events in the early embryo. Accordingly, putative human MEs are enriched for regions that are hypermethylated in the oocyte relative to sperm (termed oocyte germline differentially methylated regions ('DMRs'))[91]. Several putative MEs also overlap imprinted genes[24,42], at which PofOm regulates monoallelic expression of the maternal or paternal allele in offspring. Methylation variability at imprinted genes is particularly noteworthy because of the established role of imprinted genes in foetal growth and development[92–94].

These associations raise the possibility that methylation variation at some putative human MEs is caused by the incomplete maintenance of PofOm in the early embryo, whereby stochastic and/or environmental effects could cause a gain of methylation at the unmethylated allele and/or a loss of methylation at the methylated allele. Methylation variability could also arise due to the perturbed establishment of gamete-specific methylation states if they are able to subsequently evade reprogramming at fertilisation. This may be the case at *VTRNA2-1*[68] (discussed in 2.3.3.2), which also shows sensitivity to a range of environmental exposures that may have post-fertilisation effects[70,91,95.]. The environmental lability of regions of PofOm is further supported by growing evidence that the early developmental nutritional environment can influence methylation at imprinted genes[96–101].

The link between putative human MEs and PofOm is strengthened by their enrichment for proximal binding sites for ZFP57[3], a KZFP that maintains PofOm in early embryogenesis[102]. Individuals homozygous for a mutation that impairs ZFP57 function showed altered

methylation status at putative MEs compared to wild-type individuals[3], indicating that ZFP57 can directly influence putative ME methylation. Interestingly, a region immediately upstream of ZFP57 is associated with SoC in Gambian 2-year-olds[20] and to periconceptional and gestational maternal plasma folate during late stages of pregnancy[103–105], indicating that environmental responsiveness of ZFP57 may induce DNAm changes at ZFP57 targets[42]. Additionally, methylation variability at ZFP57 is associated with the presence of a common allele[24] and undergoes haplotype allele-specific methylation attributed to an upstream polymorphic CTCF binding site[106], suggesting that ZPF57 may integrate both genetic and environmental effects on DNAm at regions of PofOm and at some putative human MEs[24].

2.5 Putative MEs and the developmental origins of health and disease

2.5.1 Putative ME methylation can predict disease

Prospective studies using longitudinal cohorts have indicated predictive associations between putative human ME methylation and health outcomes such as cancer, BMI, bodyweight, and thyroid function in later life[24,27,49,107]. Methylation states at putative human MEs could affect phenotype by influencing gene expression in post-natal tissues. Supporting this, DNAm at several putative human MEs has been associated with gene expression in various cell/tissue types[24,49,70,108,109], and well-characterised putative human MEs show evidence of stability with age in longitudinal cohorts. For example, *VTRNA2-1* is stable over a 10-year period from childhood to early adolescence[20] and over 25 years in adulthood[70]; *POMC* is stable from birth to at least 12 years[71,110]; and *PAX8* is stable from ages 7 to 17[27].

Alternatively, some putative MEs may exert their effect on gene expression very early in development and may influence disease risk by altering developmental trajectories. Studying DNAm modification of MEs in the placenta may be particularly insightful for exploring whether ME methylation states can influence developmental trajectories in early life because of the established role of the placenta in foetal growth and development[111,112]. Clark et al.[113] analysed

the association between placental DNAm at MEs and BMI at ages one, 2 and 10. The top hit putative ME loci from this screen were annotated to growth and metabolism genes[113]. Another study found that placental methylation of the imprinted gene *MEST (Mesoderm Specific Transcript)* was associated with offspring birthweight[114]. In an epigenome-wide approach using whole genome bisulphite sequencing, placental methylation status of the putative ME *CYP2E1* was associated with autism onset in male offspring[115].

2.5.2 Putative MEs are amongst top hit loci in EWAS

Although under-represented on Illumina arrays, several putative MEs are also amongst the top hit loci in EWAS: i) methylation at the putative MEs *CYP2E1* and *DUSP22 (Dual Specificity Phosphatase 22)* measured in four immune cell blood types is associated with active and erosive rheumatoid arthritis respectively[116], ii) *DUSP22* methylation measured in whole blood and brain tissue is associated with schizophrenia risk[43], and iii) *VTRNA2-1* methylation is the top hit DMR associated with orofacial clefts[117]. However, due to the cross-sectional nature of these studies, the direction of causality between methylation and the disease has not been established.

2.5.3 Putative MEs as mediators of environmental risk on phenotype

Whilst environmental lability it not a definitive feature of an ME, the majority of the putative human MEs that have been linked to phenotypes have also been associated with early developmental exposures (**Table 2.2**). By integrating exposure, methylation, expression, and phenotype data, the causal pathways through putative human MEs mediate the environmental risk on phenotypes are beginning to be unearthed.

For example, hypomethylation of putative ME *MEST* measured in cord blood is associated with gestational bisphenol A (BPA) exposure, and with BMI at ages 1 and 6 within the same individuals[49]. Notably, *MEST* expression is upregulated upon *BPA* exposure of human mesenchymal stem cells *in vitro,* resulting in increased adipogenesis[49]. These findings indicate that hypomethylation of *MEST* in response to pre-natal BPA exposure may increase obesity risk in later life by increasing *MEST* expression.

A second well-characterised example is POMC*,* a component of the satiety-regulating melanocortin signalling pathway[118]. Hypermethylation at the intron2/exon3 boundary of *POMC*

is associated with conception in the rainy ("hungry") season in Gambian children and with increased obesity risk and BMI in European adults[71,109]. These associations were observed in melanocyte-stimulating hormone (MSH)-positive neurons, a physiologically relevant cell type involved in regulating the satiety response[71]. Taken together with evidence that hypermethylation of the intron2/exon3 boundary is associated with decreased *POMC* expression[109], these findings suggest that hypermethylation of *POMC* in early development may increase obesity risk by decreasing *POMC* gene expression in MSH-positive neurons[71].

Another notable example is PAX8, a transcription factor that regulates thyroid cell differentiation and function[119]. *PAX8* hypermethylation is observed in Gambian children conceived during the rainy season[2,20] and can predict decreased thyroid volume and decreased levels of the thyroid hormone free thyroxine (T4) [27]. Lower free T4 levels were in turn associated with increased adiposity and bone mineral density in the same cohort. This raises the possibility that *PAX8* hypermethylation in early development may increase the propensity to develop metabolic disorders in later life via the downregulation of thyroid hormones[27].

Finally, *VTNRA2-1* hypomethylation has been associated with gestational famine exposure[44], increased BMI and adiposity in childhood[70,120], and increased expression of *VTRNA2-1* RNA transcripts. These RNAs can in turn predict glucose levels in childhood and adulthood, providing a plausible causal pathway through which *VTRNA2-1* methylation influences metabolism in alter life[70].

Taken together, these studies support the notion that putative human MEs are interesting candidates for exploring the epigenetic mechanisms linking the early developmental environment to disease risk in later life[121].

2.5.4 Developmental plasticity as an adaptive mechanism

The association between several environmentally sensitive MEs and low birth weight (**Table 2.2**), may be in line with the *thrifty phenotype hypothesis* first proposed by Hales and Barker[122]. This hypothesis postulates that slowing foetal growth and altering metabolism may be an immediate survival strategy for offspring exposed to high levels of nutrient deprivation by focussing on the growth of critical organs instead of tissues such as the pancreas and muscle[122,123]. When offspring transition from a low nutrient environment to an energy rich environment, these adaptations (such as reduced insulin secretion) will be maladaptive, rendering the individuals at increased risk of developing metabolic disorders[122].

Another explanation is the *predictive adaptive response* (PAR) mechanism, which hypothesises that developmental plasticity *in utero* primes the developing embryo to its future environment, without having an immediate advantage in early life[124,125]. In cases where the postnatal environment is different to the anticipated environment, this 'prediction' will be incorrect and result in increased disease risk. The association between the putative MEs *POMC* and *PAX8* and metabolic outcomes in later life may reflect a PAR mechanism: offspring born to mothers with compromised nutritional status may be epigenetically programmed via the modulation of DNAm at the loci towards being metabolically thrifty and storing fat in later life.

It has been suggested that the evolution of these adaptive mechanisms would require epigenetic variability at environmentally responsive loci to be under genetic control., providing a heritable mechanism by which phenotypic plasticity is hardwired into the genome in order to rapidly adapt to changing local environments[126]. Both *ZFP57* and *PAX8*[24,27] are environmentally responsive and show evidence of methylation variation that is associated with a genetic variant, raising the possibility of a G x E interaction effect at these loci.

Table 2.2. Putative human MEs associated with phenotypes and early developmental exposures. Phenotypes are highlighted in yellow, and exposures are highlighted in blue.

Putative human ME	Phenotype /exposure	Tissue	Reference
VTRNA2-1 *(alias nc886) (Vault RNA2-1)*	Childhood BMI	whole blood	van Dijk et al. (2017)[106]
	Glucose and insulin levels in adolescence, adiposity in childhood	whole blood	Marttila et al. (2021)[69]
	Lung cancer & mature B-cell neoplasm	whole blood	van Baak et al. (2018)[24]
	Pre-term birth	whole blood	You et al. (2021)[126]
	Hepatocellular carcinoma (HCC) and tumour aggressiveness	HCC tumour and normal tissues	Yu et al. (2020)[127]
	Parkinson's disease	whole blood	Henderson et al. (2021)[128]
	Season-of-conception	whole blood, hair follicle	Silver et al. (2015)[20]
	Gestational famine	whole blood	Finer et al. (2016)[44]
	Gestational folate supplementation	saliva	Richmond et al. (2018)[47]
	Maternal age & socioeconomic status	whole blood	Marttila et al. (2021)[69]
	Oocyte age & preconceptual alcohol use	whole blood	Carpenter et al. (2021)[67]
POMC *(proopiomelanocortin)*	BMI & obesity risk	MSH-positive neurons, whole blood	Kühnen et al. (2016)[70]
	Adolescent depressive disorder	whole blood	Zheng et al. (2020)[129]
	Season-of-conception & maternal 1C metabolites	MSH-positive neurons, whole blood	Kühnen et al. (2016)[70]
	Prenatal alcohol exposure	Saliva	Sarkar et al. (2019)[130]
PAX8 *(Paired box8)*	Child thyroid volume & thyroid hormone levels	whole blood	Candler et al. (2021)[27]
	Mammographic density	whole blood	Lucia et al. (2022)[131]
	Sleep duration	saliva	Plante et al. (2021) [132]
	Sperm concentration, morphology, motility	sperm	Houshdaran et al. (2004)[133]
	Season-of-conception & maternal periconceptional amino acids	whole blood	Waterland et al. (2010)[2], Silver et al. (2015)[20], Candler et al. (2021)[27]
	Gestational famine	whole blood	Finer et al. (2016)[44]
	Gestational folate supplementation	saliva	Richmond et al. (2018)[47]
	Maternal smoking	cord blood	Joglekar et al. (2022)[25]
	Periconceptional micronutrient supplementation	whole blood	Saffari et al. (2020)[45]
CYP2E1 *(Cytochrome P450 2E1)*	Autism	placenta	Zhu et al. (2019)[114]
	Active rheumatoid arthritis	CD14+ monocytes, CD4 + naïve T cells	Mok et al. (2018)[115]
	Parkinson's disease	brain	Kaut et al. (2018)[134]
	Maternal gestational diabetes	cord blood	Howe et al. (2020)[64]
	Psychosocial deprivation in early childhood	buccal cells	Kumsta e al. (2016)[77]
	Prenatal antidepressant exposure	cord blood white blood cells	Gurnot et al. (2016)[135]

DUSP22 *(Dual specificity protein phosphatase 2022)*	schizophrenia	whole blood, prefrontal cortex (PFC)	Boks et al. (2018)[43]
	Erosive rheumatoid arthritis	CD14+ monocytes, CD19+ B cells, CD4+ naïve T cells, CD4 + memory T cells	Mok et al. (2018)[115]
	Alzheimer's disease	Hippocampus	Sanchez-Mut et al. (2014)[136]
	Mature B-cell neoplasm, urothelial cell carcinoma*	whole blood	van Baak et al. (2018)[24]
	gestational famine	whole blood, PFC	Boks et al. (2018)[43]
	ART	NA	Estill et al. (2016)[48]
SPATC1L *(Spermatogenesis And Centriole Associated 1 Like)*	Colorectal and prostate cancer	whole blood	van Baak et al. (2018)[24]
	BMI z-score across childhood	placenta	Clark et al. (2019)[112]
	Post-traumatic stress disorder	blood	Katrinli et al. (2021)[137]
	ART	NA	Estill et al. (2016)[48]
MEST *(Mesoderm Specific Transcript)*	BMI at age 1, & longitudinal weight gain	cord blood	Junge et al (2018)[49]
	Adult obesity	blood	El Hajj et al. (2013)[138]
	BMI z -score across childhood	placenta	Clark et al. (2019)[112]
	Birthweight	placenta	Kappil et al. (2015)[113]
	Bisphenol-A	cord blood	Junge et al (2018)[49]
	Maternal gestational diabetes	cord blood	El Hajj et al. (2013)[138]
	Periconceptional micronutrient supplementation	cord blood	Cooper et al. (2012)[46]
PLAGL1 *(Pleomorphic adenoma gene-like 1)*	Estimated foetal weight & weight at 1 year	cord blood	Azzi et al. (2014)[139]
	Sperm concentration morphology motility	sperm	Houshdaran et al. (2004)[133]
	Maternal alcohol, vitamin B2, vitamin B12	cord blood	Azzi et al. (2014)[133]
	Maternal erythrocyte folate levels	cord blood	Hoyo et al. (2014)[97]
ZFP57 *(Zinc finger protein 57)*	Colorectal cancer	whole blood	van Baak et al. (2018)[24]
	Season-of-conception	whole blood	Silver et al. (2015) [20]
	Gestational famine	whole blood	Finer et al. (2016)[44]
	Gestational folate supplementation	CD4+ and antigen-presenting cells from cord blood	Amarasekera et al. (2014)[103]
	Prenatal perfluoroalkyl substance exposure	cord blood	Miura et al. (2018)[140]

BMI = Body Mass Index, 1C = 1-carbon, ART = Artificial Reproductive Technologies, MSH = melanocyte-stimulating hormone

2.6 Conclusion

In summary, putative human MEs share several features with their murine counterparts, including associations with retrotransposons, locus-specific sensitivity to the early environment, and associations with phenotype. Several properties of putative human MEs make them interesting candidates for studying how DNAm variation contributes to phenotypic variation. Firstly, whilst many loci show tissue-specific associations between phenotype and exposures, putative human MEs are correlated across soma-wide tissues, meaning that these associations can be studied in easily accessible tissues. Secondly, since the timing of methylation establishment at MEs is constrained to the very early embryo, they are particularly promising for exploring how stochastic and environmental effects in early life influence health outcomes in later life. Thirdly, their inherent variability makes them particularly useful in EWAS by increasing power to detect associations with phenotypes and exposures at these loci, compared to other DNAm loci that show low levels of interindividual variation.

We are yet to fully understand the causal pathways through which putative human MEs influence health outcomes, the molecular mechanisms driving their establishment in early development, and whether epigenetic variability at these loci is selectively maintained in the human genome. The integration of exposure, methylation, genotype, expression, and phenotype data from longitudinal epidemiological cohorts, coupled with functional studies *in vitro* and *in vivo* using mouse models (where appropriate), will improve understanding of the functional consequences of methylation variation at putative MEs. The application of long-read sequencing data that is suitable for mapping repetitive regions of the genome will provide further insights into whether putative human ME methylation is linked to epigenetic regulation of transposable elements in early development. Further characterisation of the genomic context of putative human MEs may also improve understanding of the determinants of methylation variability. The continued study of putative human MEs will therefore advance understanding of how interindividual methylation variation arising in early development affects disease risk.

2.7 References

1. Rakyan VK, Blewitt ME, Druker R, Preis JI, Whitelaw E. Metastable epialleles in mammals. *Trends in Genetics*. 2002;18(7):348-351.
2. Waterland RA, Kellermayer R, Laritsky E, et al. Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles. *PLoS Genet*. 2010;6(12):e100252
3. Kessler NJ, Waterland RA, Prentice AM, Silver MJ. Establishment of environmentally sensitive DNA methylation states in the very early human embryo. *Sci Adv*. 2018;4(7):eaat2624.
4. Kazachenka A, Bertozzi TM, Sjoberg-Herrera MK, et al. Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell*. 2018;175(5):1259-1271.e13.
5. Wolff GL, Kodell RL, Moore SR, Cooney CA. Maternal Epigenetics and Methyl Supplements Affect Agouti Gene Expression in A vy /a Mice. *The FASEB Jounal*.1998;12(11):949-957.
6. Cooney CA, Dave AA, Wolff GL. Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. *Journal of Nutrition*.2002;132(8):2393S-2400S.
7. Waterland RA, Dolinoy DC, Lin JR, Smith CA, Shi X, Tahiliani KG. Maternal methyl supplements increase offspring DNA methylation atAxin fused. *genesis*. 2006;44(9):401-406.
8. Dolinoy DC, Weidman JR, Waterland RA, Jirtle RL. Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. *Environ Health Perspect*. 2006;114(4):567-572.
9. Dolinoy DC, Huang D, Jirtle RL. Maternal Nutrient Supplementation Counteracts Bisphenol A-Induced DNA Hypomethylation in Early Development. *Proceedings of the National Academy of Sciences*. 2007;104(32):13056-13061.
10. Anderson OS, Nahar MS, Faulk C, et al. Epigenetic responses following maternal dietary exposure to physiologically relevant levels of bisphenol A. *Environ Mol Mutagen*. 2012;53(5):334-342.
11. Kaminen-Ahola N, Ahola A, Maga M, et al. Maternal Ethanol Consumption Alters the Epigenotype and the Phenotype of Offspring in a Mouse Model. *PLoS Genet*. 2010;6(1):e1000811.
12. Neier K, Cheatham D, Bedrosian LD, Dolinoy DC. Perinatal exposures to phthalates and phthalate mixtures result in sex-specific effects on body weight, organ weights and intracisternal A-particle (IAP) DNA methylation in weanling mice. *J Dev Orig Health Dis*. 2019;10(2):176-187.
13. Oey H, Isbel L, Hickey P, Ebaid B, Whitelaw E. Genetic and epigenetic variation among inbred mouse littermates: identification of inter-individual differentially methylated regions. *Epigenetics Chromatin*. 2015;8(1):54.

14. Wolff GL, Roberts DW, Mountjoy KG. Physiological consequences of ectopic agouti gene expression: The yellow obese mouse syndrome. *Physiol Genomics*. 1999;1999(1):151-163.

15. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*. 2011;12(8):529-541.

16. Bertozzi TM, Ferguson-Smith AC. Metastable epialleles and their contribution to epigenetic inheritance in mammals. *Semin Cell Dev Biol*. 2020;97:93-105.

17. Bourc'his D. Metastable epialleles are stable in their instability. *Nat Genet*. 2021;53(8):1121-1123.

18. Byun HM, Siegmund KD, Pan F, et al. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum Mol Genet*. 2009;18(24):4808-4817.

19. Slieker RC, Bos SD, Goeman JJ, et al. Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics Chromatin*. 2013;6(1).

20. Silver MJ, Kessler NJ, Hennig BJ, et al. Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome Biol*. 2015;16(1).

21. Gunasekara CJ, Scott CA, Laritsky E, et al. A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol*. 2019;20(1):105.

22. Michels KB, Binder AM, Dedeurwaerder S, et al. Recommendations for the design and analysis of epigenome-wide association studies. *Nat Methods*. 2013;10(10):949-955.

23. Alan Harris R, Nagy-Szakal D, Kellermayer R. Human metastable epiallele candidates link to common disorders. *Epigenetics*. 2013;8(2):157-163.

24. Van Baak TE, Coarfa C, Dugué PA, et al. Epigenetic supersimilarity of monozygotic twin pairs. *Genome Biol*. 2018;19(1):2. doi:10.1186/s13059-017-1374-0

25. Joglekar R, Grenier C, Hoyo C, Hoffman K, Murphy SK. Maternal tobacco smoke exposure is associated with increased DNA methylation at human metastable epialleles in infant cord blood. *Environ Epigenet*. 2022;8(1):dvac005-dvac005.

26. Marzi SJ, Meaburn EL, Dempster EL, et al. Tissue-specific patterns of allelically-skewed DNA methylation. *Epigenetics*. 2016;11(1):24-35.

27. Candler T, Kessler N, Gunasekara C, et al. DNA methylation at a nutritionally sensitive region of the PAX8 gene is associated with thyroid volume and function in Gambian children. *Sci Adv*. 2021;7(45):eabj1561.

28. Bell JT, Pai AA, Pickrell JK, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol*. 2011;12(1):R10-R10.

29. Villicaña S, Bell JT. Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biol*. 2021;22(1). doi:10.1186/s13059-021-02347

30. Gaunt TR, Shihab HA, Hemani G, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol*. 2016.;17(1):1-14.

31. Dominguez-Salas P, Moore SE, Baker MS, et al. Maternal nutrition at conception modulates DNA methylation of human metastable epialleles. *Nat Commun*. 2014;5(1):3746.

32. Lemire M, Zaidi SHE, Ban M, et al. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat Commun*. 2015;6(1):6326.

33. Min JL, Hemani G, Hannon E, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet*. 2021;53(9):1311-1321.

34. Fraga MF, Ballestar E, Paz MF, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences*. 2005;102(30):10604-10609.

35. Poulsen P, Esteller M, Vaag A, Fraga MF. The epigenetic basis of twin discordance in age-related diseases. *Pediatr Res*. 2007;6.

36. Planterose Jiménez B, Liu F, Caliebe A, et al. Equivalent DNA methylation variation between monozygotic co-twins and unrelated individuals reveals universal epigenetic inter-individual dissimilarity. *Genome Biol*. 2021;22(1):18.

37. Grundberg E, Meduri E, Sandling JK, et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet*. 2013;93(5):876-890.

38. Hannon E, Mansell G, Walker E, et al. Assessing the co-variability of DNA methylation across peripheral cells and tissues: Implications for the interpretation of findings in epigenetic epidemiology. *PLoS Genet*. 2021;17(3):e1009443.

39. Islam SA, Goodman SJ, MacIsaac JL, et al. Integration of DNA methylation patterns and genetic variation in human pediatric tissues help inform EWAS design and interpretation. *Epigenetics Chromatin*. 2019;12(1):1-18.

40. Gunasekara CJ, MacKay H, Scott CA, et al. Systemic interindividual epigenetic variation in humans is associated with transposable elements and under strong genetic control. *bioRxiv* (2022).

41. James PT, Dominguez-Salas P, Hennig BJ, Moore SE, Prentice AM, Silver MJ. Maternal One-Carbon Metabolism and Infant DNA Methylation between Contrasting Seasonal Environments: A Case Study from The Gambia. *Curr Dev Nutr*. 2019;3(1).

42. Silver MJ, Saffari A, Kessler NJ, et al. Environmentally sensitive hotspots in the methylome of the early human embryo. Deelen J, ed. *Elife*. 2022;11:e72031.

43. Boks MP, Houtepen LC, Xu Z, et al. Genetic vulnerability to DUSP22 promoter hypermethylation is involved in the relation between in utero famine exposure and schizophrenia. *NPJ Schizophr*. 2018;4(1):1-8.

44. Finer S, Iqbal MS, Lowe R, et al. Is famine exposure during developmental life in rural Bangladesh associated with a metabolic and epigenetic signature in young adulthood? A historical cohort study. *BMJ Open*. 2016;6(11):e011768.

45. Saffari A, Shrestha S, Issarapu P, et al. Effect of maternal preconceptional and pregnancy micronutrient interventions on children's DNA methylation: Findings from the EMPHASIS study. *Am J Clin Nutr*. 2020;112(4):1099-1113.

46. Cooper WN, Khulan B, Owens S, et al. DNA methylation profiling at imprinted loci after periconceptional micronutrient supplementation in humans: results of a pilot randomized controlled trial. *The FASEB Journal*. 2012;26(5):1782-1790.

47. Richmond RC, Sharp GC, Herbert G, et al. The long-term impact of folic acid in pregnancy on offspring DNA methylation: Follow-up of the Aberdeen folic acid supplementation trial (AFAST). *Int J Epidemiol*. 2018;47(3):928-937.

48. Estill MS, Bolnick JM, Waterland RA, Bolnick AD, Diamond MP, Krawetz SA. Assisted reproductive technology alters deoxyribonucleic acid methylation profiles in bloodspots of newborn infants. *Fertil Steril*. 2016;106(3):629-639.e10.

49. Junge KM, Leppert B, Jahreis S, et al. MEST mediates the impact of prenatal bisphenol A exposure on long-term body weight development. *Clin Epigenetics*. 2018;10(1):1-12.

50. Bertozzi TM, Becker JL, Blake GET, et al. Variably methylated retrotransposons are refractory to a range of environmental perturbations. *Nat Genet*. 2021;53(8):1233-1242.

51. Steegers-Theunissen RPM, Twigt J, Pestinger V, Sinclair KD. The periconceptional period, reproduction and long-term health of offspring: The importance of one-carbon metabolism. *Hum Reprod Update*. 2013;19(6):640-655.

52. Fleming TP, Watkins AJ, Velazquez MA, et al. Origins of lifetime health around the time of conception: causes and consequences. *The Lancet*. 2018;391(10132):1842-1852.

53. Dominguez-Salas P, Hennig BJ, Moore SE, Prentice AM, Cox SE. Can seasonality of one-carbon metabolism in Gambian mothers explain differences in their offspring DNA methylation ? *Clin Chem Lab Med*. Published online 2012.

54. Fox JT, Stover PJ. Folate-mediated one-carbon metabolism. *Vitam Horm*. 2008;79:1-44.

55. Dominguez-Salas P, Moore SE, Cole D, et al. DNA methylation potential: Dietary intake and blood concentrations of one-carbon metabolites and cofactors in rural African women. *American Journal of Clinical Nutrition*. 2013. 97(6):1217-1227.

56. Brosnan ME, MacMillan L, Stevens JR, Brosnan JT. Division of labour: how does folate metabolism partition between one-carbon metabolism and amino acid oxidation? *Biochemical Journal*. 2015;472(2):135-146.

57. Rios-Avila L, Coats B, Chi YY, et al. Metabolite profile analysis reveals association of vitamin B-6 with metabolites related to one-carbon metabolism and tryptophan catabolism but not with biomarkers of inflammation in oral contraceptive users and reveals the effects of oral contraceptives on these processes. *J Nutr*. 2015;145(1):87-95.

58. Waterland RA, Jirtle RL. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol Cell Biol*. 2003;23(15):5293-5300.

59. Waterland RA, Michels KB. Epigenetic Epidemiology of the Developmental Origins Hypothesis. *Annu Rev Nutr*. 2007;27(1):363-388.

60. Canovas S, Ross PJ, Kelsey G, Coy P. DNA methylation in embryo development: epigenetic impact of ART (assisted reproductive technologies). *Bioessays*. 2017;39(11):1700106.

61. Castillo-Fernandez JE, Loke YJ, Bass-Stringer S, et al. DNA methylation changes at infertility genes in newborn twins conceived by in vitro fertilisation. *Genome Med*. 2017;9(1):28.

62. Tobi EW, Almqvist C, Hedman A, et al. DNA methylation differences at birth after conception through ART. *Human Reproduction*. 2021;36(1):248-259.

63. Tobi EW, Lumey LH, Talens RP, et al. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum Mol Genet*. 2009;18(21):4046-4053.

64. Howe CG, Cox B, Fore R, et al. Maternal gestational diabetes mellitus and newborn DNA methylation: findings from the pregnancy and childhood epigenetics consortium. *Diabetes Care*. 2020;43(1):98-105.

65. Perez MF, Lehner B. Intergenerational and transgenerational epigenetic inheritance in animals. *Nat Cell Biol*. 2019;21(2):143-151. doi:10.1038/s41556-018-0242-9

66. Messerschmidt DM, Knowles BB, Solter D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev*. 2014;28(8):812-828.

67. Takahashi Y, Valencia MM, Yu Y, et al. Transgenerational inheritance of acquired epigenetic signatures at CpG islands in mice. *Cell*. 2023;186(4):715-731.

68. Carpenter BL, Remba TK, Thomas SL, et al. Oocyte age and preconceptual alcohol use are highly correlated with epigenetic imprinting of a noncoding RNA (nc886). *Proceedings of the National Academy of Sciences*. 2021;118(12).

69. Carpenter BL, Zhou W, Madaj Z, et al. Mother–child transmission of epigenetic information by tunable polymorphic imprinting. *Proc Natl Acad Sci U S A*. 2018;115(51):E11970-E11977.

70. Marttila S, Viiri LE, Mishra PP, et al. Methylation status of nc886 epiallele reflects periconceptional conditions and is associated with glucose metabolism through nc886 RNAs. *Clin Epigenetics*. 2021;13(1):143.

71. Kühnen P, Handke D, Waterland RA, et al. Interindividual Variation in DNA Methylation at a Putative POMC Metastable Epiallele Is Associated with Obesity. *Cell Metab*. 2016;24(3):502-509.

72. Lambrot R, Chan D, Shao X, et al. Whole-genome sequencing of H3K4me3 and DNA methylation in human sperm reveals regions of overlap linked to fertility and development. *Cell Rep*. 2021;36(3):109418.

73. Chen Q, Yan W, Duan E. Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications. *Nat Rev Genet*. 2016;17(12):733-743.

74. Lismer A, Dumeaux V, Lafleur C, et al. Histone H3 lysine 4 trimethylation in sperm is transmitted to the embryo and associated with diet-induced phenotypes in the offspring. *Dev Cell*. 2021;56(5):671-686.

75. Ehrlich S, Walton E, Roffman JL, et al. Smoking, but not malnutrition, influences promoter-specific DNA methylation of the proopiomelanocortin gene in patients with and without anorexia nervosa. *The Canadian Journal of Psychiatry*. 2012;57(3):168-176.

76. van der Plaat DA, de Jong K, de Vries M, et al. Occupational exposure to pesticides is associated with differential DNA methylation. *Occup Environ Med*. 2018;75(6):427-435.

77. van der Plaat DA, Vonk JM, Terzikhan N, et al. Occupational exposure to gases/fumes and mineral dust affect DNA methylation levels of genes regulating expression. *Hum Mol Genet*. 2019;28(15):2477-2485.

78. Kumsta R, Marzi SJ, Viana J, et al. Severe psychosocial deprivation in early childhood is associated with increased DNA methylation across a region spanning the transcription start site of CYP2E1. *Transl Psychiatry*. 2016;6(6):e830-e830.

79. Jiménez-Garza O, Guo L, Byun HM, et al. Promoter methylation status in genes related with inflammation, nitrosative stress and xenobiotic metabolism in low-level benzene exposure: Searching for biomarkers of oncogenesis. *Food and Chemical Toxicology*. 2017;109:669-676.

80. Jiménez-Garza O, Baccarelli AA, Byun HM, Márquez-Gamiño S, Barrón-Vivanco BS, Albores A. CYP2E1 epigenetic regulation in chronic, low-level toluene exposure: Relationship with oxidative stress and smoking habit. *Toxicol Appl Pharmacol*. 2015;286(3):207-215.

81. Kuehnen P, Krude H. Alu elements and human common diseases like obesity. *Mob Genet Elements*. 2012;2(4):197-201. 82. Costello KR, Leung A, Trac C, et al. Sequence features of retrotransposons allow for epigenetic variability. *Elife*. 2021;10:e71104.

83. Faulk C, Kim JH, Anderson OS, et al. Detection of differential DNA methylation in repetitive DNA of mice and humans perinatally exposed to bisphenol A. *Epigenetics*. 2016;11(7):489-500.

84. Wilhelm-Benartzi CS, Houseman EA, Maccani MA, et al. In utero exposures, infant growth, and DNA methylation of repetitive elements and developmentally related genes in human placenta. *Environ Health Perspect*. 2012;120(2):296-302.

85. Tserga A, Binder AM, Michels KB. Impact of folic acid intake during pregnancy on genomic imprinting of IGF2/H19 and 1-carbon metabolism. *The FASEB Journal*. 2017;31(12):5149-5158.

86. Bertozzi TM, Elmer JL, Macfarlan TS, Ferguson-Smith AC. KRAB zinc finger protein diversification drives mammalian interindividual methylation variability. *Proceedings of the National Academy of Sciences*. 2020;117(49):31290-31300.

87. Iyengar S, Farnham PJ. KAP1 protein: An enigmatic master regulator of the genome. *Journal of Biological Chemistry*. 2011;286(30):26267-26276.

88. Ecco G, Imbeault M, Trono D. KRAB zinc finger proteins. *Development (Cambridge)*. 2017;144(15):2719-2729.

89. Elmer JL, Hay AD, Kessler NJ, Bertozzi TM, Ainscough EAC, Ferguson-Smith AC. Genomic properties of variably methylated retrotransposons in mouse. *Mob DNA*. 2021;12(1).

90. Onuchic V, Lurie E, Carrero I, et al. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. *Science (1979)*. 2018;361(6409):eaar3146.

91. Zink F, Magnusdottir DN, Magnusson OT, et al. Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nat Genet*. 2018;50(11):1542-1552.

92. Piedrahita JA. The role of imprinted genes in fetal growth abnormalities. *Birth Defects Res A Clin Mol Teratol*. 2011;91(8):682-692.

93. Moore GE, Ishida M, Demetriou C, et al. The role and interaction of imprinted genes in human fetal growth. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2015;370(1663):20140074.

94. Reik W, Walter J. Genomic imprinting: Parental influence on the genome. *Nat Rev Genet*. 2001;2(1):21-32.

95. Dugué PA, Yu C, Mckay T, et al. VTRNA2-1: Genetic Variation, Heritable Methylation and Disease Association. *Int J Mol Sci*. 2001. 22(5):2535.

96. Heijmans BT, Tobi EW, Stein AD, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A*. 2008;105(44):17046-17049.

97. Steegers-Theunissen RP, Obermann-Borst SA, Kremer D, et al. Periconceptional maternal folic acid use of 400 µg per day is related to increased methylation of the IGF2 gene in the very young child. *PLoS One*. 2009;4(11).

98. Hoyo C, Murtha AP, Schildkraut JM, et al. Methylation variation at IGF2 differentially methylated regions and maternal folic acid use before and during pregnancy. *Epigenetics*. 2011;6(7):928-936.

99. James P, Sajjadi S, Tomar AS, et al. Candidate genes linking maternal nutrient exposure to offspring health via DNA methylation: A review of existing evidence in humans with specific focus on one-carbon metabolism. *Int J Epidemiol*. 2018;47(6):1910-1937.

100. Lee HS, Barraza-Villarreal A, Biessy C, et al. Dietary supplementation with polyunsaturated fatty acid during pregnancy modulates DNA methylation at IGF2/H19 imprinted genes and growth of infants. *Physiol Genomics*. 2014;46(23):851-857.

101. Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nat Rev Genet*. 2019;20(4):235-248.

102. Takahashi N, Gray D, Strogantsev R, et al. ZFP57 and the targeted maintenance of postfertilization genomic imprints. In: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol 80. Cold Spring Harbor Laboratory Press; 2016:177-187.

103. Gonseth S, Roy R, Houseman EA, et al. Periconceptional folate consumption is associated with neonatal DNA methylation modifications in neural crest regulatory and cancer development genes. *Epigenetics*. 2015;10(12):1166-1176.

104. Amarasekera M, Martino D, Ashley S, et al. Genome-wide DNA methylation profiling identifies a folate-sensitive region of differential methylation upstream of ZFP57-imprinting regulator in humans. *The FASEB Journal*. 2014;28(9):4068-4076.

105. Irwin RE, Thursby SJ, Ondičová M, et al. A randomized controlled trial of folic acid intervention in pregnancy highlights a putative methylation-regulated control element at ZFP57. *Clin Epigenetics*. 2019;11(1):31.

106. Do C, Do C, Dumont ELP, et al. Allele-specific DNA methylation is increased in cancers and its dense mapping in normal plus neoplastic cells increases the yield of disease-associated regulatory SNPs. *Genome Biol*. 2020;21(1).

107. van Dijk SJ, Tellam RL, Morrison JL, Muhlhausler BS, Molloy PL. Recent developments on the role of epigenetics in obesity and metabolic disease. *Clin Epigenetics*. 2015;7(1):1-13.

108. Lee K, Kunkeaw N, Jeon SH, et al. Precursor miR-886, a novel noncoding RNA repressed in cancer, associates with PKR and modulates its activity. *Rna*. 2011;17(6):1076-1089.

109. Kuehnen P, Mischke M, Wiegand S, et al. An Alu Element–Associated Hypermethylation Variant of the POMC Gene Is Associated with Childhood Obesity. *PLoS Genet*. 2012;8(3):e1002543.

110. Yoo JY, Lee S, Lee HA, et al. Can proopiomelanocortin methylation be used as an early predictor of metabolic syndrome? *Diabetes Care*. 2014;37(3):734-739.

111. Woods L, Perez-Garcia V, Hemberger M. Regulation of placental development and its impact on fetal growth—new insights from mouse models. *Front Endocrinol (Lausanne)*. 2018;9:570.

112. Koukoura O, Sifakis S, Spandidos DA. DNA methylation in the human placenta and fetal growth. *Mol Med Rep*. 2012;5(4):883-889.

113. Clark J, Martin E, Bulka CM, et al. Associations between placental CpG methylation of metastable epialleles and childhood body mass index across ages one, two and ten in the Extremely Low Gestational Age Newborns (ELGAN) cohort. *Epigenetics*. 2019;14(11):1102-1111.

114. Kappil MA, Green BB, Armstrong DA, et al. Placental expression profile of imprinted genes impacts birth weight. *Epigenetics*. 2015;10(9):842-849.

115. Zhu Y, Mordaunt CE, Yasui DH, et al. Placental DNA methylation levels at CYP2E1 and IRS2 are associated with child outcome in a prospective autism study. *Hum Mol Genet*. 2019;28(16):2659-2674

116. Mok A, Rhead B, Holingue C, et al. Hypomethylation of CYP2E1 and DUSP22 Promoters Associated With Disease Activity and Erosive Disease Among Rheumatoid Arthritis Patients. *Arthritis and Rheumatology*. 2018;70(4):528-536.

117. Gonseth S, Shaw GM, Roy R, et al. Epigenomic profiling of newborns with isolated orofacial clefts reveals widespread DNA methylation changes and implicates metastable epiallele regions in disease risk. *Epigenetics*. 2019;14(2):198-213.

118. Mountjoy KG. Pro-opiomelanocortin (POMC) neurones, POMC-derived peptides, melanocortin receptors and obesity: how understanding of this system has changed over the last decade. *J Neuroendocrinol*. 2015;27(6):406-418.

119. Fernandez LP, Lopez-Marquez A, Santisteban P. Thyroid transcription factors in development, differentiation and disease. *Nat Rev Endocrinol*. 2015;11(1):29-42.

120. van Dijk SJ, Peters TJ, Buckley M, et al. DNA methylation in blood from neonatal screening cards and the association with BMI and insulin sensitivity in early childhood. *Int J Obes*. 2018;42(1):28-35.

121. Tobi EW, Slieker RC, Luijk R, et al. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Sci Adv*. 2018;4(1):eaao4364.

122. Hales CN, Barker DJP. The thrifty phenotype hypothesis: Type 2 diabetes. *Br Med Bull*. 2001;60(1):5-20.

123. Prentice AM, Rayco-Solon P, Moore SE. Insights from the developing world: thrifty genotypes and thrifty phenotypes. *Proceedings of the Nutrition Society*. 2005;64(2):153-161.

124. Bateson P, Gluckman P, Hanson M. The biology of developmental plasticity and the Predictive Adaptive Response hypothesis. *J Physiol*. 2014;592(11):2357-2368.

125. Gluckman PD, Hanson MA, Spencer HG. Predictive adaptive responses and human evolution. *Trends Ecol Evol*. 2005;20(10):527-533.

126. Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences*. 2010;107(suppl 1):1757-1764.

127. You YA, Kwon EJ, Hwang HS, Choi SJ, Choi SK, Kim YJ. Elevated methylation of the vault RNA2-1 promoter in maternal blood is associated with preterm birth. *BMC Genomics*. 2021;22(1):528.

128. Yu MC, Lee CW, Lin CH, et al. Differential hypermethylation of the VTRNA2-1 promoter in hepatocellular carcinoma as a prognostic factor: Tumor marker prognostic study. *International Journal of Surgery*. 2020;79:282-289.

129. Henderson AR, Wang Q, Meechoovet B, et al. DNA methylation and expression profiles of whole blood in parkinson's disease. *Front Genet*. 2021;12:640266.

130. Zheng D, Bi X, Zhang T, et al. Epigenetic alterations of the promoter region of the POMC gene in adolescent depressive disorder patients with nonsuicidal self-injury behaviors. *Psychol Res Behav Manag*. 2020;13:997.

131. Sarkar DK, Gangisetty O, Wozniak JR, et al. Persistent Changes in Stress-Regulatory Genes in Pregnant Women or Children Exposed Prenatally to Alcohol. *Alcohol Clin Exp Res*. 2019;43(9):1887-1897.

132. Lucia RM, Huang WL, Alvarez A, et al. Association of mammographic density with blood DNA methylation. *Epigenetics*. 2022;17(5):531-546.

133. Plante DT, Papale LA, Madrid A, Cook JD, Prairie ML, Alisch RS. PAX8/PAX8-AS1 DNA methylation levels are associated with objective sleep duration in persons with

unexplained hypersomnolence using a deep phenotyping approach. *Sleep*. 2021;44(10):zsab108.

134. Houshdaran S, Cortessis VK, Siegmund K, Yang A, Laird PW, Sokol RZ. Widespread Epigenetic Abnormalities Suggest a Broad DNA Methylation Erasure Defect in Abnormal Human Sperm. *PLoS One*. 2007;2(12):e1289-.

135. Kaut O, Schmitt I, Wüllner U. Genome-scale methylation analysis of Parkinson's disease patients' brains reveals DNA hypomethylation and increased mRNA expression of cytochrome P450 2E1. *Neurogenetics*. 2012;13(1):87-91.

136. Gurnot C, Martin-Subero I, Mah SM, et al. Prenatal antidepressant exposure associated with CYP2E1 DNA methylation change in neonates. *Epigenetics*. 2015;10(5):361-372.

137. Sanchez-Mut JV, Aso E, Heyn H, et al. Promoter hypermethylation of the phosphatase DUSP22 mediates PKA-dependent TAU phosphorylation and CREB activation in Alzheimer's disease. *Hippocampus*. 2014;24(4):363-368.

138. Katrinli S, Zheng Y, Gautam A, et al. PTSD is associated with increased DNA methylation across regions of HLA-DPB1 and SPATC1L. *Brain Behav Immun*. 2021;91:429-436.

139. el Hajj N, Pliushch G, Schneider E, et al. Metabolic programming of MEST DNA methylation by intrauterine exposure to gestational diabetes mellitus. *Diabetes*. 2013;62(4):1320-1328.

140. Azzi S, Sas TCJ, Koudou Y, et al. Degree of methylation of ZAC1 (PLAGL1) is associated with prenatal and post-natal growth in healthy infants of the EDEN mother child cohort. *Epigenetics*. 2014;9(3):338-345.

141. Miura R, Araki A, Miyashita C, et al. An epigenome-wide study of cord blood DNA methylations in relation to prenatal perfluoroalkyl substance exposure: The Hokkaido study. *Environ Int*. 2018;115:21-28.

Chapter 3

3. Tissue- and ethnicity- hypervariable DNA methylation states show evidence of establishment in the early human embryo

Rationale of chapter

This chapter focuses on the identification of tissue- and ethnicity- independent hypervariable CpGs ('hvCpGs') using Illumina450K array methylation data. We reasoned that CpGs showing consistently high inter-individual variation in a large number of datasets that vary by tissue-type and ethnicity would be robust to tissue-specific and ethnicity-specific drivers of methylation variation, as well as dataset-specific technical artefacts, thereby revealing insights into fundamental mechanisms driving biological methylation variation. We also explore our hypothesis that methylation establishment at these loci may have occurred in the early human embryo and examine the proximity of hvCpGs to ERV1 and ERVK transposable elements as these features have been previously associated with putative human MEs.

Summary of chapter

(Abstract from published paper)

We analysed DNA methylation data from 30 datasets comprising 3474 individuals, 19 tissues and 8 ethnicities at CpGs covered by the Illumina450K array. We identified 4143 hypervariable CpGs ('hvCpGs') with methylation in the top 5% most variable sites across multiple tissues and ethnicities. hvCpG methylation was influenced but not determined by genetic variation, and was not linked to probe reliability, epigenetic drift, age, sex or cell heterogeneity effects. hvCpG methylation tended to covary across tissues derived from different germ-layers and hvCpGs were enriched for proximity to ERV1 and ERVK retrovirus elements. hvCpGs were also enriched for loci previously associated with periconceptional environment, parent-of-origin-specific methylation, and distinctive methylation signatures in monozygotic twins. Together, these properties position hvCpGs as strong candidates for studying how stochastic and/or environmentally influenced DNA methylation states which are established in the early embryo and maintained stably thereafter can influence life-long health and disease.

Notes

- This paper was published in Nucleic Acids Research (May 2022) and is available at: 0.1093/nar/gkac503
- The Supplementary Figures and Tables are available at: https://academic.oup.com/nar/article/50/12/6735/6609816#supplementary-data
- The Supplementary Figures have also been included in Appendix A of this thesis.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	lsh1803699	**Title**	Ms
First Name(s)	Maria		
Surname/Family Name	Derakhshan		
Thesis Title	Identification and characterisation of tissue- and ethnicity-independent hypervariable DNA methylation loci in humans		
Primary Supervisor	Matt Silver		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Nucleic Acids Research		
When was the work published?	May 2022		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	**No**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I performed all the data analysis in the paper.Dr Matt Silver and Dr Garrett Hellenthal gave guidance on these analyses. The samples from the foetal multi-tissue dataset had been previously collected and processed by co-authors Dr Miho Ishida and Dr Charalambos Demetriou (UCL Institute of Child Health). I processed the raw IDAT files to generate methylation Beta values for th is dataset. The methylation Beta values for the internal Gambian dataset had been previously generated by the EMPHASIS study group (PI Prof Caroline Fall), and methylation Beta values for Bornean and Kenyan samples had been generated by Dr Nicolas Brucato. I wrote the first draft of the paper and and all authors reviewed and contributed to the final manuscript.

SECTION E

Student Signature	[redacted]
Date	29/11/2022

Supervisor Signature	[redacted]
Date	2/12/22

Nucleic Acids Research, 2022 **1**
https://doi.org/10.1093/nar/gkac503

Tissue- and ethnicity-independent hypervariable DNA methylation states show evidence of establishment in the early human embryo

Maria Derakhshan[1], Noah J. Kessler[2], Miho Ishida[3], Charalambos Demetriou[3], Nicolas Brucato[4], Gudrun E. Moore[3], Caroline H.D. Fall[5], Giriraj R. Chandak[6], Francois-Xavier Ricaut[4], Andrew M. Prentice[7], Garrett Hellenthal[8,*] and Matt J. Silver[1,7,*]

[1]London School of Hygiene and Tropical Medicine, UK, [2]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK, [3]UCL Great Ormond Street Institute of Child Health, UK, [4]Laboratoire Évolution and Diversité Biologique (EDB UMR 5174), Université de Toulouse Midi-Pyrénées, CNRS, IRD, UPS,Toulouse, France, [5]MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, UK, [6]Genomic Research on Complex Diseases (GRC Group), CSIR-Centre for Cellular and Molecular Biology,Hyderabad, India, [7]Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, The Gambia and [8]UCL Genetics Institute, University College London, Gower Street, London WC1E 6BT, UK

Received December 16, 2021; Revised May 06, 2022; Editorial Decision May 26, 2022; Accepted May 27, 2022

ABSTRACT

We analysed DNA methylation data from 30 datasets comprising 3474 individuals, 19 tissues and 8 ethnicities at CpGs covered by the Illumina450K array. We identified 4143 hypervariable CpGs ('hvCpGs') with methylation in the top 5% most variable sites across multiple tissues and ethnicities. hvCpG methylation was influenced but not determined by genetic variation, and was not linked to probe reliability, epigenetic drift, age, sex or cell heterogeneity effects. hvCpG methylation tended to covary across tissues derived from different germ-layers and hvCpGs were enriched for proximity to ERV1 and ERVK retrovirus elements. hvCpGs were also enriched for loci previously associated with periconceptional environment, parent-of-origin-specific methylation, and distinctive methylation signatures in monozygotic twins. Together, these properties position hvCpGs as strong candidates for studying how stochastic and/or environmentally influenced DNA methylation states which are established in the early embryo and maintained stably thereafter can influence life-long health and disease.

INTRODUCTION

DNA methylation (DNAm) plays a critical role in mammalian development, underpinning X-chromosome inactivation, genomic imprinting, silencing of repetitive regions and cell differentiation (1). DNAm states that vary between individuals have been a focus of Epigenome-Wide Association Studies (EWAS) due to their potential to drive phenotypic variation (2,3). Factors influencing interindividual methylation differences include genetic variation (4,5), cell heterogeneity effects (6,7), sex (8,9), age (10,11) and pre- and post-natal environment (12–14). Growing evidence from studies investigating DNAm patterns in multiple tissues suggests that these factors can have both shared and tissue-specific influences on DNAm variation (12,15–18).

In this study, we identified and characterized hypervariable CpGs ('hvCpGs') covered on the widely used Illumina HumanMethylation450K (hereafter 'Illumina450K') array (19) that showed high interindividual variation in multiple datasets covering 19 different tissue/cell types and 8 ethnicities spanning a wide range of ages. We reasoned that identified loci would be robust to both tissue-specific drivers of methylation variability such as those mentioned above and dataset-specific technical artefacts (20–23), thereby revealing insights into biological mechanisms influencing methylation variation across system-wide tissues.

Tissue-independent methylation variation has been previously observed at a class of loci at which methylation not only varies between individuals but is also *correlated* across tissues derived from different germ layers within a given individual. Also described as 'systemic inter-individual variation' or SIV, this property is attributed to stochastic methylation establishment in the pre-gastrulation embryo (24–29). Accordingly, SIV CpGs overlap loci showing 'epigenetic supersimilarity' (ESS) indicating establishment before cleavage in monozygotic (MZ) twins (27) and show sensi-

*To whom correspondence should be addressed. Tel: +44 020 7679 2189; Email: g.hellenthal@ucl.ac.uk
Correspondence may also be addressed to Matt J. Silver. Tel: +44 020 7927 2557; Email: matt.silver@lshtm.ac.uk

© The Author(s) 2022. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

tivity to the periconceptional environment (24,25,27,28,30). Several SIV loci have been associated with phenotypic traits and disease, including obesity (31), cancer (25,27), rheumatoid arthritis (32), autism (33), Alzheimer's disease (34), Parkinson's disease (35) and thyroid volume and function-related differences in body fat and bone mineral density (36). SIV loci are therefore promising candidates for exploring the developmental origins of disease, with the additional advantage that easily accessible tissues can be used as proxies for pathologically relevant but inaccessible tissues (37).

We investigated whether hvCpGs showed evidence of establishment in the early embryo and sensitivity to the periconceptional environment. We also examined the genomic context of hvCpGs by exploring their association with multi-tissue histone marks, transposable elements and regions of parent-of-origin specific methylation. Finally, we probed putative functional roles of hvCpGs by interrogating EWAS trait associations and by performing gene ontology enrichment analysis.

Our curated set of hvCpGs show methylation variation that is not explained by probe reliability, age, sex, cell heterogeneity or genetic effects. Instead, hvCpGs show evidence of establishment in the early embryo and correlation across tissues. They therefore serve as a useful resource for studying the influence of early environmental and/or stochastic effects on DNAm in diverse tissues and ethnicities, and for studying the impact of DNAm differences on life-long health and disease.

MATERIALS AND METHODS

Methylation data used for identifying hvCpGs

Publicly available methylation Beta matrices were downloaded from The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/tcga) and the Gene Expression Omnibus (GEO) (38) (Supplementary Tables 1 and 2). Methylation Beta matrices were analysed instead of .idat files because Beta matrices are readily available in public databases, and because analysis of multiple datasets with different processing pipelines should strengthen the robustness of our findings of shared high methylation variance across datasets. The TCGA database was used a resource for downloading methylation data from a large number of tissues. TCGA methylation data were downloaded using the *TCGAbiolinks (v2.18.0)* R package (39–41), selecting only samples annotated as 'Solid Tissue Normal'. Of the 33 TCGA datasets, 10 were selected for our study as these had methylation data in at least 20 samples. GEO methylation Beta matrices were downloaded using the *GEOquery (v2.58.0)* R package (42) from 11 unique accessions that were selected to expand both the number of tissues and ethnicities used in our study. Where available, detection p-values (measuring signal intensity), and metadata on age, sex and disease status were also downloaded. We split GEO beta matrices into separate groups based on ethnicity and tissue/cell type and refer to the resulting 17 separated groups as 'datasets'. Non-public datasets internal to this study include IlluminaEPIC (43) array data from whole blood samples from Gambian 8–9-year olds (ISRCTN14266771 (44)) and Illumina450K data from Bornean and Kenyan saliva samples (45) (Supplementary Table S3). These datasets were chosen to expand the number of ethnicities considered in this study. For IlluminaEPIC datasets we selected probes covered on the Illumina450K array. In total, we analysed 30 datasets (3 internal, 10 TCGA and 17 GEO) that covered 8 ethnicities and 19 different tissue/cell types (Supplementary Table S4).

Methylation data processing

For each methylation dataset used in our main analysis, we used the *ChAMP (v2.20.1)* R package (46) to remove: (i) probes with a detection *P*-value >0.01 in >5% samples (where detection *P*-values were available), (ii) probes mapping to multiple genomic positions (47), (iii) probes mapping to the X and Y chromosomes and (iv) single nucleotide polymorphism (SNP)-related probes identified by Zhou *et al*. (48) that contain SNPs (MAF > 1%) that are within 5 bp of the CpG interrogation site and/or SNPs effecting probe hybridization. Where ethnicity information was available, we removed probes with population-specific SNPs identified by Zhou *et al.* using 1000 Genomes populations (MAF > 1%), otherwise we removed the General Recommended Probes Probes (48). Probes that had a missing value in any of the samples in a specific dataset were removed from that dataset. To reduce technical biases introduced by differing type I and type II probe designs on the Illumina450K and IlluminaEPIC arrays, we applied Beta Mixture Quantile normalization (BMIQ) (49) to all datasets using the *champ.norm* function from the *ChAMP* R package. All datasets were adjusted for the first 10 principal components (PCs) of variation to account for methylation variability driven by known and/or unknown technical artefacts (such as plate and array position) and cell heterogeneity. Methylation values were adjusted for these 10 PCs, age (where available) and sex by taking the residuals from a single linear regression model on methylation M values, where M is defined as $\log_2(\text{beta}/1 - \text{beta})$ (see Supplementary Tables 1–3 for details of the linear model applied to each dataset). Adjusted M values were transformed back into Beta values by applying the transformation exp(adjusted_M)/(1 + exp(adjusted_M)). Finally, for each probe, we removed outlier methylation values, defined according to Tukey's outer fences (Q1 – 3*IQR and Q3 + 3*IQR). The hg19 reference genome was used throughout all relevant analyses as the Illumina450K array metadata manifest uses this version.

Identification of hvCpGs

We defined an hvCpG in the following way:

(1) in ≥65% of datasets in which the CpG is covered (following quality control), it has methylation variance in the top 5% of all (non-removed) CpGs.
(2) is covered in at least 15 of the 30 datasets.

While the (5%, 65%) threshold in (1) is arbitrary, we note that ~80% of the resulting set of hvCpGs were also captured when using a different (20%, 90%) threshold, indicating that these hvCpGs are in the top 20% of variable CpGs in ≥90% of datasets they are covered in (Supplementary Figure S1).

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

Table 1. Main CpG sets used in this study

CpG set	*n*	Notes
hvCpGs	4143	CpGs within top 5% methylation Beta variance in at least 65% datasets in which the CpG is covered, requiring the hvCpG to be covered in at least 15 datasets and to be reported as technically reliable.
array background	406 306	CpGs covered in at least 15 of the 30 datasets used in this study.
distribution-matched controls	3566	Array background CpGs with similar methylation Beta distributions to hvCpGs in the 'Blood_Cauc' dataset, requiring each control CpG to be technically reliable.
de-clustered hvCpGs	2640	A set of hvCpGs in which no CpGs is within 4 kb of another CpG.
mQTL-matched controls	3722	CpGs reported by the GoDMC meta-GWAS (51) with the same number of mQTL associations and similar mean % variance explained by an mQTL, requiring each control CpG to be present in at least as many datasets as the hvCpG.

Probe reliability

Technically unreliable probes were identified by examining intra-class correlation coefficients (ICCs) from two studies. The first study compared methylation consistency between the Illumina450K and IlluminaEPIC platforms using 365 blood DNA samples, defining poor quality probes as those with ICC $\leq$0.4 (23). The second study examined methylation reliability between technical replicates from 265 African American peripheral blood leukocyte samples on the Illumina450K platform, defining poor quality probes as those with ICC $\leq$0.37 (50). We defined technically unreliable probes as those reported as being poor quality in at least one of these two studies.

Methylation quantitative trait locus (mQTL) analysis

mQTL summary statistics from the Genetics of DNA Methylation Consortium (GoDMC), a meta-GWAS of 36 European blood cohorts ($N = 27\,750$) generated using imputed genotype data ($\sim$10 million SNPs) and $\sim$420 000 CpGs (51) were used for this analysis. Significance thresholds of $P < 1 \times 10^{-8}$ and $P < 1 \times 10^{-14}$ were applied for *cis* and *trans* mQTLs, respectively (51), giving 271 724 significant SNP-CpG associations comprising 190 102 CpGs and 224 648 SNPs. The variance in DNA methylation explained by a given mQTL was estimated as $2 * \beta * \text{MAF}(1 - \text{MAF})$, where β is the effect size and MAF is the minor allele frequency (52). To investigate mQTL effects acting at a given CpG, we first calculated the % variance explained by each associated mQTL before calculating the mean % variance explained across all mQTLs associated with the CpG.

Monozygotic twin discordance

We analysed CpGs identified as being 'equivalently variable' between MZ co-twins and between unrelated individuals ('evCpGs (blood)') by Planterose Jiménez *et al.*(53) using Illumina450K data in whole blood. 154 of these evCpGs replicated in adipose tissue from 97 MZ twin pairs ('evCpGs (blood & adipose)'). evCpGs are candidates for methylation states that are established stochastically after MZ twin splitting and are used in our study to indicate CpGs at which genetic effects do not play a large role in methylation variation.

Control CpG sets

Distribution-matched controls. To ensure that several of our analyses are not biased by distributional properties of hvCpGs such as their high variability and enrichment for intermediate methylation states (Supplementary Figure S2), we constructed a set of CpGs with similar distribution of methylation Beta values to hvCpGs in the Caucasian blood dataset ('Blood_Cauc', Supplementary Table S1). This dataset was chosen as it has the highest number of post-natal samples and because several downstream analyses leverage published studies that used blood methylation data. For each of the 4108 hvCpGs covered in the 'Blood_Cauc' dataset, a two-sided Kolmogorov–Smirnov (KS) test (*ks.test* in R) was used to test for the divergence in methylation Beta distributions between the hvCpG and each technically reliable (see 'Probe reliability', Methods) background CpG, selecting the background CpG with the greatest *P*-value (requiring a *P*-value > 0.1). In total, 3566 hvCpGs were each matched to a control CpG ('distribution-matched controls', Table 1, Supplementary Figure S3).

mQTL-matched controls. To determine the degree to which hypervariability at hvCpGs is explained by mQTL effects, each hvCpG was matched to a CpG amongst those reported in the GoDMC meta-analysis (51). Controls were selected to have (i) the same number of mQTL associations, (ii) a similar mean % variance explained by mQTL (across all significant mQTL) and (iii) presence in at least as many datasets as the hvCpG (Table 1, Supplementary Figure S4).

Identification of hvCpG clusters

hvCpG clusters were identified by considering the decay of methylation correlation with distance at hvCpGs. To do this, we calculated the average pairwise Spearman correlation (ρ) across hvCpG pairs with inter-CpG distance falling within 100 bp bins, for datasets with at least 100 samples (Supplementary Figure S5A). The distance threshold for defining hvCpG clusters was chosen to be 4000 bp as this is approximately the point at which pairwise correlations levelled out (Supplementary Figure S5A). In total, 2219 (54%) hvCpGs fell into 716 clusters comprising at least two CpGs, with the remaining 1924 (46%) hvCpGs falling outside of these clusters (Supplementary Figure S5B). In 563 (79%) of these clusters, the average Spearman corre-

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

lation (ρ) across hvCpG pairs was >0.5 (Supplementary Figure S5C).

'De-clustering' of hvCpGs

To account for the possibility that our analyses may be biased by the non-random distribution and inter-dependence of hvCpGs in CpG clusters, we generated a de-clustered set of hvCpGs in which no CpG was within 4 kb of another CpG. 2640 de-clustered hvCpGs were generated by randomly selecting one CpG from each of the clusters and then including all 'singleton' CpGs falling outside of clusters.

Age stability

To examine temporal stability of hvCpGs we used published intra-class correlation coefficients (ICCs) for probes on the Illumina450K array determined using white blood cell samples taken ~6 years apart (54). The ICC scores compare within-sample variability (across the two time-points) to between-sample variability, with ICC ≥0.5 defined as temporally stable by Flanagan *et al.* (54). To account for the possibility that high ICC scores might be driven by the high variability of hvCpGs, we compared ICC scores at hvCpGs to those at CpGs with similar methylation Beta distributions to hvCpGs at the first time point (Supplementary Figure S6A). These CpGs were matched to each hvCpG using the same Kolmogorov-Smirnov method detailed in 'Distribution-matched controls' but using publicly available Flanagan *et al*. methylation data (GSE61151) instead of the 'Blood_Cauc' dataset (54).

While we regressed out the effect of age in those datasets where this covariate was available, we checked for potential residual age effects on both methylation mean and variance ('epigenetic drift') by (a) performing sub-analyses of infant and cord blood datasets (Supplementary Table S10); and (b) determining the proportion of hvCpGs that overlap a published set of 6108 CpGs identified using whole blood Illumina450K data from 3295 individuals aged 18–88 years that show an increased methylation variability with age of >5% every 10 years (11) (Supplementary Figure S6B).

Sex effects

We regressed out the effect of sex in those datasets where this was available. However, since sex is an important potential driver of inter-individual methylation differences, we further examined the potential for hvCpG methylation to be driven by sex-specific effects in 8 datasets that had an approximately equal number of male and female samples and a sample size >80 (Blood_Japan, Blood_Mexican, Blood_Gamb, CD4+_Estonian, CD8+_Estonian, Saliva_Cauc, Buccals_Sing_9mo, Buccals_Cauc). We split each dataset by sex to generate 8 'male-only' and 'female-only' sub-datasets. We then calculated the proportion of hvCpGs that had methylation variance in the top 5% in each of these sub-datasets.

Published CpG sets used to investigate early embryo establishment

We used the following publicly available data to examine evidence that methylation states at hvCpGs are established in the early embryo. See Table 2 for a summary of these datasets. We note that not all hvCpGs may have been covered in the array background of all of these studies.

Systemic Interindividual Variation ('SIV') CpGs. SIV-CpGs were collated from four published datasets that used either whole genome bisulfite sequencing (WGBS) or Illumina450K data from multiple tissues derived from different germ layers to identify CpGs displaying high interindividual variation and low intra-individual (cross-tissue) variation. These properties are suggestive of variable methylation establishment before germ layer differentiation (26–29). Further details on the four SIV screens used in this study are given in Supplementary Table S6.

Epigenetic supersimilarity ('ESS') CpGs. Epigenetic supersimilarity (ESS) loci were identified by van Baak *et al.* (27) using Illumina450K data from adipose tissue from 97 MZ and 162 dizygotic (DZ) twin pairs (55). In that study, 1580 ESS sites were identified within the top decile of methylation variance, with an interindividual methylation range >0.4 and greater-than-expected concordance in MZ twins vs DZ twins. This supersimilarity amongst MZ twins is attributed to methylation establishment before MZ twin splitting.

MZ twinning CpGs. Van Dongen *et al.* (56) performed an epigenome-wide association analysis on each of 6 cohorts with methylation data from both MZ and DZ twins (five blood and one buccal) to identify probes differentially methylated between MZ twins and DZ (dizygotic) twins. A meta-analysis was then performed using the blood datasets to identify 834 Bonferroni-significant differentially methylated CpGs, which we refer to as 'MZ twinning CpGs'.

Season of conception ('SoC') CpGs. Silver *et al.* (57) used Illumina450K data to identify 259 CpGs associated with season-of-conception ('SoC') in Gambian 2-year olds, each of which showed a minimum methylation difference of 4% between individuals conceived in the peaks of the Gambian rainy and dry seasons.

Transposable elements and telomeres. Locations of ERV1 and ERVK transposable elements determined by RepeatMasker were downloaded from the UCSC annotations repository as previously described (28). Telomere coordinates were downloaded from the UCSC hg19 annotations repository (http://genome.ucsc.edu).

Imprinted genes, parent-of-origin-specific methylation (PofOm). Imprinted genes classified as 'predicted' or 'known' were downloaded from https://www.geneimprint.com. Parent-of-origin-specific CpGs were identified by Zink *et al.* (58) using WGBS data from peripheral blood from Icelandic individuals.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

Table 2. Published CpG sets used in this study

CpG set	Description	n. CpGs overlapping array background	Reference
SIV	Interindividual methylation variation with concordant methylation across tissues derived from different germ layers within a given individual. See Supplementary Table S6.	3089	Harris *et al.*, van Baak *et al.,* Kessler *et al.,* Gunasekara *et al.* (26–29)
ESS	Greater-than-expected methylation similarity between MZ co-twins.	1217	van Baak *et al.* (27)
MZ twinning CpGs	Probes differentially methylated between MZ and DZ twins.	728	van Dongen *et al.* (56)
evCpGs	MZ co-twin methylation discordance that is equivalent to methylation discordance between unrelated individuals in whole blood. A subset of these replicated in adipose tissue.	317 (blood) 145 (blood & adipose)	Planterose Jimnez *et al.* (53)
SoC	CpGs at which methylation is associated with season of conception in Gambian children.	242	Silver *et al.* (57)
PofOm	Regions of parent-of-origin-specific methylation identified in peripheral blood from Icelandic individuals.	732 CpGs in 116 PofOm regions	Zink *et al.* (58)

SIV = systemic interindividual variation, ESS = epigenetic supersimilarity, evCpGs = equivalently variable CpGs, SoC = season-of-conception, PofOm = parent-of-origin-specific methylation.

SIV power calculation

To assess power to detect SIV in previous screens with small numbers of samples, we analysed the 4-individual multi-tissue dataset used by van Baak *et al.* (27,59). We downloaded this dataset from GEO (GSE50192), selecting the same tissues (gall bladder, abdominal aorta sciatic nerve) used by van Baak *et al.* (27). For each of the 1042 SIV-CpGs reported by van Baak *et al.*, we generated methylation values for three tissues for each simulated individual by randomly sampling from a 3D multivariate normal distribution, with mean equal to the mean of each tissue's sampled methylation values at the CpG, and standard deviation specified by a 3 × 3 cross-tissue co-variance matrix of the sampled methylation values at the CpG. For each SIV-CpG, we sampled four simulated individuals and determined if this random sample met the SIV definition specified by van Baak *et al.* (27), repeating this process 1000 times to give a power estimate (Supplementary Figure S7).

Processing and analysis of fetal multi-tissue dataset

The unpublished fetal multi-tissue dataset comprised 60 samples, corresponding to 30 individuals that each have methylation data from two tissues derived from different germ layers (ectoderm: brain, spinal cord, skin; mesoderm: kidney, rib, heart, tongue; endoderm: intestine, gut, lung, liver). These fetal tissues were obtained from the 'Moore Fetal Cohort' from the termination of pregnancies at Queen Charlotte's and Chelsea Hospital (London, UK). Ethical approval for obtaining fetal tissues was granted by the Research Ethics Committee of the Hammersmith, Queen Charlotte's and Chelsea and Acton Hospitals (2001/6028). DNA was extracted from fetal tissues using the AllPrep DNA/RNA/Protein Mini Kit (Qiagen) and bisulfite conversion was carried out using EZ DNA Methylation Kits (Zymo Research). Samples were then processed using the Illumina InfiniumEPIC array. Derived methylation data were imported as *.idat* files into R and analysed using the *meffil* R package (v 1.1.2) (60) with default parameters. Briefly, methylation predicted sex was used to remove two sex outliers (samples with methylation > 5 SDs from mean). Next, 1 sample was removed for which the predicted median methylation signal was more than 3 SDs from the expected signal, leaving 57 samples. 515 probes with detection-*P*-value value >0.1 and 307 probes with bead number <3 in >20% of samples respectively were removed. Array data were then corrected for dye-bias and background effects and functional normalization was applied, specifying the number of PCs to be 7 (the PC at which the variance explained at control probes levelled out). Next, the *ChAMP (v2.20.1)* R package (46) was used to remove cross-hybridizing and multi-mapping probes, probes on XY chromosomes, and SNP-related probes, leaving 746 492 CpGs. We selected the 452 016 probes that overlapped the Illumina450K array and the 27 individuals for which both tissue samples passed quality control. This included nine individuals with methylation data from endoderm and mesoderm, 10 individuals with methylation data from endoderm and ectoderm and eight individuals with methylation data from mesoderm and ectoderm (see Supplementary Table S7). Methylation was then adjusted for predicted sex and batch using a linear model. The mean gestational age of these individuals was 12.5 weeks.

For the nine individuals with available endoderm–mesoderm samples, we calculated the Pearson *r* between germ layer methylation values for each hvCpG and repeated this for individuals with endoderm–ectoderm and mesoderm–ectoderm samples. The inter-germ layer correlation was then defined as the average Pearson r across these three comparisons. We calculated interindividual variation using the same metric as van Baak *et al.* (27): for each CpG, we took the mean methylation value across the two germ-layer derived tissues for every individual (giving 27 values for each CpG) and defined interindividual variation of the CpG as the range of these means.

Chromatin states at hvCpGs

Chromatin states were predicted by a ChromHMM 15-state model(61) using Chromatin Immunoprecipitation Sequencing (ChIP-Seq) data generated by the Roadmap Epigenomics Consortium (62). These data were downloaded for H1 ESCs (E003), fetal brain (E071), fetal muscle (E090), fe-

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

tal small intestine (E085), foreskin fibroblasts (E055), adipose (E063) and primary mononuclear cells (E062) from the Washington University Roadmap repository. Chromatin states were collapsed into eight states for clarity (Supplementary Table S8).

EWAS trait associations at hvCpGs

hvCpG trait associations were determined using the EWAS catalogue (http://ewascatalog.org/), which details significant results (P-value $< 1 \times 10^{-4}$) from published EWAS studies. Considering only those traits for which at least 1% of hvCpGs overlapped associated CpGs (highlighted in green in Supplementary Table S9), we first extracted the array background CpGs overlapping the 'Blood_Cauc' dataset that were associated with each trait. We then calculated enrichment odds ratios of hvCpGs relative to blood distribution-matched controls (Table 1) and determined the significance of the enrichment using Fisher's Exact Tests.

GTEx transcription levels

The median gene transcription levels for 54 tissues were downloaded from the GTEx portal (https://gtexportal.org/home/datasets). Transcription levels were examined at 416 out of the 425 genes that were annotated to a hvCpG cluster in the Illumina450K manifest.

Gene ontology term enrichment analysis

Gene Ontology (GO) term enrichment analysis was performed using the *missMethyl* R package (*v1.24.0*) (63) using the *gometh* function, setting arguments sig.cpg = hvCpGs, all.cpg = array.background, sig.genes = T, collection = 'GO', array.type = '450K' and prior.prob = T to adjust for variation in the number of 450K probes mapping to each gene.

Bootstrapped confidence intervals

All bootstrapped 95% confidence intervals were calculated over 1000 bootstrap samples.

RESULTS

Identification of hypervariable CpGs

We analysed methylation data from 3474 individuals across 30 datasets (28 Illumina450K and 2 EPIC array) comprising 19 unique tissue/cell types and 8 ethnicities covering a range of ages (Supplementary Tables 1–4). We focussed on CpGs covered by the Illumina450K array and began by excluding probes with poor detection p-values, cross-hybridizing probes, probes on the X and Y chromosomes and probes associated with known SNPs (see Materials and Methods for details).

We aimed to identify CpGs with consistently high interindividual variation in methylation across diverse datasets, so minimizing the effects of dataset-specific drivers of variability including those related to different normalization methods and processing pipelines. Reasoning that removal of unmeasured technical, batch and cell heterogeneity effects would maximize power to detect true variable methylation states, we adjusted all methylation values for the first ten principal components (PCs) of methylation variation, and additionally adjusted for sex (in datasets with both sexes) and age (where available).

Our strategy for identifying tissue- and ethnicity- independent hypervariable CpGs ('hvCpGs') is summarized in Figure 1 and detailed in 'Materials and Methods'. We defined hvCpGs as CpGs with methylation Beta variance in the top 5% of all CpGs in at least 65% of datasets in which the CpG was covered (Table 1), yielding 4330 hvCpGs. Note that no CpGs are expected to meet these criteria if the top 5% most variable CpGs in each dataset are entirely independent of those in the others. These thresholds are arbitrary but were chosen in order to select CpGs that met our required criteria of being highly variable in a large number of tissues (median = 13, IQR = [10,15]) and ethnicities (median = 7, IQR = [[6,7]) (Figure 1B). Further, we note that ~80% of identified hvCpGs were within the top 20% of variable CpGs in at least 90% of datasets (Supplementary Figure S1), meaning that the majority of hvCpGs are within the top 20% of variable loci in almost all covered datasets.

We next compared the set of 4330 hvCpGs with an alternative set obtained using the same method but without prior adjustment of each dataset for the first ten PCs. This alternative set contained only 1302 CpGs, which confirmed our intuition that PC adjustment maximizes power to identify true dataset-independent hypervariability by removing unwanted technical variation (Supplementary Figure S8). Finally, we used reported measures of methylation variability among technical replicates (23,50) to remove 187 technically unreliable probes (see Methods), leaving a final set of 4143 hvCpGs (Table 1; Supplementary Table S5).

hvCpGs are enriched for intermediate methylation values in all datasets compared to the array background (Supplementary Figure S2; see Table 1 for definition of array background) and are distributed throughout the genome (Supplementary Figure S9A, B), with 2219 (54%) falling within 716 'clusters' containing two or more hvCpGs separated by <4 kb (Supplementary Figure S5B). To account for the possibility that our downstream analyses may be biased by these distributional properties, we generated a set of controls that were distribution-matched in a whole blood dataset (Supplementary Figure S3) and a set of 'declustered hvCpGs' (Table 1, 'Materials and Methods').

hvCpG variability is not driven by age, sex or cell heterogeneity

Evidence from multiple studies suggests that methylation variability can increase with age (termed epigenetic drift) (11,64), raising the possibility that cross-dataset hypervariability of hvCpGs is driven in part by a large proportion of adult/elderly samples. However, 3815 (92%) out of 4122 hvCpGs with methylation measured in cord blood and/or buccal samples from infants showed methylation variance within the top 5% of CpGs in those datasets (Supplementary Table S10), suggesting that high variability at hvCpGs arises in early life. We further probed age stability of hvCpGs by leveraging two studies of age effects in blood.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

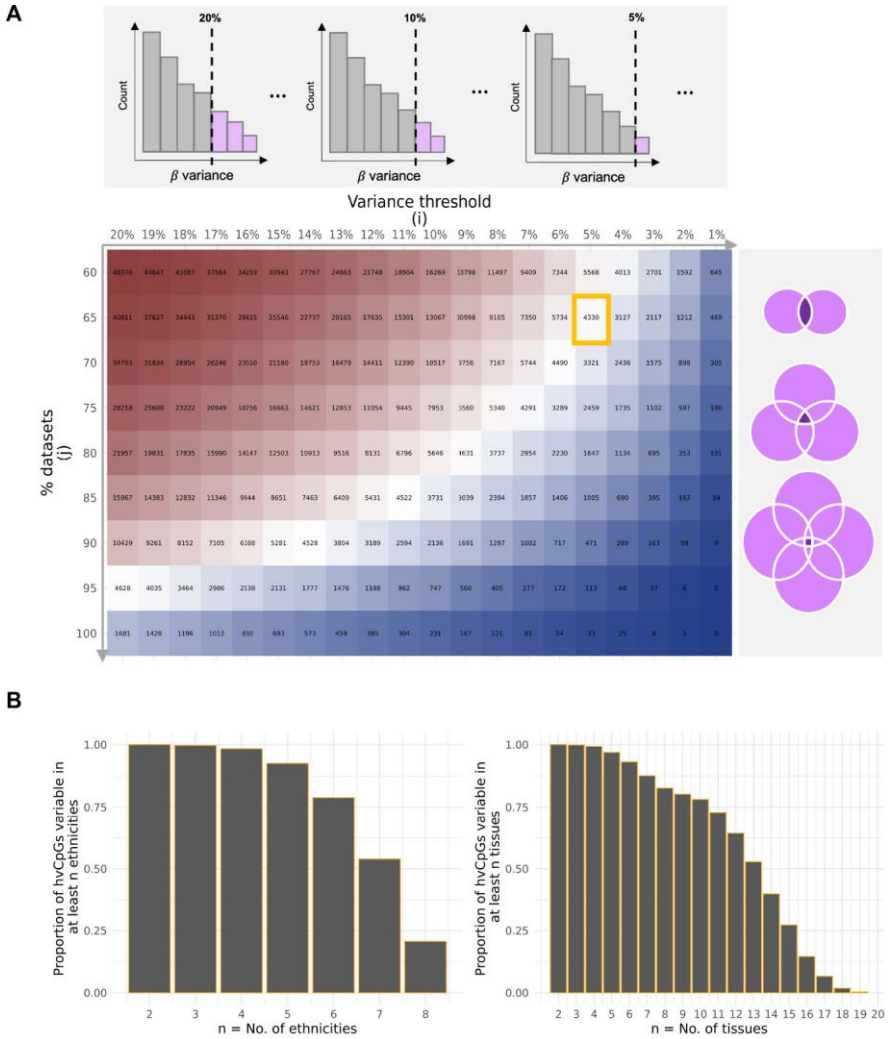


Figure 1. Identification of tissue- and ethnicity- independent hypervariable CpGs ('hvCpGs'). (**A**) CpGs with methylation Beta variance within the top 20%, 19%, 18%… 1% of variable CpGs were first extracted from each of the 30 methylation datasets used in this study. The intersection of these CpGs was then taken over an increasing proportion of datasets, requiring each CpG to be present in a minimum of 15 of the datasets analysed. The heatmap shows the number of CpGs within the top *i* % of variable sites by methylation Beta variance ('variance threshold') overlapping at least *j* % of datasets. To identify hypervariable CpGs ('hvCpGs'), we set a threshold at $i,j = [5,65]$, marked by the orange box. (**B**) Bar charts showing the proportion of the set of 4330 hvCpGs identified using $i,j = [5,65]$ that have top 5% methylation Beta variance in $\geq n$ ethnicities (left) and tissues (right). See Supplementary Table S4 for groupings of datasets by tissue type and ethnicity.

The first study reported methylation consistency in individuals sampled at two time points six years apart using intraclass correlation coefficients (ICCs) (54). Because ICCs increase with CpG variability, we compared temporal stability of hvCpGs to controls with similar methylation Beta distributions selected at the first time point ('Methods'). The temporal stability of hvCpGs was significantly greater than that of controls (Wilcox paired signed-rank test P-value $< 5.7 \times 10^{-81}$), with 95% of hvCpGs considered temporally stable versus 89% of controls (Supplementary Figure S6A). The second measured epigenetic drift in a cross-sectional study of 3295 whole blood samples from individuals aged 18 to 88 (11). Only 7% of hvCpGs overlapped CpGs that show increased methylation variability with age, compared to 16.5% of blood distribution-matched controls (Supplementary Figure S6B). This suggests that that the majority of hvCpGs are stable over a broad time period in whole blood and further supports the notion that hypervariability of hvCpGs in multiple datasets is not an artefact of epigenetic drift effects.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

Methylation values were pre-adjusted for the first ten PCs and for sex in all datasets where sex was available as a covariate (24 out of 30 datasets). We investigated the potential for unaccounted-for sex effects to drive methylation variance at hvCpGs by constructing male-only and female-only datasets ('Materials and Methods'). 100% of hvCpGs were in the top 5% of CpGs by methylation variance in at least one of the 'male-only' and 'female-only' datasets analysed. Similarly, 3548 (96%) of the 3678 hvCpGs covered in purified CD4+ and CD8+ datasets had methylation variance among the top 5% in at least one dataset (Supplementary Table S10), suggesting that methylation variation at hvCpGs was not driven by unaccounted-for cell heterogeneity effects amongst the heterogeneous tissue types studied.

Together, these data strongly suggest that variability at hvCpGs is not driven by sex, age or cell heterogeneity effects.

Hypervariability is not driven by genetic variants

Genetic variation is an important driver of interindividual methylation differences (4,5). There is evidence that mQTLs can be shared across different tissues (15,16,65,66) and ethnic groups (5), raising the possibility that 'universal' (multi-tissue and multi-ethnic) mQTLs might drive cross-dataset variability at hvCpGs. We therefore investigated the potential influence of methylation quantitative trait loci (mQTL) on methylation variability at hvCpGs by leveraging a recently published large meta-GWAS (36 cohorts, $n = 27\,750$ individuals) that identified common genetic variants associated with methylation in blood from Europeans (51), reasoning that by definition 'universal' mQTLs would be included in this meta-analysis.

We considered multiple methylation variance thresholds (5%, 10% and 20%) and observed a positive relationship between hypervariability and both the probability of a significant mQTL association and the mean mQTL effect size (Figure 2A). Amongst the set of 4143 hvCpGs, there were 6985 *cis* mQTL (covering 3635 hvCpGs and 6417 SNPs) and 971 *trans* mQTL (covering 713 hvCpGs and 753 SNPs). Overall, 3722 (90%) hvCpGs were reported to be associated with at least one (*cis* or *trans*) mQTL. The median of the mean % variance explained by mQTLs was 4% (Figure 2B), suggesting that additive genetic effects explain a small to moderate proportion of methylation variability at the majority of these hypervariable loci in blood. Noting that the statistical power to detect mQTL associations will be greater at loci that are inherently variable, we matched hvCpGs to CpGs with the same number of mQTL associations and similar mean % variance explained by mQTL ('mQTL-matched controls', Table 1, Supplementary Figure S4). hvCpGs showed an average 5-fold increase in methylation variance compared to mQTL-matched controls across datasets (Figure 2C), further supporting the notion that methylation variation at hvCpGs is not principally driven by universal genetic effects.

To further probe the influence of genetic effects on hvCpG methylation we examined the overlap between hvCpGs and a published set of CpGs that show DNAm variation between monozygotic (MZ) co-twins that is *equivalently variable (ev)* to that between unrelated individuals, suggestive of genetically independent variable methylation establishment after MZ twin splitting (53). In total, hvCpGs comprise 122 (42%) of the 317 evCpGs identified in blood (1.9-fold enrichment relative to distribution-matched controls) and 62% of those that were replicated as evCpGs in adipose tissue (2.8-fold enrichment relative to controls) (Supplementary Table S11), supporting the notion that hvCpGs are likely influenced but not determined by genetic variation in multiple tissues.

hvCpGs show covariation across tissues derived from different germ layers

DNAm states that are variable in different tissues and that are influenced but not determined by genotype may have been established before germ layer separation in early embryonic development and may therefore covary across tissues derived from different germ layers (28). None of the 30 datasets used to identify hvCpGs had multi-tissue data from the same individuals. We therefore examined the overlap between hvCpGs and 3089 CpGs that show systemic (cross-tissue) interindividual variation (SIV), collated from four published sources (26–29) (Supplementary Table S6). Because both SIV-CpGs and hvCpGs are enriched for intermediate methylation states (28), we used the set of blood distribution-matched controls (Table 1) as a comparator to ensure that our analysis was not biased by this shared property. 24% of hvCpGs overlap a known SIV-CpG, showing a ~5-fold enrichment for SIV-CpGs relative to blood distribution-matched controls (Figure 3A, Supplementary Table S12, Supplementary Figure S10A). We note that a further 540 (13%) hvCpGs are within 1 kb of a SIV-CpG, ~5-fold greater than array background CpGs. This suggests that many hvCpGs directly overlap or co-localize with a known SIV-CpG.

The set of all hvCpGs comprises 32.1% of the 3089 CpGs reported as SIV in any of the four independent studies analysed despite comprising <1% of the 450K array. When considering 'high-confidence' SIV-CpGs reported in at least two or three of the four screens, the proportion identified rises to 76.5% and 95.1% respectively (Figure 3B).This suggests that our approach of identifying hypervariable loci across multiple datasets may be a more powerful method for identifying putative SIV loci, compared to existing SIV screens that necessarily rely on rare datasets with multi-tissue, multi-germ layer methylation data from small numbers of individuals. To confirm this, we estimated the power to detect SIV using the multi-tissue data from four individuals analysed by van Baak *et al.* (27). Using a permutation framework ('Methods'), we estimated the mean power to detect SIV as 56% (median [IQR] = 0.58 [0.44, 0.72]; Supplementary Figure S7). As expected, given the small sample size of this multi-tissue dataset, a large proportion of hvCpGs (75%) did not meet the minimum interindividual variation threshold of 0.2 used by van Baak *et al.* to define SIV. On the assumption that hvCpGs are highly enriched for true SIV, this could explain why hvCpGs constitute 61.7% of the van Baak *et al.* SIV-CpGs, while just

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

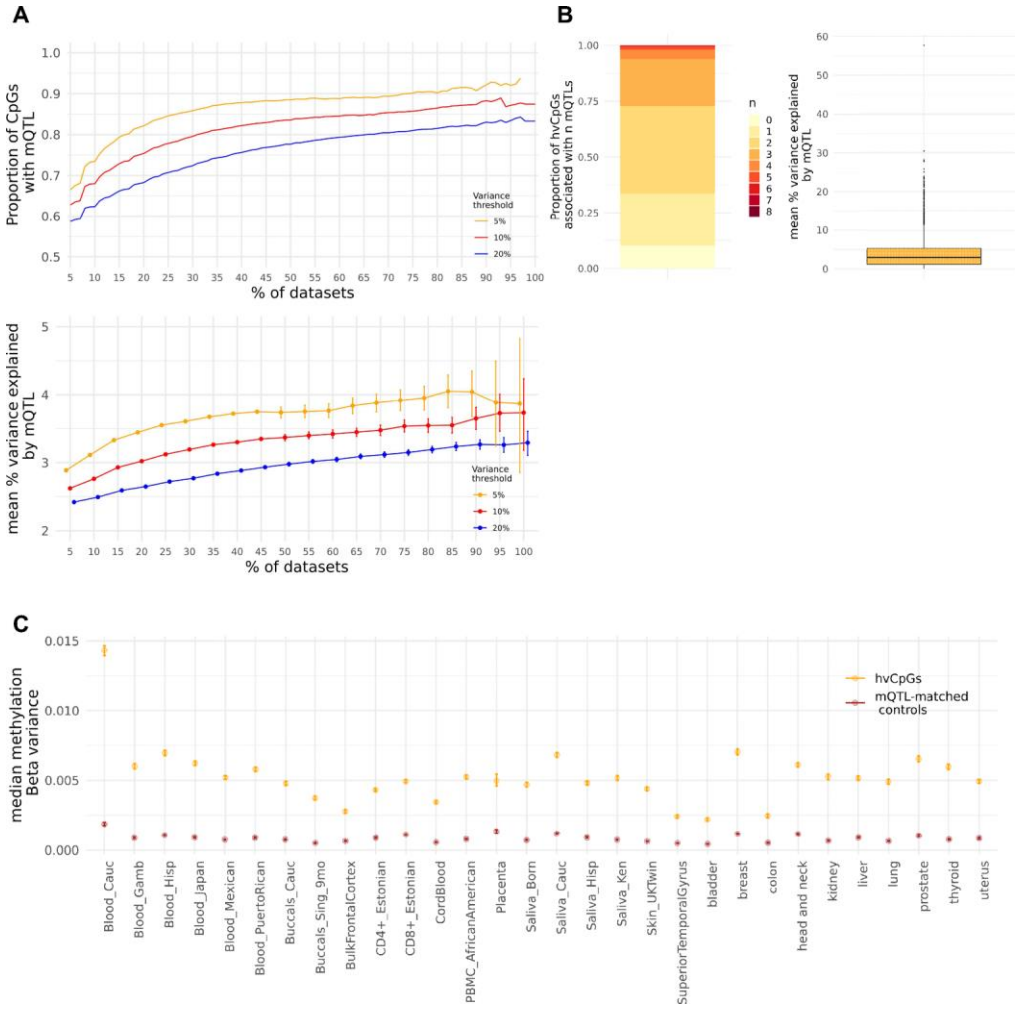


Figure 2. Genetic effects at hvCpGs using mQTL data from a large meta GWAS in blood (Min et al. 2021). (**A**) The relationship between hypervariability and the proportion of CpGs with at least one mQTL association (top) and the mean mQTL effect size (bottom). Coloured curves represent CpGs with top 5% (orange), 10% (red) and 20% (blue) methylation Beta variance in at least *x%* of datasets. (**B**) mQTL effects at hvCpGs**.** Left: the proportion of hvCpGs that are associated with *n* mQTLs. Right: the distribution of the mean % variance explained by mQTLs at 3722 hvCpGs that are associated with at least one mQTL. (**C**) Median methylation Beta variance at 3722 hvCpGs overlapping the 'Blood_Cauc' dataset (orange) and corresponding controls matched on number of mQTL associations and mean % explained by mQTL ('mQTL-matched controls', Table 1; Supplementary Figure S4), in each dataset. See Supplementary Tables 1–4 for further details on tissues and ethnicities. Error bars in A and C are bootstrapped 95% confidence intervals. Note, error bars in (C) are very small.

13.5% of hvCpGs are identified as SIV-CpGs in the van Baak *et al*. analysis.

To directly test our hypothesis that hvCpGs comprise previously unidentified SIV loci, we analysed a dataset of fetal tissues from 27 individuals, each with methylation data from two tissues derived from different germ layers (see Supplementary Table S7). Inter-germ layer correlations at hvCpGs had a median average Pearson *r* of 0.42, compared to array background CpGs which had a median average Pearson *r* of 0.05 (Figure 3C left). Of the 3878 hvCpGs covered in this fetal multi-tissue dataset, 1653 (42%) had an average inter-germ layer Pearson $r \geq 0.5$. Of these, 58% did not overlap previously identified SIV loci, suggesting that hvCpGs comprise novel SIV loci. A comparison of the average inter-germ layer correlation at hvCpGs and at previously identified SIV-CpGs showed that hvCpGs and SIV-CpGs had similar inter-germ layer correlations (Figure 3C, right).

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

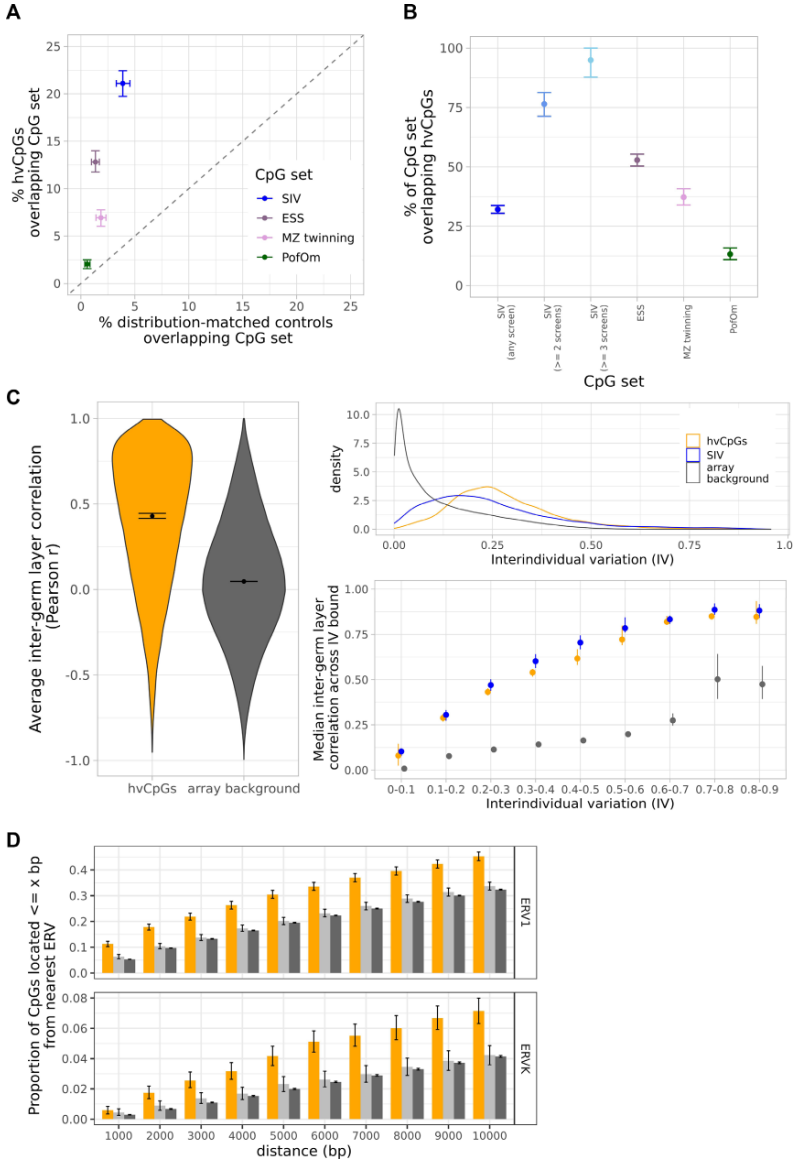


Figure 3. hvCpGs are enriched for loci and genomic features linked to variable methylation establishment in early development. (**A**) The proportion of 3566 hvCpGs (y-axis) vs corresponding distribution-matched controls (x-axis) covered in the 'Blood_Cauc' dataset that overlap 3089 SIV-CpGs, 1217 ESS CpGs identified by van Baak *et al.* (2018), 728 'MZ twinning' CpGs identified by van Dongen *et al.* (2021) and 732 PofOm CpGs identified by Zink *et al.* (2018). (**B**) The proportion of SIV-CpGs, ESS CpGs, MZ twinning CpGs and PofOm CpGs that are hvCpGs. SIV-CpGs identified in at least two or three independent screens were also included in this plot. (**C**) Inter-germ layer correlations at hvCpGs using a fetal multi-tissue dataset that comprises methylation data from 10 individuals with endoderm- and ectoderm- derived tissues, 9 individuals with endoderm- and mesoderm- derived tissues and 8 individuals with mesoderm- and ectoderm- derived tissues (see Supplementary Table S7). Left: The distribution of average inter-germ layer correlations at 3878 hvCpGs (orange) and 372 571 array background CpGs (excluding previously identified SIV CpGs and hvCpGs) (dark grey) covered in the fetal multi-tissue dataset. Top right: Interindividual variation at 3878 hvCpGs (orange), 4076 previously identified SIV loci (blue) covered in the fetal multi-tissue dataset, and 372 571 array background CpGs (see 'Materials and Methods' for definition of interindividual variation). Bottom right: Comparison of average inter-germ layer correlations at hvCpGs, SIV- CpGs and array background CpGs, stratified by interindividual variation. Each point indicates the median average inter-germ layer correlation for those CpGs with interindividual variation falling within each bound specified on the x-axis. (**D**) The proportion of 3566 hvCpGs, distribution-matched controls and array background CpGs that are $\leq x$ bp from the nearest ERV1 and ERVK transposable elements determined by RepeatMasker. Error bars in all panels are bootstrapped 95% confidence intervals. SIV = systemic interindividual variation, ESS = epigenetic supersimilarity, PofOm = parent-of-origin-specific methylation.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

hvCpGs are enriched for loci with distinctive methylation patterns in MZ twins

We further investigated evidence for establishment of hvCpG methylation states in the early embryo by testing the overlap between hvCpGs and a published set of 1217 'epigenetic supersimilarity' (ESS) CpGs overlapping array background. ESS CpGs show high interindividual variation with greater-than-expected methylation concordance between monozygotic co-twins in adipose tissue, suggestive of methylation establishment in the early zygote before MZ cleavage (27). 13% of hvCpGs overlap an ESS CpG, showing a ∼9.5-fold enrichment for ESS CpGs relative to distribution-matched controls (Figure 3A, Supplementary Table S12, Supplementary Figure S10B).

We next examined the overlap between hvCpGs and a published set of CpGs showing a unique methylation signature in adult tissues from MZ vs DZ twins ('MZ twinning CpGs', Table 2), linked to MZ twin splitting events in early development (56). 7% of hvCpGs overlap an MZ twinning CpG, showing a 3.7-fold enrichment for MZ twinning CpGs compared to distribution-matched controls (Figure 3A, Supplementary Table S12).

Notably, 54% of ESS and 37% of MZ twinning CpGs overlapping array background are hvCpGs (Figure 3B).

Reconciling the timing of variable methylation establishment at hvCpGs

The enrichments that we observe for SIV, ESS, evCpGs and MZ twinning CpGs offer a potential insight into the timing of methylation establishment at hvCpGs. In total, 38% of hvCpGs overlap at least one of these CpG sets (Supplementary Figure S11A) and enrichment is stronger amongst CpGs that show at least two of these properties (Supplementary Figure S11B). In particular, hvCpGs comprise 78% of SIV-ESS loci and 65% of SIV-MZ twinning loci, suggesting that SIV loci with evidence of establishment in the pre-gastrulation embryo are enriched for hvCpGs.

Variable methylation states identified at evCpGs are thought to originate in embryonic development and/or early post-natal life (53). We note that 41 out of 317 evCpGs overlap SIV and/or MZ twinning CpGs, suggesting that at least a subset may be established in the pre-gastrulation embryo. hvCpGs comprise 67% of evCpGs that overlap SIV-CpGs, and 76% of that overlap MZ twinning CpGs (Supplementary Figure S11B).

hvCpGs are enriched for parent-of-origin methylation and proximal TEs

In mice, variable methylation states have been associated with the Intracisternal A Particle (IAP) class of endogenous retrovirus (67,68), with growing evidence that methylation variability may in part be driven by incomplete silencing of IAPs in early development (69,70). In humans, SIV-CpGs are enriched for proximal endogenous retrovirus elements (ERVs), including the subclasses ERV1 and ERVK (28). This is also the case with hvCpGs: 45% and 7% of hvCpGs are located within 10 kb of an ERV1 and ERVK element respectively, representing a ∼1.3-fold and ∼1.7-fold enrichment relative to both array background and blood distribution-matched controls (Figure 3D, Supplementary Figure S10 C, Supplementary Table S12). Approximately 4.7% of hvCpGs are also located within 1Mb of telomeric regions, showing a 1.8-fold enrichment relative to distribution-matched controls and array background CpGs (Supplementary Table S12).

Maintenance of parent of origin-specific methylation (PofOm) in the pre-implantation embryo is critical for genomic imprinting (71), and several previously identified SIV loci have been found to be associated with imprinted genes and/or PofOm (25,27,57). 58 hvCpGs (1.4%) were annotated to 32 imprinted genes (Supplementary Table S13), no more than expected by chance since 1.9% of array background CpGs are annotated to imprinted genes. 10 hvCpGs were annotated to the polymorphically imprinted non-coding RNA *VTRNA2-1*, a well-established SIV locus that is associated with periconceptional environmental exposures (25,27,30,72,73). Although only a small proportion (2.2%) of hvCpGs overlap regions of PofOm identified in peripheral blood (58), this overlap represents a 3.5-fold and 11-fold enrichment relative to distribution-matched controls and array background respectively that is maintained after de-clustering (Figure 3A, Supplementary Figure S10D, Supplementary Table S12). This overlap constitutes 13% of all PofOm CpGs overlapping array background (Figure 3B).

hvCpGs show sensitivity to pre-natal environment

Variable methylation states established in early development that are sensitive to environmental perturbation are promising candidates for exploring the developmental origins of health and disease (74–76). We explored whether hvCpGs show sensitivity to pre-natal environment by examining their overlap with loci associated with season of conception ('SoC') in a rural Gambian population exposed to seasonal fluctuations in diet and other factors (77–79). hvCpGs comprise 70 (29%) out of 242 previously identified SoC-CpGs (57) overlapping array background, an approximately 3-fold enrichment relative to distribution-matched controls (Supplementary Table S11).

We next leveraged a recent meta-analysis of 2365 cord blood samples that modelled genetic (G), genetic by environment (GxE) and additive genetic and environment (G + E) effects at variably methylated probes, where E represents a range of prenatal exposures including pre-pregnancy BMI, maternal smoking, gestational age, hypertension, anxiety and depression (14). Of the 703 hvCpGs overlapping the neonatal blood variably methylated regions explored in that study, G, GxE, and G + E effects were the 'winning' models for 30%, 30% and 40% of probes respectively, representing an increase in G + E effects compared to array background (Supplementary Figure S12). This analysis supports our intuition that hvCpGs are influenced but not determined by genetic variation, with pre-natal environment as an additional influencing factor.

Chromatin states at hvCpGs

Compared to array background, hvCpGs are enriched within intergenic regions and CpG island 'shores' but are

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

depleted within gene bodies and regions directly upstream of transcription start sites (Supplementary Figure S9C). We predicted chromatin states at hvCpGs by examining the overlaps of hvCpGs with histone modifications using the chromHMM 15-state model (61) for seven tissues including embryonic stem cells (H1 ESCs), and fetal and adult tissues (62). Although many hvCpGs were associated with regulatory elements in all tissues, hvCpGs were generally depleted in these regions compared to array background, except within predicted enhancers in H1 ESCs (Supplementary Figure S13).

Gene expression and ontology analysis

409 hvCpG clusters (corresponding to 1282 CpGs) are annotated to 425 genes in the Illumina450k manifest. Analysis of GTEx expression data reveals that these are expressed in a diverse range of tissues. (Supplementary Figure 14). Gene ontology enrichment analysis revealed that hvCpGs were significantly enriched for terms associated with cell-cell adhesion (Figure 4A), which is largely driven by the colocalization of 3.3% of hvCpGs to clustered protocadherin (*cPCDH*) genes on chromosome 5. This region comprises three clusters of protocadherin genes (*cPCDHα, cPCDHβ, cPCDHγ*), each containing many variable exons whose promoter choice is determined stochastically via differential methylation by DNA-methyltransferase 3 beta (DNMT3B) in early embryonic development (80,81), resulting in the expression of distinct *cPCDH* isoforms of cell-surface proteins that are critical for establishing neuronal circuits (82). The *cPCDH* gene locus has also been found to be influenced by age (11,83–85). Accordingly, although a minority (5%) of hvCpGs showed evidence of epigenetic drift in blood (11), these are enriched within the *cPCDH* locus relative to those that did not show evidence of epigenetic drift (Fisher's Exact Test (FET) P-value $= 9.4 \times 10^{-9}$, OR = 4.02). Hypervariable methylation states at the *cPCDH* gene locus may therefore be driven by early developmental and/or aging effects. Noting that evCpGs and MZ twinning CpGs (Table 2) have also been reported to colocalize with this locus (53,86), hvCpGs annotated to *cPCDH* genes were ~8.5-fold enriched for MZ twinning CpGs (FET P-value $= 1.04 \times 10^{-22}$) and ~3-fold enriched for evCpGs (FET P-value $= 1.6 \times 10^{-3}$) relative to hvCpGs that were not.

Association of hvCpGs with reported EWAS trait associations

To probe the potential functional role of hvCpGs, we analysed their overlap with traits reported in the epigenome-wide association studies (EWAS) catalogue (http://ewascatalog.org/). 86% of hvCpGs show significant associations (reported P-value $< 1 \times 10^{-4}$) with one or more of 231 unique traits covered in the catalogue (Supplementary Table S9). However, compared to blood distribution-matched controls, a suitable comparator given that the majority of EWAS have been carried out in blood, we found that hvCpGs were enriched amongst CpGs associated with sex, Alzheimer's disease and inflammatory bowel disease only (Figure 4B).

Noting that all sex-associated hvCpGs have top 5% methylation Beta variance in at least one of our eight generated female-only and male-only datasets respectively ('Materials and Methods'), and that a similar proportion of SIV-CpGs are also associated with sex (23% of hvCpGs and 20% of the 3089 SIV-CpGs considered in our study), we speculate that the association with sex may be a feature of variable methylation states established in early development. Amongst the 64 hvCpGs associated with Alzheimer's disease, 25 overlap previously identified SIV and/or ESS loci, 9 of which annotated to *CYP2E1*, a gene that has also been associated with Parkinson's disease and rheumatoid arthritis (32,87). Amongst the 200 hvCpGs associated with inflammatory bowel disease, 87 overlap a SIV/ESS locus, 9 of which are annotated to *C21orf56*, a gene at which offspring methylation has been associated with maternal folate levels in pregnancy (88).

hvCpGs were notably depleted amongst age-related traits relative to distribution-matched controls (Figure 4B), in agreement with our earlier findings that hvCpGs are largely stable with age (Supplementary Figure S6). hvCpGs are also depleted amongst CpGs that are differentially methylated between buccal cells and peripheral blood mononucleocytes ('Tissue' in Figure 4B), supporting the notion that hvCpGs may be established before cell differentiation and that the method used to identify the hvCpGs is robust to tissue-specific methylation variation.

DISCUSSION

We have identified and characterized tissue- and ethnicity-independent hypervariable methylation states at CpGs covered by the 450k array. Our methodological approach was designed to be robust to dataset-specific drivers of methylation variability, including sex, age, cell type heterogeneity and technical artefacts. We identified 4143 hvCpGs and found strong evidence that methylation states at many hvCpGs are likely to be established in the early embryo and are stable postnatally. Our analysis positions hvCpGs as tissue- and ethnicity- independent age-stable biomarkers of early stochastic and/or environmental effects on DNA methylation.

hvCpGs cover ~1% of the 450K array and were in the top 5% variable methylation states in an average of 13 distinct tissues and 7 ethnicities. Our study is not the first to investigate DNAm patterns in multiple tissues. Previous studies have identified CpGs that are differentially methylated between tissues (89–91); determined the extent to which variable methylation states in accessible tissues (such as blood) reflect those in inaccessible tissues such as brain (65,90–93); compared methylation patterns between peripheral tissues (66,94,95); directly identified SIV loci using tissues derived from different germ layers (24–26,28,29); functionally characterized tissue-specific variably methylated regions (96); and examined the extent to which common drivers of methylation variation, such as genetics, age, sex and environment, are tissue-specific (8,12,15–18,97,98). The majority of these studies used a comparatively small number of tissues or cell-types, and few have used multi-tissue datasets from different ethnicities (15). To our knowledge, ours is the first study to explore the extent to which vari-

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

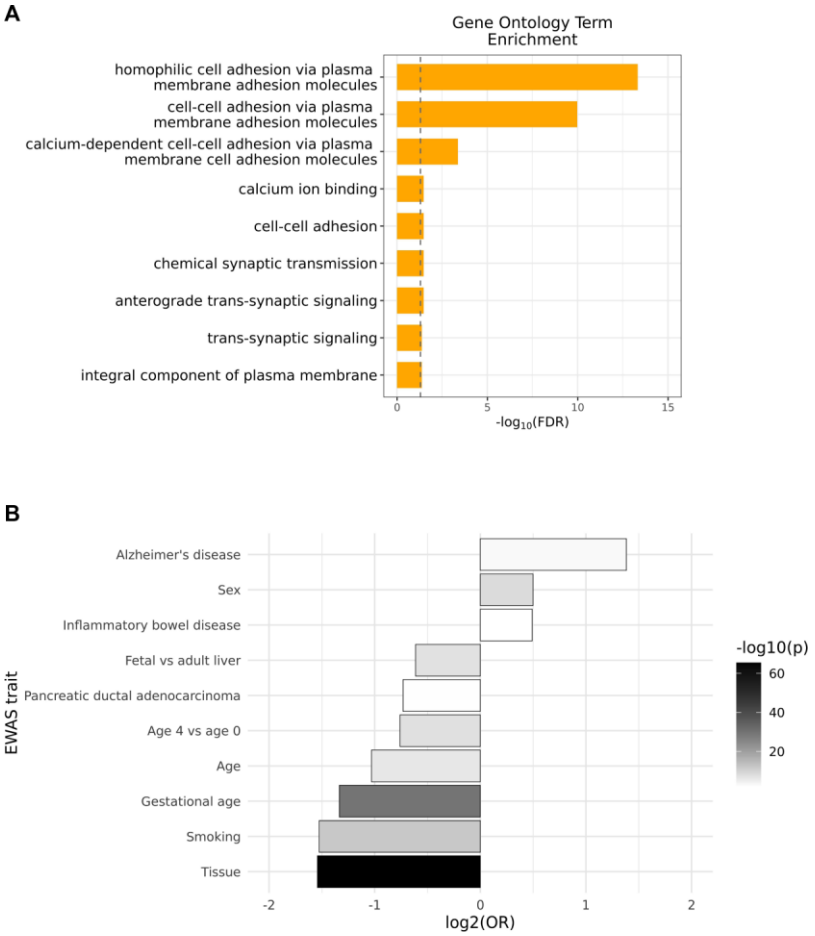


Figure 4. Functional annotation of hvCpGs. (**A**) Gene ontology term enrichment analysis at hvCpGs. Vertical line indicates a significance threshold of FDR < 0.05. (**B**) EWAS trait enrichment of hvCpGs relative to blood-distribution controls (Table 1) for traits overlapping at least 1% of hvCpGs (see 'Materials and Methods' and Supplementary Table S9 for further details). X-axis gives enrichment odds ratios and bar colour gives Fisher's Exact Test (FET) significance *P*-values. Shown are the 10 traits with FET *P*-value ≤ 0.01.

ably methylated CpGs are shared across diverse tissues and ethnicities in the human genome.

The majority of hvCpGs were associated with at least one mQTL suggesting that genetic effects influence methylation at these loci. Although data on the total proportion of methylation variance explained by all mQTLs associated with each hvCpG were not available, our comparison with mQTL-matched controls together with evidence of enrichment for sensitivity to periconceptional environment suggests that stochastic and/or environmental effects have a relatively large influence on methylation variability at hvCpGs. This is supported by evidence of methylation discordance between MZ twins, although we note that MZ discordance can be driven by *de novo* genetic mutations after MZ twinning events (99). A large proportion of hvCpGs show evidence of systemic interindividual variation (SIV), that is, intra-individual correlation in methylation across tissues derived from different germ layers. Whilst loci that covary across different tissue types are enriched for mQTL effects (16,65,66,94), it has been suggested that SIV loci are putative human metastable epialleles with variable methylation states established before gastrulation that are influenced but not determined by genetic variation (28).

Our fetal multi-tissue analysis supports the notion that SIV at hvCpGs arises during early development and is likely not, for example, driven by post-natal environmental influences that act across many tissues. hvCpGs were also highly enriched for epigenetic supersimilarity loci and MZ twinning-associated CpGs, both of which have been linked to establishment of methylation in the cleavage stage preimplantation embryo (27,56). The degree of overlap between variably methylated regions in different cell types has also been linked to their common developmental origin (96). If this pattern holds true, it follows that stochastic

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

and/or environmentally influenced variably methylated loci that are shared across a large number of diverse tissues are likely to have originated before germ-layer differentiation. Definitive proof of this would require an analysis of methylation variation at multiple stages in a sufficient number of pre-gastrulation embryos.

Although examination of EWAS trait associations revealed no evidence that hvCpGs are enriched for post-natal environmental effects, it is possible that cross-tissue and cross-ethnicity variable methylation states at some hvCpGs are influenced by later gestational or post-natal environmental effects. Such effects may act in addition to or independently of early environmental effects across multiple tissues, as has been suggested at the *VTRNA2-1* locus in the context of folate supplementation in pregnancy (100), maternal age at delivery (73), and smoking (101).

An interesting feature of hvCpGs is their enrichment for intermediate methylation values relative to the array background. A previous read-level analysis of SIV loci from human embryos using reduced representation bisulphite sequencing data indicated that intermediate methylation states at SIV loci are driven by stochastic, cell variegation effects rather than allele-specific methylation (28), and this may also be the case for hvCpGs.

The association of hvCpGs with parent-of-origin-specific methylation and proximal ERV1 and ERVK elements is notable because these features have been linked to SIV-CpGs (28). This suggests that genomic regions targeted by epigenetic silencing or maintenance mechanisms during early embryonic reprogramming may be enriched for stochastic and/or environmentally influenced methylation variation. For example, it has been suggested that regions of PoFOm may be vulnerable to stochastic or environmentally-sensitive loss of methylation on the usually-methylated allele or gain of methylation on the usually-unmethylated allele at a later time-point, leading to interindividual methylation variation (57,71,102). Similarly, certain IAP elements (a class of ERVK LTR retrotransposon) show methylation variation between isogenic mice (67,68) that in several cases can be influenced by pre-natal environment (103–105). Whilst transposable elements are usually silenced to prevent insertion events from damaging the genome, recent evidence suggests that methylation variability at IAP elements is partly driven by low-affinity binding of *trans-a*cting Krüppel-associated box (KRAB)-containing zinc finger proteins (KZFPs) (69) and by sequence variation in KZFP-binding sites (69,106). Whilst KZFPs are known to target TEs in humans (107,108), the extent of their role in driving methylation variation is an ongoing area of research.

The large overlap between hvCpGs and 'high confidence' SIV-CpGs identified in at least two independent screens suggests that the identification of hvCpGs might constitute a high-powered method for detecting novel SIV loci. Supporting this, the largest SIV screen to date with 10 individuals was reported to be underpowered to detect the well-established SIV locus at the non-coding RNA gene *VTRNA2-1* (29) (represented by 10 hvCpGs), and we found that a 4-individual multi-tissue dataset analysed by van Baak *et al*. (27) had limited power to detect SIV loci. Another consideration is that SIV screens to date have used different sets of tissues. Since loci that covary between one pair of tissues do not necessarily covary between another pair (65), the enrichment for high confidence SIV loci (i.e. those reported in multiple independent SIV screens) might reflect the fact that methylation states at hvCpGs covary across a large number of tissues. Importantly, our analysis of a fetal multi-tissue dataset offers a strong validation of previously unreported SIV at hvCpGs.

Our analysis of EWAS trait associations revealed a moderate enrichment for hvCpGs amongst CpGs associated with Alzheimer's disease and inflammatory bowel disease. SIV loci have been linked to this and other disease outcomes including autism, cancer and obesity (27,31,109). For example, 10 hvCpGs overlap the *PAX8* gene which is a known SIV locus. *PAX8* methylation measured in peripheral blood of Gambian 2-year olds was recently shown to be correlated with thyroid volume and hormone levels in the same children in mid-childhood, and the latter was associated with changes in body fat and bone mineral density (36). This suggests that hvCpGs are interesting candidates for exploring how stochastic and/or environmentally influenced DNAm states established in early development might influence lifelong health.

We identified hvCpGs that are variable in diverse ethnicities, raising the possibility that regions of hypervariable methylation may be a conserved feature in the human genome. It is also possible that there are ethnicity-specific regions of hypervariable methylation that would not have been captured in our analysis. Conserved variable methylation patterns established in the early embryo that are sensitive to early environment and that are able to influence gene expression might mediate a *predictive-adaptive-response* mechanism that senses the pre-natal environment in order to prime the developing embryo to its post-natal environment (75,76). One hypothesis suggests that stochastic methylation states that are genetically hardwired into the human genome could provide a means of rapid adaptation to changing local environments on a scale much faster than is attainable through Darwinian evolution (110). Alternatively, stochastic methylation arising in early development independently of environmental factors may increase population fitness by expanding the range of phenotypes in a given generation (111). Associations between genotype and methylation variance have been previously reported, for example at the putative metastable epiallele *PAX8* (36) at the master regulator of genomic imprinting *ZFP57* (27) and at several probes in the major histocompatibility complex (MHC) region associated with rheumatoid arthritis (112). Interestingly, 4% of hvCpGs are located within the MHC, representing an enrichment relative to the array background (FET P-value $= 2.7 \times 10^{-10}$, OR $= 1.7$). Further analysis of genotype-methylation variance effects is required to determine if this region, which contains a large amount of sequence variation and is implicated in many immune-mediated diseases (113), or others contain additional examples of genetically-driven phenotypic plasticity that is mediated by DNA methylation.

By selecting CpGs within the top 5% of methylation variance in at least 65% of datasets we were able to identify

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

CpGs that were highly variable across multiple tissues and ethnicities. A comparison of CpGs identified using slightly different thresholds suggested that the set of hvCpGs is relatively insensitive to these parameters, but we note that the final set of hvCpGs is nevertheless dependent on the choice of thresholds.

Our method of adjusting for the first 10 PCs of variation increased power to detect consistent variable methylation states across datasets by reducing technical artefacts in each dataset, although some true biological variation may have been removed by doing so. It remains the case that we may not have controlled for all non-biological sources of variation within each dataset. As such, any remaining inter dataset differences due to unaccounted for technical variation and/or different pre-processing and normalization steps already applied to public methylation data would result in a loss of power to detect hvCpGs. Conversely, if technical/normalization issues were to cause a CpG to be in the top 5% of variance in one dataset, this CpG would be unlikely to be in the top 5% of variance across a majority of datasets. Inherent control of false positives arising from residual technical differences between datasets is therefore a strength of our approach.

We analysed methylation data covering 19 different tissues and 8 ethnicities. While these data were sufficiently powerful to identify several thousand hvCpGs, future analyses are likely to identify additional loci through the inclusion of larger datasets from diverse tissues and ethnicities as they become available. Furthermore, the vast majority of publicly available methylation datasets use the Illumina 450K array. Therefore, a major limitation of this study is that we were only able to analyse the small proportion of the methylome covered by this array, which has been found to miss a disproportionate amount of variable CpGs (29). However, we note that our method for identifying hypervariable CpGs can easily be applied to whole methylome sequencing data which is becoming increasingly available.

Through the joint analysis of methylation data from multiple tissues, we have identified a large set of hypervariable loci on the 450K array that are present across multiple tissues and ethnicities. Comparisons with a diverse range of data sources reveal that stochastic and/or environmentally responsive methylation states at these loci are likely to have been established in the early embryo and appear to be stable with age, making them interesting candidates for studying the developmental origins of life-long health and disease.

DATA AVAILABILITY

The large majority of datasets analysed in this paper are in the public domain. GEO accession numbers and/or further details are provided in Supplementary Tables. A small number of analysed datasets have restricted access. Requests to access these should be submitted to the corresponding authors in the first instance with researcher access requiring an application to the relevant institutional review boards.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

GoDMC meta-GWAS summary statistics were kindly provided by Jordana Bell. This data has now been published (Min *et al.* 2021).

FUNDING

Biotechnology and Biological Sciences Research Council [BB/M009513/1, BB/R01356X/1]; UK Medical Research Council [MR/M01424X/1, MC_PC_MR/RO20183/1 and MR/N006208/1]; Wellcome Trust/Royal Society [098386/Z/12/Z]; French National Research Agency [ANR OCEOADAPTO and ANR PAPUAEVOL 20-CE12-0003-01]; Department of Biotechnology, Ministry of Science and Technology, India [BT/IN/DBT-MRC/DFID/24/GRC/2015–16]; the fetal tissues supplied are part of the UCL-Imperial Baby Bio Bank collection. https://directory.biobankinguk.org. Funding for open access charge: UKRI.

Conflict of interest statement. None declared.

REFERENCES

1. Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204–220.
2. Rakyan,V.K., Down,T.A., Balding,D.J. and Beck,S. (2011) Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, **12**, 529–541.
3. Lappalainen,T. and Greally,J.M. (2017) Associating cellular epigenetic models with human phenotypes. *Nat. Rev. Genet.*, **18**, 441–451.
4. 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 Biol.*, **17**, 61.
5. Bell,J.T., Pai,A.A., Pickrell,J.K., Gaffney,D.J., Pique-Regi,R., Degner,J.F., Gilad,Y. and Pritchard,J.K. (2011) DNA methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biol.*, **12**, R10.
6. Jaffe,A.E. and Irizarry,R.A. (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.*, **15**, R31.
7. Houseman,E.A., Kile,M.L., Christiani,D.C., Ince,T.A., Kelsey,K.T. and Marsit,C.J. (2016) Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinf.*, **17**, 259.
8. Singmann,P., Shem-Tov,D., Wahl,S., Grallert,H., Fiorito,G., Shin,S.-Y., Schramm,K., Wolf,P., Kunze,S., Baran,Y. *et al.* (2015) Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics Chromatin*, **8**, 43.
9. Yousefi,P., Huen,K., Davé,V., Barcellos,L., Eskenazi,B. and Holland,N. (2015) Sex differences in DNA methylation assessed by 450 k beadchip in newborns. *BMC Genomics*, **16**, 911.
10. Horvath,S. (2013) DNA methylation age of human tissues and cell types. *Genome Biol.*, **14**, R115.
11. Slieker,R.C., van Iterson,M., Luijk,R., Beekman,M., Zhernakova,D.v, Moed,M.H., Mei,H., van Galen,M., Deelen,P. and Bonder,M.J. (2016) Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol.*, **17**, 191.
12. Busche,S., Shao,X., Caron,M., Kwan,T., Allum,F., Cheung,W.A., Ge,B., Westfall,S., Simon,M.-M. and Barrett,A. (2015) Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. *Genome Biol.*, **16**, 290.
13. Hannon,E., Knox,O., Sugden,K., Burrage,J., Wong,C.C.Y., Belsky,D.W., Corcoran,D.L., Arseneault,L., Moffitt,T.E., Caspi,A. *et al.* (2018) Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genet.*, **14**, e1007544.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

14. Czamara,D., Eraslan,G., Page,C.M., Lahti,J., Lahti-Pulkkinen,M., Hämäläinen,E., Kajantie,E., Laivuori,H., Villa,P.M., Reynolds,R.M. *et al.* (2019) Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns. *Nat. Commun.*, **10**, 2548.
15. Smith,A.K., Kilaru,V., Kocak,M., Almli,L.M., Mercer,K.B., Ressler,K.J., Tylavsky,F.A. and Conneely,K.N. (2014) Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*, **15**, 145.
16. Lin,D., Chen,J., Perrone-Bizzozero,N., Bustillo,J.R., Du,Y., Calhoun,V.D. and Liu,J. (2018) Characterization of cross-tissue genetic-epigenetic effects and their patterns in schizophrenia. *Genome Med.*, **10**, 13.
17. Day,K., Waite,L.L., Thalacker-Mercer,A., West,A., Bamman,M.M., Brooks,J.D., Myers,R.M. and Absher,D. (2013) Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol.*, **14**, R102.
18. Slieker,R.C., Relton,C.L., Gaunt,T.R., Slagboom,P.E. and Heijmans,B.T. (2018) Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics Chromatin*, **11**, 25.
19. Bibikova,M., Barnes,B., Tsan,C., Ho,V., Klotzle,B., Le,J.M., Delano,D., Zhang,L., Schroth,G.P., Gunderson,K.L. *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.
20. Price,E.M. and Robinson,W.P. (2018) Adjusting for batch effects in DNA methylation microarray data, a lesson learned. *Front. Genet.*, **0**, 83.
21. Leek,J.T., Scharpf,R.B., Bravo,H.C., Simcha,D., Langmead,B., Johnson,W.E., Geman,D., Baggerly,K. and Irizarry,R.A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.
22. Harper,K.N., Peters,B.A. and Gamble,M.v. (2013) Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. *Cancer Epidemiol. Biomarkers Prev.*, **22**, 1052–1060.
23. Sugden,K., Hannon,E.J., Arseneault,L., Belsky,D.W., Corcoran,D.L., Fisher,H.L., Houts,R.M., Kandaswamy,R., Moffitt,T.E., Poulton,R. *et al.* (2020) Patterns of reliability: assessing the reproducibility and integrity of DNA methylation measurement. *Patterns*, **1**, 100014.
24. Waterland,R.A., Kellermayer,R., Laritsky,E., Rayco-Solon,P., Harris,R.A., Travisano,M., Zhang,W., Torskaya,M.S., Zhang,J., Shen,L. *et al.* (2010) Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles. *PLoS Genet.*, **6**, e1001252.
25. Silver,M.J., Kessler,N.J., Hennig,B.J., Dominguez-Salas,P., Laritsky,E., Baker,M.S., Coarfa,C., Hernandez-Vargas,H., Castelino,J.M., Routledge,M.N. *et al.* (2015) Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome Biol.*, **16**, 118.
26. Harris,Alan, Nagy-Szakal,R. and Kellermayer,R. (2013) Human metastable epiallele candidates link to common disorders. *Epigenetics*, **8**, 157–163.
27. 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 Biol.*, **19**, 2.
28. 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. *Sci. Adv.*, **4**, eaat2624.
29. Gunasekara,C.J., Scott,C.A., Laritsky,E., Baker,M.S., MacKay,H., Duryea,J.D., Kessler,N.J., Hellenthal,G., Wood,A.C., Hodges,K.R. *et al.* (2019) A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol.*, **20**, 105.
30. Finer,S., Iqbal,M.S., Lowe,R., Ogunkolade,B.W., Pervin,S., Mathews,C., Smart,M., Alam,D.S. and Hitman,G.A. (2016) Is famine exposure during developmental life in rural bangladesh associated with a metabolic and epigenetic signature in young adulthood? A historical cohort study. *BMJ Open*, **6**, e011768.
31. Kühnen,P., Handke,D., Waterland,R.A., Hennig,B.J., Silver,M., Fulford,A.J., Dominguez-Salas,P., Moore,S.E., Prentice,A.M., Spranger,J. *et al.* (2016) Interindividual variation in DNA methylation at a putative POMC metastable epiallele is associated with obesity. *Cell Metab.*, **24**, 502–509.
32. Mok,A., Rhead,B., Holingue,C., Shao,X., Quach,H.L., Quach,D., Sinclair,E., Graf,J., Imboden,J., Link,T. *et al.* (2018) Hypomethylation of CYP2E1 and DUSP22 promoters associated with disease activity and erosive disease among rheumatoid arthritis patients. *Arthritis Rheumatol.*, **70**, 528–536.
33. Zhu,Y., Mordaunt,C.E., Yasui,D.H., Marathe,R., Coulson,R.L., Dunaway,K.W., Jianu,J.M., Walker,C.K., Ozonoff,S., Hertz-Picciotto,I. *et al.* (2019) Placental DNA methylation levels at CYP2E1 and IRS2 are associated with child outcome in a prospective autism study. *Hum. Mol. Genet.*, **28**, 2659–2674.
34. Sanchez-Mut,J., Heyn,H., Silva,B.A., Dixsaut,L., Garcia-Esparcia,P., Vidal,E., Sayols,S., Glauser,L., Monteagudo-Sánchez,A., Perez-Tur,J. *et al.* (2018) PM20D1 is a quantitative trait locus associated with alzheimer's disease. *Nat. Med.*, **24**, 598–603.
35. Young,J.I., Sivasankaran,S.K., Wang,L., Ali,A., Mehta,A., Davis,D.A., Dykxhoorn,D.M., Petito,C.K., Beecham,G.W., Martin,E.R. *et al.* (2019) Genome-wide brain DNA methylation analysis suggests epigenetic reprogramming in Parkinson disease. *Neurology: Genetics*, **5**, e342.
36. Candler,T., Kessler,N., Gunasekara,C., Ward,K., James,P., Laritsky,E., Baker,M., Dyer,R., Elango,R. and Jeffries,D. (2021) DNA methylation at a nutritionally sensitive region of the PAX8 gene is associated with thyroid volume and function in gambian children. *Sci. Adv.*, **7**, eabj1561.
37. Gunasekara,C.J. and Waterland,R.A. (2019) A new era for epigenetic epidemiology. *Epigenomics*, **11**, 1647–1649.
38. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
39. Silva,T.C., Colaprico,A., Olsen,C., D'Angelo,F., Bontempi,G., Ceccarelli,M. and Noushmehr,H. (2016) TCGA workflow: analyze cancer genomics and epigenomics data using bioconductor packages. *F1000Res*, **5**, 1542.
40. Mounir,M., Lucchetta,M., Silva,T.C., Olsen,C., Bontempi,G., Chen,X., Noushmehr,H., Colaprico,A. and Papaleo,E. (2019) New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput. Biol.*, **15**, e1006701.
41. Colaprico,A., Silva,T.C., Olsen,C., Garofano,L., Cava,C., Garolini,D., Sabedot,T.S., Malta,T.M., Pagnotta,S.M., Castiglioni,I. *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.
42. Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the gene expression omnibus (GEO) and bioconductor. *Bioinformatics*, **23**, 1846–1847.
43. Moran,S., Arribas,C. and Esteller,M. (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, **8**, 389–399.
44. Chandak,G.R., Silver,M.J., Saffari,A., Lillycrop,K.A., Shrestha,S., Sahariah,S.A., di Gravio,C., Goldberg,G., Tomar,A.S., Betts,M. *et al.* (2017) Protocol for the EMPHASIS study; epigenetic mechanisms linking maternal pre-conceptional nutrition and children's health in india and sub-saharan africa. *BMC Nutrition*, **3**, 81.
45. Brucato,N., Fernandes,V., Mazières,S., Kusuma,P., Cox,M.P., Ng'ang'a,J.W., Omar,M., Simeone-Senelle,M.-C., Frassati,C., Alshamali,F. *et al.* (2018) The comoros show the earliest austronesian gene flow into the swahili corridor. *Am. J. Hum. Genet.*, **102**, 58–68.
46. 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.
47. Nordlund,J., Bäcklin,C.L., Wahlberg,P., Busche,S., Berglund,E.C., Eloranta,M.-L., Flaegstad,T., Forestier,E., Frost,B.-M., Harila-Saari,A. *et al.* (2013) Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol.*, **14**, r105.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

48. Zhou,W., Laird,P.W. and Shen,H. (2017) Comprehensive characterization, annotation and innovative use of infinium DNA methylation beadchip probes. *Nucleic Acids Res.*, **45**, e22.
49. Teschendorff,A.E., Marabita,F., Lechner,M., Bartlett,T., Tegner,J., Gomez-Cabrero,D. and Beck,S. (2013) A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k DNA methylation data. *Bioinformatics*, **29**, 189–196.
50. Bose,M., Wu,C., Pankow,J.S., Demerath,E.W., Bressler,J., Fornage,M., Grove,M.L., Mosley,T.H., Hicks,C., North,K. *et al.* (2014) Evaluation of microarray-based DNA methylation measurement using technical replicates: the atherosclerosis risk in communities (ARIC) study. *BMC Bioinf.*, **15**, 312.
51. Min,J.L., Hemani,G., Hannon,E., Dekkers,K.F., Castillo-Fernandez,J., Luijk,R., Carnero-Montoro,E., Lawson,D.J., Burrows,K. and Suderman,M. (2021) Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat. Genet.*, **53**, 1311–1321.
52. Bonilla,C., Bertoni,B., Min,J.L., Hemani,G. and Genetics of DNA Methylation ConsortiumGenetics of DNA Methylation Consortium and Elliott,H.R. (2021) Investigating DNA methylation as a potential mediator between pigmentation genes, pigmentary traits and skin cancer. *Pigment Cell Melanoma Res.*, **34**, 892–904.
53. Planterose Jiménez,B., Liu,F., Caliebe,A., Montiel González,D., Bell,J.T., Kayser,M. and Vidaki,A. (2021) Equivalent DNA methylation variation between monozygotic co-twins and unrelated individuals reveals universal epigenetic inter-individual dissimilarity. *Genome Biol.*, **22**, 18.
54. Flanagan,J.M., Brook,M.N., Orr,N., Tomczyk,K., Coulson,P., Fletcher,O., Jones,M.E., Schoemaker,M.J., Ashworth,A., Swerdlow,A. *et al.* (2015) Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study. *Cancer Epidemiol. Biomarkers Prev.*, **24**, 221–229.
55. Grundberg,E., Meduri,E., Sandling,J.K., Hedman,A.K., Keildson,S., Buil,A., Busche,S., Yuan,W., Nisbet,J., Sekowska,M. *et al.* (2013) Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.*, **93**, 876–890.
56. van Dongen,J., Gordon,S.D., McRae,A.F., Odintsova,V.v., Mbarek,H., Breeze,C.E., Sugden,K., Lundgren,S., Castillo-Fernandez,J.E., Hannon,E. *et al.* (2021) Identical twins carry a persistent epigenetic signature of early genome programming. *Nat. Commun.*, **12**, 5618.
57. Silver,M.J., Saffari,A., Kessler,N.J., Chandak,G.R., Fall,C.H.D., Issarapu,P., Dedaniya,A., Betts,M., Moore,S.E., Routledge,M.N. *et al.* (2022) Environmentally sensitive hotspots in the methylome of the early human embryo. *Elife*, **11**, e72031.
58. Zink,F., Magnusdottir,D.N., Magnusson,O.T., Walker,N.J., Morris,T.J., Sigurdsson,A., Halldorsson,G.H., Gudjonsson,S.A., Melsted,P., Ingimundardottir,H. *et al.* (2018) Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nat. Genet.*, **50**, 1542–1552.
59. Lokk,K., Modhukur,V., Rajashekar,B., Märtens,K., Mägi,R., Kolde,R., Koltšina,M., Nilsson,T.K., Vilo,J., Salumets,A. *et al.* (2014) DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol.*, **15**, r54.
60. Min,J.L., Hemani,G., Smith,G.D., Relton,C. and Suderman,M. (2018) Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics*, **34**, 3983.
61. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
62. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
63. Phipson,B., Maksimovic,J. and Oshlack,A. (2016) MissMethyl: an r package for analyzing data from illumina's humanmethylation450 platform. *Bioinformatics*, **32**, 286–288.
64. Fraga,M.F., Ballestar,E., Paz,M.F., Ropero,S., Setien,F., Ballestar,M.L., Heine-Suñer,D., Cigudosa,J.C., Urioste,M. and Benitez,J. (2005) Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 10604–10609.
65. Hannon,E., Lunnon,K., Schalkwyk,L. and Mill,J. (2015) Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics*, **10**, 1024–1032.
66. Islam,S.A., Goodman,S.J., MacIsaac,J.L., Obradović,J., Barr,R.G., Boyce,W.T. and Kobor,M.S. (2019) Integration of DNA methylation patterns and genetic variation in human pediatric tissues help inform EWAS design and interpretation. *Epigenetics Chromatin*, **12**, 1.
67. Rakyan,V.K., Blewitt,M.E., Druker,R., Preis,J.I. and Whitelaw,E. (2002) Metastable epialleles in mammals. *Trends Genet.*, **18**, 348–351.
68. Kazachenka,A., Bertozzi,T.M., Sjoberg-Herrera,M.K., Walker,N., Gardner,J., Gunning,R., Pahita,E., Adams,S., Adams,D. and Ferguson-Smith,A.C. (2018) Identification, characterization, and heritability of murine metastable epialleles: implications for Non-genetic inheritance. *Cell*, **175**, 1259–1271.
69. Bertozzi,T.M., Elmer,J.L., Macfarlan,T.S. and Ferguson-Smith,A.C. (2020) KRAB zinc finger protein diversification drives mammalian interindividual methylation variability. *Proc. Natl. Acad. Sci.*, **117**, 31290–31300.
70. Elmer,J.L., Hay,A.D., Kessler,N.J., Bertozzi,T.M., Ainscough,E.A.C. and Ferguson-Smith,A.C. (2021) Genomic properties of variably methylated retrotransposons in mouse. *Mob DNA*, **12**, 6.
71. Monk,D., Mackay,D.J.G., Eggermann,T., Maher,E.R. and Riccio,A. (2019) Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nat. Rev. Genet.*, **20**, 235–248.
72. Carpenter,B.L., Remba,T.K., Thomas,S.L., Madaj,Z., Brink,L., Tiedemann,R.L., Odendaal,H.J. and Jones,P.A. (2021) Oocyte age and preconceptual alcohol use are highly correlated with epigenetic imprinting of a noncoding RNA (nc886). *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2026580118.
73. Markunas,C.A., Wilcox,A.J., Xu,Z., Joubert,B.R., Harlid,S., Panduri,V., Håberg,S.E., Nystad,W., London,S.J., Sandler,D.P. *et al.* (2016) Maternal age at delivery is associated with an epigenetic signature in both newborns and adults. *PLoS One*, **11**, e0156361.
74. Gluckman,P.D., Hanson,M.A. and Low,F.M. (2011) The role of developmental plasticity and epigenetics in human health. *Birth Defects Res. C - Embryo Today: Reviews*, **93**, 12–18.
75. Low,F.M., Gluckman,P.D. and Hanson,M.A. (2012) Developmental plasticity, epigenetics and human health. *Evol. Biol.*, **39**, 650–665.
76. Fleming,T.P., Watkins,A.J., Velazquez,M.A., Mathers,J.C., Prentice,A.M., Stephenson,J., Barker,M., Saffery,R., Yajnik,C.S., Eckert,J.J. *et al.* (2018) Origins of lifetime health around the time of conception: causes and consequences. *Lancet North Am. Ed.*, **391**, 1842–1852.
77. Moore,S.E., Cole,T.J., Collinson,A.C., Poskitt,E.M.E., McGregor,I.A. and Prentice,A.M. (1999) Prenatal or early postnatal events predict infectious deaths in young adulthood in rural africa. *Int. J. Epidemiol.*, **28**, 1088–1095.
78. Dominguez-Salas,P., Moore,S.E., Cole,D., da Costa,K.A., Cox,S.E., Dyer,R.A., Fulford,A.J.C., Innis,S.M., Waterland,R.A., Zeisel,S.H. *et al.* (2013) DNA methylation potential: dietary intake and blood concentrations of one-carbon metabolites and cofactors in rural african women. *Am. J. Clin. Nutr.*, **97**, 1217–1227.
79. James,P.T., Dominguez-Salas,P., Hennig,B.J., Moore,S.E., Prentice,A.M. and Silver,M.J. (2019) Maternal one-carbon metabolism and infant DNA methylation between contrasting seasonal environments: a case study from the Gambia. *Curr. Dev. Nutr.*, **3**, nzy082.
80. el Hajj,N., Dittrich,M. and Haaf,T. (2017) Epigenetic dysregulation of protocadherins in human disease. *Semin. Cell Dev. Biol.*, **69**, 172–182.
81. Toyoda,S., Kawaguchi,M., Kobayashi,T., Tarusawa,E., Toyama,T., Okano,M., Oda,M., Nakauchi,H., Yoshimura,Y., Sanbo,M. *et al.* (2014) Developmental epigenetic modification regulates stochastic expression of clustered protocadherin genes, generating single neuron diversity. *Neuron*, **82**, 94–108.
82. Flaherty,E. and Maniatis,T. (2020) The role of clustered protocadherins in neurodevelopment and neuropsychiatric diseases. *Curr. Opin. Genet. Dev.*, **65**, 144–150.
83. Salpea,P., Russanova,V.R., Hirai,T.H., Sourlingas,T.G., Sekeri-Pataryas,K.E., Romero,R., Epstein,J. and Howard,B.H. (2012) Postnatal development- and age-related changes in

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

DNA-methylation patterns in the human genome. *Nucleic Acids Res.*, **40**, 6477–6494.
84. McClay,J.L., Aberg,K.A., Clark,S.L., Nerella,S., Kumar,G., Xie,L.Y., Hudson,A.D., Harada,A., Hultman,C.M., Magnusson,P.K.E. *et al.* (2014) A methylome-wide study of aging using massively parallel sequencing of the methyl-CpG-enriched genomic fraction from blood in over 700 subjects. *Hum. Mol. Genet.*, **23**, 1175–1185.
85. Kim,S., Wyckoff,J., Morris,A.-T., Succop,A., Avery,A., Duncan,G.E. and Jazwinski,S.M. (2018) DNA methylation associated with healthy aging of elderly twins. *GeroScience*, **40**, 469–484.
86. van Dongen,J., Nivard,M.G., Willemsen,G., Hottenga,J.J., Helmer,Q., Dolan,C., Ehli,E.A., Davies,G.E., van Iterson,M., Breeze,C.E. *et al.* (2016) Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat. Commun.*, **7**, 11115.
87. Kaut,O., Schmitt,I. and Wüllner,U. (2012) Genome-scale methylation analysis of parkinson's disease patients' brains reveals DNA hypomethylation and increased mRNA expression of cytochrome P450 2E1. *Neurogenetics*, **13**, 87–91.
88. Amarasekera,M., Martino,D., Ashley,S., Harb,H., Kesper,D., Strickland,D., Saffery,R. and Prescott,S.L. (2014) Genome-wide DNA methylation profiling identifies a folate-sensitive region of differential methylation upstream of ZFP57-imprinting regulator in humans. *FASEB J.*, **28**, 4068–4076.
89. Byun,H.M., Siegmund,K.D., Pan,F., Weisenberger,D.J., Kanel,G., Laird,P.W. and Yang,A.S. (2009) Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum. Mol. Genet.*, **18**, 4808–4817.
90. Slieker,R.C., Bos,S.D., Goeman,J.J., Bovée,J.V., Talens,R.P., van der Breggen,R., Suchiman,H.E.D., Lameijer,E.W., Putter,H., van den Akker,E.B. *et al.* (2013) Identification and systematic annotation of tissue-specific differentially methylated regions using the illumina 450k array. *Epigenetics Chromatin*, **6**, 26.
91. Davies,M.N., Volta,M., Pidsley,R., Lunnon,K., Dixit,A., Lovestone,S., Coarfa,C., Harris,R.A., Milosavljevic,A., Troakes,C. *et al.* (2012) Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol.*, **13**, R43–R43.
92. Walton,E., Hass,J., Liu,J., Roffman,J.L., Bernardoni,F., Roessner,V., Kirsch,M., Schackert,G., Calhoun,V. and Ehrlich,S. (2016) Correspondence of DNA methylation between blood and brain tissue and its application to schizophrenia research. *Schizophr. Bull.*, **42**, 406–414.
93. Braun,P.R., Han,S., Hing,B., Nagahama,Y., Gaul,L.N., Heinzman,J.T., Grossbach,A.J., Close,L., Dlouhy,B.J., Howard,M.A. *et al.* (2019) Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. *Transl. Psychiatry*, **9**, 47.
94. Hannon,E., Mansell,G., Walker,E., Nabais,M.F., Burrage,J., Kepa,A., Best-Lane,J., Rose,A., Heck,S., Moffitt,T.E. *et al.* (2021) Assessing the co-variability of DNA methylation across peripheral cells and tissues: implications for the interpretation of findings in epigenetic epidemiology. *PLoS Genet.*, **17**, e1009443.
95. Jiang,R., Jones,M.J., Chen,E., Neumann,S.M., Fraser,H.B., Miller,G.E. and Kobor,M.S. (2015) Discordance of DNA methylation variance between two accessible human tissues. *Sci. Rep.*, **5**, 8257.
96. 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 Genet.*, **14**, e1007707.
97. Gordon,L., Joo,J.E., Powell,J.E., Ollikainen,M., Novakovic,B., Li,X., Andronikos,R., Cruickshank,M.N., Conneely,K.N., Smith,A.K. *et al.* (2012) Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome Res.*, **22**, 1395–1406.
98. Do,C., Lang,C.F., Lin,J., Darbary,H., Krupska,I., Gaba,A., Petukhova,L., Vonsattel,J.P., Gallagher,M.P., Goland,R.S. *et al.* (2016) Mechanisms and disease associations of haplotype-dependent allele-specific DNA methylation. *Am. J. Hum. Genet.*, **98**, 934–955.
99. Vadgama,N., Pittman,A., Simpson,M., Nirmalananthan,N., Murray,R., Yoshikawa,T., de Rijk,P., Rees,E., Kirov,G. and Hughes,D. (2019) De novo single-nucleotide and copy number variation in discordant monozygotic twins reveals disease-related genes. *Eur. J. Hum. Genet.*, **27**, 1121–1133.
100. Richmond,R.C., Sharp,G.C., Herbert,G., Atkinson,C., Taylor,C., Bhattacharya,S., Campbell,D., Hall,M., Kazmi,N., Gaunt,T. *et al.* (2018) The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the aberdeen folic acid supplementation trial (AFAST). *Int. J. Epidemiol.*, **47**, 928–937.
101. Ambatipudi,S., Cuenin,C., Hernandez-Vargas,H., Ghantous,A., le Calvez-Kelm,F., Kaaks,R., Barrdahl,M., Boeing,H., Aleksandrova,K. and Trichopoulou,A. (2016) Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics*, **8**, 599–618.
102. James,P., Sajjadi,S., Tomar,A.S., Saffari,A., Fall,C.H.D., Prentice,A.M., Shrestha,S., Issarapu,P., Yadav,D.K., Kaur,L. *et al.* (2018) Candidate genes linking maternal nutrient exposure to offspring health via DNA methylation: a review of existing evidence in humans with specific focus on one-carbon metabolism. *Int. J. Epidemiol.*, **47**, 1910–1937.
103. Wolff,G.L., Kodell,R.L., Moore,S.R. and Cooney,C.A. (1998) Maternal epigenetics and methyl supplements affect agouti gene expression in a vy /a mice.
104. Cooney,C.A., Dave,A.A. and Wolff,G.L. (2002) Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. *J. Nutr.*, **132**, 2393S–2400S.
105. Dolinoy,D.C., Huang,D. and Jirtle,R.L. (2007) Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. *Proc. Natl. Acad. Sci.*, **104**, 13056–13061.
106. 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**, e71104.
107. Imbeault,M., Helleboid,P.Y. and Trono,D. (2017) KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, **543**, 550–554.
108. Shen,P., Xu,A., Hou,Y., Wang,H., Gao,C., He,F. and Yang,D. (2021) Conserved paradoxical relationships among the evolutionary, structural and expressional features of KRAB zinc-finger proteins reveal their special functional characteristics. *BMC Mol. Cell Biol.*, **22**, 7.
109. Zhu,Y., Mordaunt,C.E., Yasui,D.H., Marathe,R., Coulson,R.L., Dunaway,K.W., Jianu,J.M., Walker,C.K., Ozonoff,S., Hertz-Picciotto,I. *et al.* (2019) Placental DNA methylation levels at CYP2E1 and IRS2 are associated with child outcome in a prospective autism study. *Hum. Mol. Genet.*, **28**, 2659–2674.
110. Feinberg,A.P. and Irizarry,R.A. (2010) Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl. Acad. Sci. U.S.A.*, 107, 1757–1764.
111. Ballouz,S., Pena,M.T., Knight,F.M., Adams,L.B. and Gillis,J.A. (2019) The transcriptional legacy of developmental stochasticity. bioRxiv doi: https://doi.org/10.1101/2019.12.11.873265, 12 December 2019, preprint: not peer reviewed.
112. Liu,Y., Aryee,M.J., Padyukov,L., Fallin,M.D., Hesselberg,E., Runarsson,A., Reinius,L., Acevedo,N., Taub,M., Ronninger,M. *et al.* (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.*, **31**, 142–147.
113. Matzaraki,V., Kumar,V., Wijmenga,C. and Zhernakova,A. (2017) The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.*, **18**, 76.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac503/6609816 by guest on 21 June 2022

Supplementary Analysis

Our working hypothesis is that methylation establishment at hvCpGs occurs before gastrulation, which is supported by findings in Chapter 3 that many hvCpGs show systemic interindividual variation (SIV) across somatic tissues. To gain further insight into inter-tissue correlations at hvCpGs, we examined whether methylation states at hvCpGs covary between blood and sperm. Sperm DNAm patterns are established after the global erasure of DNAm in primordial germ cells and are highly distinctive relative to somatic tissues[114]. We leveraged blood-sperm correlation data generated by Åsenius et al. using a cross-sectional IlluminaEPIC dataset of matched human blood and sperm samples[114]. These data were available for 837 hvCpGs that had met the minimum variability threshold in both blood and sperm set in their study (see Åsenius et al. for further details). 34 (4%) of these 837 hvCpGs showed significant correlations between the two tissues (applying the same p value threshold as Åsenius et al. of $p < 9 \times 10^{-8}$), with a median Pearson r of 0.83 (IQR = [0.77, 0.89]). Genetic variation has been argued to be the underlying cause of methylation covariation between blood and sperm[114], suggesting that methylation variability at some hvCpGs may be attributed to genetic effects rather than stochastic/and or environmental effects in the early embryo[114].

114. Åsenius, F., Gorrie-Stone, T.J., Brew, A., Panchbhaya, Y., Williamson, E., Schalkwyk, L.C., Rakyan, V.K., Holland, M.L., Marzi, S.J. and Williams, D.J., 2020. The DNA methylome of human sperm is distinct from blood with little evidence for tissue-consistent obesity associations. *PLoS genetics*, *16*(10), p.e1009035.

Chapter 4

4. KZFP-mediated silencing of transposable elements as a mechanism influencing hypervariable methylation in early development

Rationale of chapter

In Chapter 3, we reported that ~17% of hvCpGs are associated with at least one *trans* mQTL. In a subsequent analysis examining the genomic locations of these mQTLs (included in Chapter 4), we noted their enrichment amongst genes encoding Krüppel associated Box (KRAB) zinc finger proteins (KZFPs). Whilst KZFPs usually epigenetically silence transposable elements (TEs) in the early embryo by recruiting KAP1 (KRAB-associated protein), recent evidence suggests that the establishment and maintenance methylation variability at murine MEs may be attributed to variable binding of KZFPs to TEs in a sequence-dependent manner. In this chapter, we were therefore motivated to explore whether KZFP expression/and or activity in early development might also influence methylation at hvCpGs. We did this by examining whether the mQTLs located in KZFP genes might also influence KZFP expression, and by examining the proximity of hvCpGs to KZFP and KAP1 binding sites.

Summary of findings

Hypervariable CpGs ('hvCpGs') are enriched for *trans* mQTLs located in KZFP genes. Many of these mQTLs overlap KZFP eQTLs, suggesting that methylation at hvCpGs may be influenced by KZFP expression. hvCpGs are enriched for proximal KAP1 binding sites in embryonic stem cells (ESCs), further suggesting a link between hvCpG methylation and KZFP-KAP1 epigenetic silencing machinery in the early embryo. hvCpGs are also enriched for proximity to specific retrotransposon subfamilies, and for KAP1-bound retrotransposons, indicating that methylation at hvCpGs may be linked to KZFP-mediated silencing of TEs. hvCpGs are additionally notably enriched for proximal binding sites of ZFP57 and ZNF445: KZFPs that maintain parent-of-origin specific methylation (PofOm) in the early embryo. Together, these findings: i) strengthen evidence given in Chapter 3 that DNAm states at many hvCpGs are linked to the early embryo, and ii) indicate that KZFP-KAP1 epigenetic silencing mechanisms at retrotransposons and regions of PofOm may play a role in methylation variability at hvCpGs in early development.

Notes

The Supplementary Figures and Tables for this chapter can be found in Appendix B.

4.1 Background

4.1.1. KRAB zinc finger proteins and transposable elements

KRAB zinc finger proteins (KZFPs) are a diverse family of vertebrate-specific transcription factors that emerged 420 million years ago in tetrapods[1–3]. KZFPs comprise a C terminal multi-C2H2-zinc finger domain that binds to DNA in a sequence-specific manner[4,5], and an N terminal KRAB domain[6] that recruits the co-factor KAP1 (KRAB-associated protein 1)[7–9]. KAP1 acts as a scaffold for an epigenetic silencing complex comprising the histone methyltransferase SETDB1[10], heterochromatin protein 1 (HP1)[11], and the nucleosome remodelling and deacetylation (NuRD) complex[12]. KAP1 also recruits *de novo* DNA methyltransferases (DNMTs) in the early embryo[13,14]. Together, these proteins deposit DNA methylation and heterochromatin inducing H3K9me3 marks, resulting in epigenetic silencing of the target region (**Figure 4.1**).

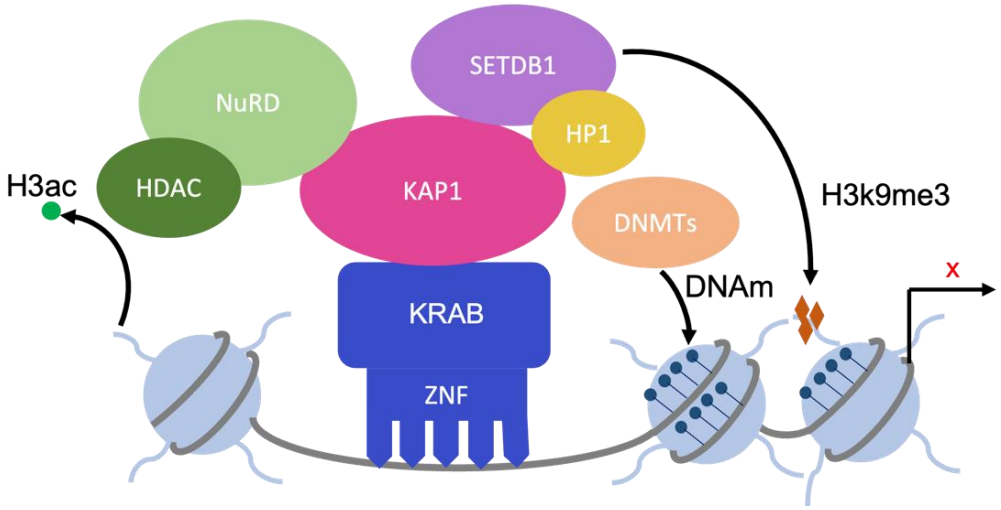


Figure 4.1. KZFP-mediated epigenetic silencing via the induction of heterochromatin formation and DNA methylation. KRAB zinc finger proteins contain an N-terminal KRAB domain and a C-terminal array of DNA-binding zinc fingers ('ZNF'). The KRAB domain recruits KAP1 which acts as a scaffold for the histone methyltransferase SETDB1 and NuRD/HDAC which catalyse the addition of H3K9me3 and removal of H3K9ac respectively. HP1 recognises H3K9me3 which further acts to silence the region. DNMTs recruited by KAP1 deposit methylation at CpG sites. KRAB = Krüppel associated Box, ZNF = zinc finger array, KAP1 = KRAB-associated protein, DNMTs = de novo methyltransferases, SETDB1 = SET domain bifurcated histone lysine methyltransferase 1, HP1 = heterochromatin protein 1, NuRD = nucleosome remodelling and deacetylation, HDAC = histone deacetylase.

KZFPs target a range of genomic features, most commonly transposable elements (TEs)[3,15]. TEs are parasitic elements that comprise approximately half of mammalian genomes[16,17] and, when activated, can harm the host though several mechanisms, including genome damage via transposition[18–20], activation of aberrant expression of nearby genes[21–23] and induction of the innate immune response[24]. TEs are therefore epigenetically silenced by KZFPs and RNA-based mechanisms[25] in embryonic stem cells (ESCs), to prevent their activation during the global wave of demethylation that occurs shortly after fertilisation[26–31]. DNAm states established by KZFPs in ESCs are stably inherited through mitotic cell division[26,30,32]. KZFPs can also recognise their TE targets in somatic cells although, unlike in ESCs, KZFPs epigenetically regulate TEs in differentiated cells using histone modifications alone rather than DNA methylation[26,33,34].

KZFPs and their TE targets co-evolve in an evolutionary arms race[35–38], whereby genetic drift of TE sequences allows them to escape KZFP recognition and in turn provides a selective pressure on the diversification of KZFP zinc-finger domains. Supporting this, waves of KZFP expansion are correlated with waves of TE expansion, with many KZFPs and their TE targets having similar evolutionary ages[2,3,36]. More recent evidence indicates that context-specific epigenetic regulation of TEs throughout development by KZFPs can also aid the co-option of advantageous TE-based regulatory features such as promoters, enhancers, insulators, and transcription factor (TF) binding sites[39–41] by the host genome. In this manner, KZFPs and their TE targets are postulated to contribute to the wiring of gene regulatory networks[3,42–44].

4.1.2. The link between KZPFs, TEs and metastable epialleles

Murine metastable epialleles (MEs) have been identified at Intracisternal A Particle elements (`IAPs`) in the mouse reference genome (C57BL/6J). IAPs are an evolutionarily young class of murine-specific Long-Terminal Repeat (LTR) retrotransposon[45,46]. Recent evidence has linked methylation variation at murine MEs, also referred to as Variably Methylated IAPs ('VM-IAPs'), to incomplete epigenetic silencing by a cluster of *trans*-acting KZFPs on chromosome 4. Deletion of this cluster causes hypomethylation of VM-IAPs[47]. VM-IAPs are enriched for specific sequences[45,47,48], flanking CTCF binding sites[45,46,48] and have high CpG density[48], suggesting that the genomic context of IAPs influences their susceptibility to being variably methylated by KZFPs[45,48]. A recent study has proposed that low methylation states at VM-IAPs

can be protected from further silencing via the recruitment of ZF-CxxC containing proteins that recognise CpG dense loci[48].

If this model holds true, variable binding of KZFPs to IAP elements in early development will result in cell variegation: a given IAP that is bound by KZFPs in one cell will be methylated, whilst the same IAP element that is not bound by KZFPs in another cell will be hypomethylated (**Figure 4.2**). These methylation states will then be inherited by cell progeny throughout tissue differentiation, resulting in epigenetic mosaicism in the adult mice (**Figure 4.2**). Given that this methylation establishment is stochastic, the proportion of methylated cells will vary between individuals, resulting in interindividual methylation variation (**Figure 4.2**).

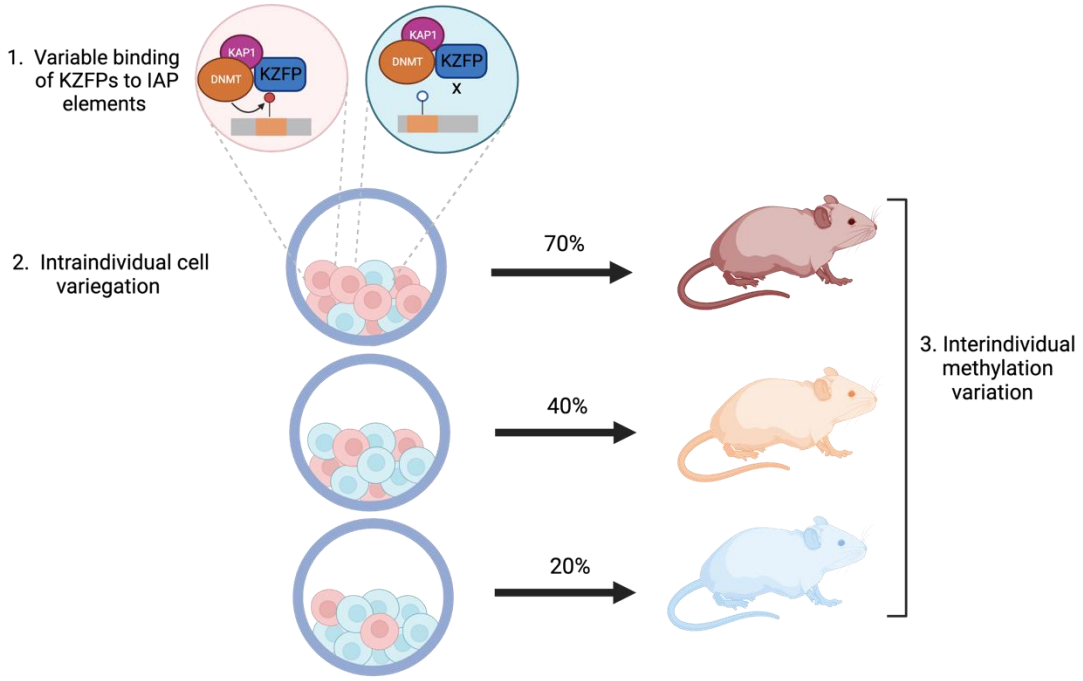


Figure 4.2. Model hypothesising that variable binding of KZFPs to IAP elements results in stochastic setting of methylation states at the cellular level, driving interindividual methylation variation at metastable epialleles. In some cells, the KZFP binds with high affinity, resulting in hypermethylation of the target (red cells) due to the recruitment of DNMTs by KAP1. In others, the KZFP binds with low affinity, resulting in hypomethylation of the target (blue cells). The aggregate methylation levels in a given sample reflect the proportion of cells that were methylated at this time point. This proportion will vary between individuals, resulting in interindividual methylation variation between isogenic mice. Image created with BioRender.com.

Whilst the human genome lacks LTR-containing IAP elements, putative human MEs and hvCpGs have been previously associated with LTR-containing retrotransposons of the classes ERV1 and ERVK[49–51], raising the possibility that KZFP-mediated silencing mechanisms may have a role in driving methylation variability in the early human embryo also. In addition, KZFPs have been found to be trans-acting transcriptional regulators that can influence DNA methylation in humans[52,53].

4.1.3. Aims of chapter

We reasoned that the association between hvCpGs and KZFP mQTLs does not contradict our hypothesis that methylation variation at hvCpGs is only partly attributed to genetic variation and may instead suggest a link between KZFP expression and/or activity and hvCpG methylation. In this chapter, we explore this by: i) examining whether the mQTLs located in KZFP genes are also eQTLs for KZFPs, and ii) examining the proximity of hvCpGs to KAP1 and KZFP binding sites to investigate whether hvCpGs are proximal to regions that are targeted by KZFP-KAP1 machinery. We also begin the chapter by examining the proximity of hvCpGs to additional classes of retrotransposon beyond ERV1 and ERVK.

We find evidence that hvCpG methylation may be influenced by KZFP expression/activity and that hvCpGs colocalise with regions that are epigenetically silenced by KZFPs in the pre-implantation embryo. These findings suggest that methylation variation at some hvCpGs may be linked to KZFP-mediated silencing of TEs and regions of parent-of-origin specific methylation in early development.

4.2 Methods

4.2.1 CpG sets

This chapter uses several of the same CpG sets used in Chapter 3. We describe them briefly here but please refer to the Methods section of Chapter 3 for further details.

hvCpGs

4,143 Illumina450K CpGs that have top 5% methylation variance in many tissues and ethnicities, and which show several features linking them to the early embryo (see Chapter 3 and Derakhshan *et al.*[54]).

De-clustered hvCpGs

To account for the possibility that some analyses may be biased by the non-random distribution and inter-dependence of hvCpGs, we used a set of 2,640 hvCpGs in which no CpG was within 4kb of any other CpG.

Array background CpGs

406,306 Illumina450K array CpGs covered in at least 15 of the 30 datasets we used in Chapter 3.

4.2.2 Retrotransposon nomenclature and genomic coordinates

The retrotransposon classification system is shown in **Figure B.1**. Briefly, retrotransposons can be split into Long Terminal Repeat (LTR)-containing retrotransposons and non-LTR retrotransposons. LTR-containing retrotransposons comprise endogenous retroviruses (ERVs) and non-LTR retrotransposons comprise long interspersed repeats (LINEs) and short interspersed repeats (SINEs). Each of these ERV, LINE and SINE classes can be further split into families and subfamilies based on sequence similarity (**Figure B.1**). We downloaded the hg19 genomic coordinates of 698 subfamilies (belonging to 16 families) from the UCSC RepeatMasker database using the 'Table Browser' tool (**Figure B.2**). We refer to each instance of a retrotransposon in the genome as an 'element'.

4.2.3 Retrotransposon evolutionary age

We obtained the evolutionary clade of each retrotransposon subfamily from the Dfam database[55]. These clades are assigned by Dfam according to NCBI taxonomy nomenclature, with each clade containing the descendants in which the retrotransposon was present. For example, retrotransposon subfamilies assigned to the clade 'Mammalia' will be found in all mammals[56]. Dfam subfamily names differ slightly to those in RepeatMasker because some elements are split into portions in the former, such as 3'-end and 5'end LTRs[57]. We matched RepeatMasker ids to Dfam ids and extracted clade information using the /famdb.py extension: *famdb.py -i ./Dfam.h5 family -f embl_meta $subfamily*

4.2.4 Retrotransposon CpG density

Retrotransposon coordinate ranges were obtained from RepeatMasker and corresponding sequences were extracted from the hg19 reference genome fasta file. The CpG density of each retrotransposon element was calculated as the number of GC or GC dinucleotides divided by the length of the retrotransposon element. For each subfamily, we compared CpG densities of retrotransposons within 3kb of an hvCpG to those within 3kb of an array background CpG (excluding hvCpGs).

4.2.5 KRAB-zinc finger protein (KZFP) and KAP1 binding sites

To investigate if hvCpGs are proximal to regions targeted by KZFP-KAP1 silencing machinery, we downloaded KZFP and KAP1 binding site data (hg19 coordinates) from the Gene Expression Omnibus (GEO) (GSE78099) using the *GEOquery (v2.58.0)* R package[58]. These binding sites had been generated by chromatin immunoprecipitation with exonuclease digestion (ChIP-exo) by Imbeault *et al.*[3]. ChIP-exo uses lambda exonuclease to digest double-stranded DNA, resulting in the resolution of transcription factor (TF) binding sites to a near single nucleotide accuracy[59]. ChIP-exo data for KZFPs were generated after inducing the expression of KZFPs in human embryonic kidney 293T (HEK293T) cells lines[3]. In total, 236 ChIP-exo experiments were downloaded to analyse binding site data for 221 unique KZFPs, as several experiments were repeated. We used this dataset as it is the only publicly available KZFP ChIP-exo dataset. ChIP-exo data for endogenous KAP1 binding sites were obtained from H1

embryonic stem cells (H1-ESCs), derived from the inner cell mass of the pre-implantation embryo[60]. Further details on ChIP-exo quality control steps are described elsewhere[3].

4.2.6 Regions of parent-of-origin specific methylation (PofOm)

229 regions of parent-of-origin differentially methylated regions ('PofOm DMRs') were identified in whole blood from Icelandic individuals in a genome-wide screen by Zink *et al.*[61]. Many of these PofOm DMRs had not been identified in previous studies. In this chapter, we examined the overlap between this updated set of PofOm DMRs and KZFP ChIP-exo peaks.

4.2.7 hvCpG proximity to features of interest

We used two-sided Fishers Exact Tests (FET) to compare the enrichment/depletion of hvCpGs for proximal features of interest (KAP1 binding sites, KZFP binding sites, retrotransposon subfamily elements), relative to array background CpGs. We used a de-clustered set of hvCpGs (see 'CpG sets') to reduce potential biases caused by the clustering of hvCpGs. We examined the proximity of hvCpGs to each of the 698 retrotransposon subfamilies, and the 231 KZFP ChIP-exo experiments, and considered FDR < 0.05 significant. We set our proximity threshold at 3kb, because DNAm has been found to 'spread' 3-5kb from the KAP1 docking site in ESCs[26]. We also repeated the KZFP ChIP-exo analysis for additional proximity thresholds set at 2kb, 1kb and 0kb (i.e overlapping). Note, we did not consider these additional thresholds for the retrotransposon analysis because retrotransposons are poorly covered on the Illumina450K array.

4.2.8 Histone modification analysis

H3K modifications in H1-ESCs were examined using the *annotatr (v1.10.0)* R package, using ChIP-sequencing ('ChIP-seq') data generated by the Roadmap Epigenomic Consortium ('E003'). We used 'broadPeak' data because H3K9me3 histone modifications are diffuse[62]. 3,750 hvCpGs and 318,699 array background CpGs were annotated to these 'broadPeaks' respectively. We tested their enrichment for each type of H3 modification at hvCpGs relative to array background CpGs using two-sided Fishers Exact Tests.

4.2.9 Association between hvCpG methylation and genetic variation at KZFP genes

To examine whether hvCpG methylation is influenced by the expression and/or function of KZFPs, we examined the association between methylation at hvCpGs and genetic variation at KZFP genes located *in trans*. This analysis used publicly available summary data from a large mQTL meta-analysis conducted in Europeans[63] (*trans* mQTLs defined as $p < 1 \times 10^{-14}$ by Min *et al.*[63]). We first identified *trans* mQTLs associated with hvCpGs and array background CpGs respectively and then used a one-sided Fishers Exact Test to examine whether mQTLs associated with hvCpGs were enriched amongst KZFP genes relative to mQTLs associated with the array background CpGs. Gene names for 366 KZFP genes were obtained from two previous studies that had identified KZFP genes by screening for genes proximal to both C2H2 zinc finger arrays and KRAB domains (see Shen *et al.*[64] and Imbeault *et al.*[3] for further details) (**Figure B.3**). We obtained the hg19 coordinates for these genes and their promoters (defined as the region spanning 1000 bp upstream of the transcription start site) using the *TxDb.Hsapiens.UCSC.hg19.knownGene (v3.2.2)* R package. In this chapter, we refer to *trans* mQTLs located in KZFP genes as 'KZFP mQTLs'.

4.2.10 Characterising KZFP mQTLs

To functionally characterise KZFP mQTLs, we mapped their genomic locations to gene transcripts and protein coding sequences of KZFP genes using the *EnsDb.Hsapienes.v75 (v2.99.0)* R package[65] and *gnm_tx_output()* and *gnm_prt_output()* functions from the *Gviz (v 1.43.0)* R package[66]. To explore evidence that these mQTLs may influence KZFP expression we examined their overlap with eQTLs using the HaploReg database[67] .

4.3 Results

4.3.1 hvCpGs are enriched for proximity to certain retrotransposon subfamilies

In Chapter 3, we found that hvCpGs are enriched for proximity to ERV1 and ERVK retrotransposon families, which belong to the Long Terminal Repeat (LTR) class of retrotransposon. These specific families were chosen because they had been previously associated with putative human MEs[50]. In this chapter, we extend this analysis by examining the proximity of hvCpGs to 698 retrotransposon subfamilies belonging to the ERV class and to the non-LTR-containing LINE and SINE retrotransposon classes (**Figure B.1**).

hvCpGs were enriched for proximity (within 3kb) to 7 subfamilies relative to the array background (FDR < 0.05) (**Figure 4.3**). This includes one ERV1 subfamily (LTR12C), one ERVL-MaLR subfamily (MLT1K), and five LINE-1 subfamilies (L1Meg, L1MC3, L1PA5, L1PA7, L1M5) (**Supplementary Table B.1**). The strongest evidence for enrichment was observed for L1MEg (OR = 2.19, q-value = 1.3×10^{-3}) and LTR12C (OR = 2.24, q-value = 6.1×10^{-3}) (**Figure 4.3**). LTR12C is a human-specific retrotransposon that has been recently linked to regions of systemic interindividual variation (SIV)[48]. 1.3% of de-clustered hvCpGs were within 3kb of an LTR12C element (**Supplementary Table B.1**). hvCpGs were also depleted for proximity to 7 subfamilies belonging to LINE-2 (L2b, L2c), SINE-Alu (Alujb,) and SINE-MIR (MIRb, MIRc, MIR, MIR3) families (**Supplementary Table B.1**).

Murine Variably Methylated IAP elements have increased CpG density compared to non-variable IAPs[48]. Of the 7 subfamilies significantly associated with hvCpGs (**Figure 4.3**), LTR12C has the highest CpG density (**Supplementary Figure B.4**). However, there is no difference in CpG density between hvCpG-associated elements compared to array background-associated elements (**Supplementary Figure B.4**), suggesting that retrotransposon CpG density alone is not sufficient to explain methylation variation at nearby hvCpGs.

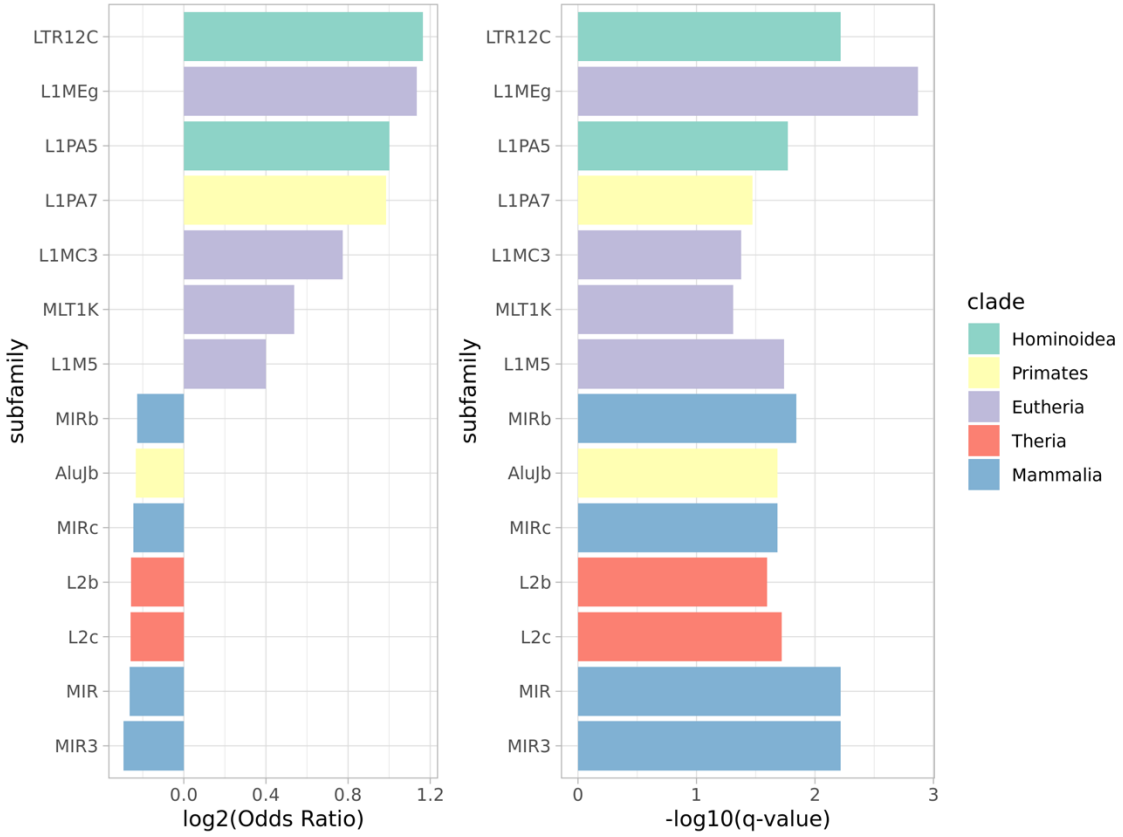


Figure 4.3. Enrichment for proximal retrotransposon subfamilies at hvCpGs. Left: Odds ratios of enrichment/depletion for proximal retrotransposon subfamilies (within 3kb) at a de-clustered set of 2,640 hvCpGs relative to array background CpGs. **Right:** Fishers Exact Test (FET) p-values, adjusted for the False Discovery Rate (FDR). Only subfamilies with FDR< 0.05 and with at least 1% of hvCpGs within 3kb are shown. Bar colours give the evolutionary clade of each subfamily obtained from Dfam database. See **Supplementary Table B.1** for groupings of these subfamilies by retrotransposon family and class.

4.3.2 Coverage of retrotransposons on the Illumina450K array

hvCpGs were identified on the Illumina450K array which is biased towards promoters and genic regions and is not suitable for profiling repetitive regions due to high rates of hybridisation artefacts [68]. We assessed retrotransposon coverage on the Illumina450K array by calculating the % of elements belonging to a given subfamily that were proximal to an Illumina450K CpG at varying proximity thresholds (0kb, 1kb and 3kb) (**Figure B.5**). 203 (29%) subfamilies had at least 10% of elements within 3kb of an array background CpG, but only 8 of these had at least 20% of elements proximal to an array background CpG (**Figure B.5**). The 203 'covered'

subfamilies were enriched for SINEs (OR ratio = 152, FET p-value = 7.76 x 10^{-27}) but depleted for ERVs (OR = 0.55, FET p-value = 8.14 x 10^{-4}) and LINEs (OR = 0.49, FET p-value = 1.44 x 10^{-3}) relative to 'uncovered' subfamilies. These findings confirm that the Illumina450K array is under-represented at transposable elements and particularly depleted for TE classes that are associated with hvCpGs and putative human MEs (ERVs and LINEs). We were therefore unable to examine whether hvCpG methylation states were associated with methylation levels at proximal retrotransposons.

4.3.3 mQTLs of hvCpGs are enriched in KZFP genes

In Chapter 3, we had used a high-powered mQTL meta-analysis carried out in Europeans[63] (see Methods) to explore genetic effects at hvCpGs, reasoning that any 'universal' mQTLs (multi-tissue, multi-ethnic) would by definition be present in European blood. Using a list of 366 KZFP genes collated from two studies (see Methods), we noted that the *trans* mQTLs for hvCpGs were enriched within KZFP genes compared to the *trans* mQTLs of the array background (10.8% compared to 4.6%, OR = 2.30, FET p-value = 5.54 x 10^{-12}). In total, there were 142 associations ($p < 1 \times 10^{-14}$) between 137 hvCpGs (falling into 92 clusters) and 105 KZFP mQTLs (located in 65 KZFP genes) (**Figure 4.4B-C, Supplementary Table B.2**). The median of the mean % methylation variance explained by these KZFP mQTLs was 1.18% (IQR = [0.8 %, 1.6%]). 12 KZFP mQTLs were associated with more than one hvCpG cluster (**Supplementary Table B.3**), indicating that these genetic variants can influence methylation at hvCpGs across the genome.

The majority (125/137) of hvCpGs associated with a KZFP mQTL were also associated with at least one other (*cis* or *trans*) mQTL. The median of the mean % methylation variance explained by these mQTLs was 2.03% [IQR = 0.86%, 4.19%], indicating that genetic variants located outside of KZFP genes can also influence hvCpG methylation. Of the 105 KZFP mQTLs associated with hvCpGs, 43 (31.4%) were also associated with at least one of 157 array background CpGs. Therefore, whilst our findings suggest that hvCpGs are more likely to be influenced by KZFP genetic variation than array background CpGs, the association with KZFP mQTLs does not fully explain the hypervariability of hvCpGs.

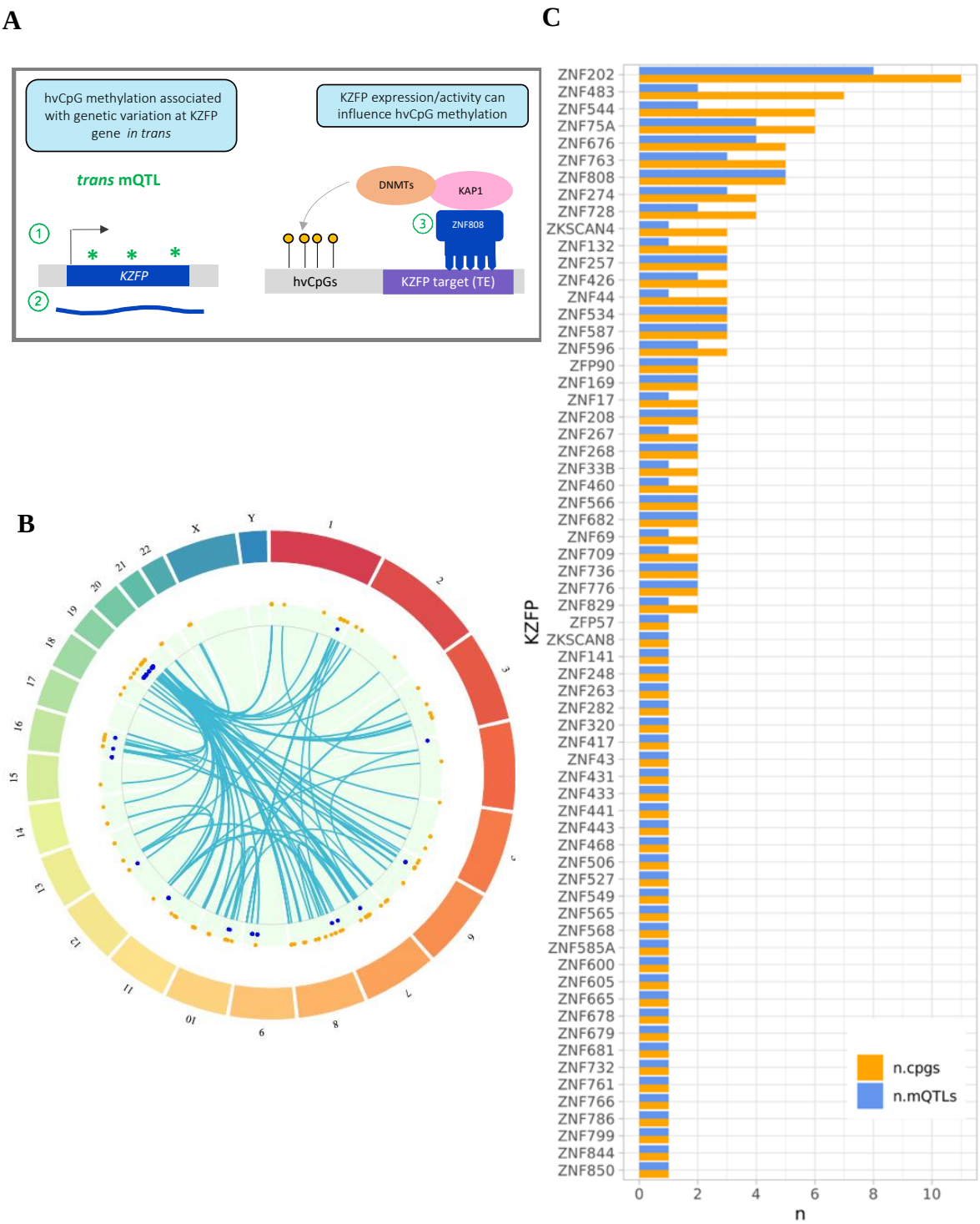


Figure 4.4. hvCpGs are enriched for *trans* mQTLs located in KZFP genes relative to the array background. **A)** Rationale for analysing KZFP mQTLs associated with hvCpGs. These mQTLs may influence 1) transcription, 2) mRNA processing and 3) protein function, indicating that KZFP expression/and or activity may influence methylation at hvCpGs. **B)** The genomic locations of 137 hvCpGs (orange points) that are associated with 105 KZFP mQTLs (dark blue points). Light blue lines represent associations between hvCpGs and their corresponding KZFP mQTLs. **C)** Bar chart showing the number of KZFP mQTLs within each KZFP gene (blue) and the number of associated hvCpGs (orange). For example, ZNF202 comprises 8 mQTLs that are associated with 11 hvCpGs.

4.3.4 Linking KZFP *trans* mQTLs to KZFP expression and activity

We reasoned that genetic variation at KZFP genes might affect KZFP expression and/or function, leading to altered binding to genomic targets and recruitment of DNMTs, with a consequent effect on hvCpG methylation (**Figure 4.4A**). The 105 KZFP mQTLs map to 65 KZFP genes (**Figure 4.4**). The majority of mQTLs (72/105) were in transcribed regions, 4 were in promoters, 18 were in exon start sites and 7 were in a KZFP protein coding sequence (1 in the KRAB domain and 6 in the C2H2 zinc finger domain) (**Figure B.6**), suggesting that these mQTLs may influence KAP1 recruitment and DNA binding affinity. To probe the influence of the 105 identified KZFP mQTLs on gene expression, we examined their overlap with eQTLs, reasoning that because 24% of associated hvCpGs show evidence of systemic interindividual variation in previous studies (SIV) (Chapter 3), KZFP mQTLs identified in blood may be eQTLs in other tissues. 89/104 *trans* mQTLs were an eQTL in at least one tissue type, and 52 of these were an eQTL for the KZFP gene in which they were located, in a median of 2 GTEx tissues (IQR = [2,7]) (**Figure B.7**). This suggests that KZFP genetic variants that alter methylation at hvCpGs *in trans* are also able to alter the post-natal expression of the given KZFP.

We next examined the proximity of hvCpGs that were associated with KZFP mQTLs to binding sites for the respective KZFP. Because we had ChIP-exo data available for 221 KZFPs, we restricted this analysis to the 85 hvCpGs associated with a *trans* mQTL located in one of these 221 KZFP genes. We found that 29/85 (34.1%) hvCpGs were located within 3kb of the respective KZFP ChIP-exo binding site. This supports the notion that the KZFPs may influence hvCpG methylation by directly binding to a nearby region.

4.3.5 hvCpGs are enriched for proximal KAP1 binding sites and KAP1-bound retrotransposons

To explore our hypothesis that hvCpG methylation variation is linked to KZFP-mediated recruitment of KAP1 in the early embryo, we examined the proximity of hvCpGs to ChIP-exo peaks for endogenous KAP1 obtained in H1-ESCs[3]. 10.1% of hvCpGs were within 3kb of a KAP1 ChIP-exo peak compared to 4.2% of the array background, representing a significant enrichment (OR = 2.62, FET p-value = 1.30×10^{-60}) (**Figure 4.5A**) that is maintained after de-clustering hvCpGs (**Figure B.8**).

Whilst 86% (30984/36801) of genome-wide KAP1 ChIP-exo peaks overlap a retrotransposon element, this only corresponds to 1.27% of all retrotransposon elements. 5.6% of hvCpGs were within 3kb of a KAP1-bound retrotransposon, representing an enrichment relative to unbound KAP1 retrotransposon elements (OR = 2.47, FET p-value = 1.71×10^{-7}). This indicates that hvCpGs are particularly enriched for proximal retrotransposons that are targeted by KZFPs in the early embryo, raising the possibility that methylation variability at some hvCpGs may be linked to KZFP-mediated silencing of retrotransposons. This finding is notable given that methylation variability at some murine MEs is postulated to be attributed to variable binding of KZFPs to IAP elements in a stochastic manner, resulting in methylation variability between isogenic littermates (**Figure 4.2**)[47].

4.3.6 Histone modifications in embryonic stem cells

In addition to recruiting DNA methyltransferases, KAP1 also recruits SETDB1 which deposits H3 Lys9 trimethylation (H3K9me3)[10]. Accordingly, hvCpGs were enriched for overlaps with H3K9me3 modifications and for dual/tri histone modifications that included H3K9me3 in H1-ESCs compared to the array background (**Figure 4.5B-C**). hvCpGs were also notably depleted for solo H3K36me3 modifications which are usually found in gene bodies of transcriptionally active genes[69] (OR = 0.21, FET p-value = 7.2×10^{-72}) (**Figure 4.5B-C**). The enrichment for H3K9me3 in H1-ESCs supports the notion that hvCpGs are enriched for proximity to regions targeted by KZFP-KAP1 epigenetic silencing machinery in early development.

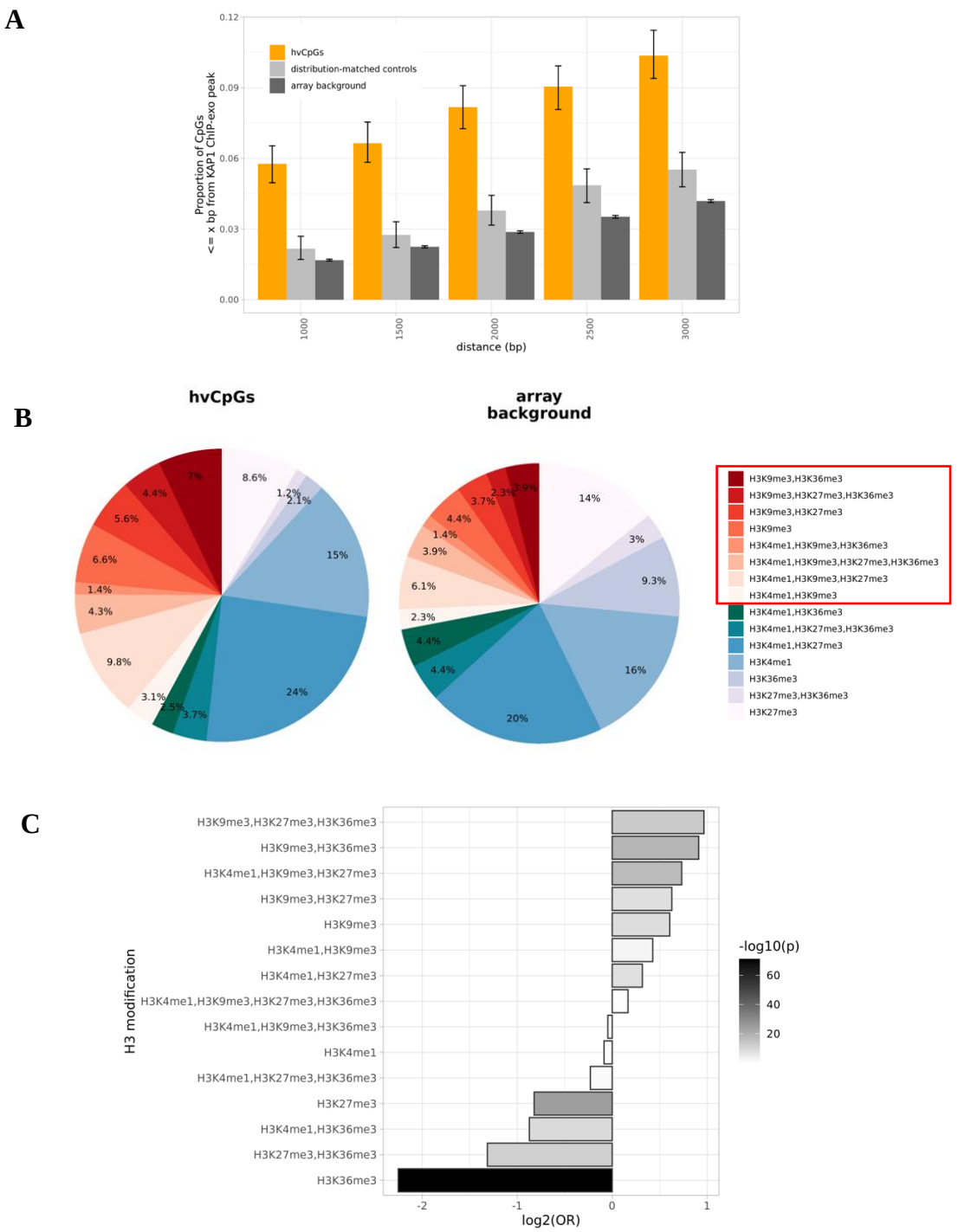


Figure 4.5. hvCpGs are enriched for proximal KAP1 binding sites and H3K9me3 modifications in H1-ESCs relative to the array background. A) The proportion of 3566 hvCpGs (orange), 3566 distribution-matched controls (light grey) and array background CpGs (dark grey) located ≤x bp from nearest KAP1 ChIP-exo peak identified in H1-ESCs by Imbeault *et al.* Error bars are bootstrapped confidence intervals. **B)** Pie charts showing the proportion of 3750 hvCpGs (left) and 318699 array background CpGs (right) overlapping each H3 modification signature using 'broadPeaks' ENCODE data obtained in H1-ESCs. **C)** Odds ratios for the enrichment/depletion of hvCpGs overlapping each H3 modifications signature relative to the array background. Bar shade gives the Fishers Exact Test (FET) p-values. Only modifications with FDR-adjusted p-values ('q-values') < 0.05 are shown.

4.3.7 hvCpGs are enriched for proximal binding sites of KZFPs linked PofOm

To further investigate KZFP-KAP1 epigenetic silencing and links to hvCpG methylation, we next examined hvCpG proximity to KZFP binding sites, using ChIP-exo data for 221 KZFPs obtained in HEK293T cells[3] (see Methods). 72% of hvCpGs were located within 3kb of a ChIP-exo peak for any KZFP, no different from the array background. We next assessed whether hvCpGs are enriched for proximal binding sites for specific KZFPs using our de-clustered set of hvCpGs to minimise potential biases introduced by clustering of hvCpGs. We observed a significant enrichment (FDR < 0.05) for proximity to ZFP57 binding sites (proximity thresholds set at 0kb, 1kb and 2kb, **Figure B.9**, **Supplementary Table B.4**). We also observed an enrichment for proximity to ZNF445 binding sites (proximity thresholds at 0kb, 1kb, and 2kb) and ZNF2 binding sites (proximity thresholds at 2kb and 3kb) (**Figure B.9**, **Supplementary Table B.4**).

ZFP57 and ZNF445 have known roles in regulating parent of origin specific methylation (PofOm)[70,71], and in Chapter 3 we found that hvCpGs are enriched for regions of PofOm. The function of ZNF2 is unknown. Using a published set of regions showing PofOm[61], we observed that ZNF2 binding sites overlap the next highest proportion of PofOm regions after ZFP57 and ZNF445 (**Figure B.10A-B**). ZNF2 appears to have both shared and unique targets with these KZFPs (**Figure B.10C**), raising the possibility that it may also be involved in regulating PofOm in the early embryo.

4.3.8 Investigating hvCpG subsets based on notable features

In Chapter 3 we reported that hvCpGs are enriched for several features including SIV CpGs, ESS CpGs, MZ twinning CpGs and regions of PofOm. In this chapter, we further reported enrichments for proximal retrotransposons (**Figure 4.3**), KAP1 ChIP-exo peaks (**Figure 4.5A**) and KZFP ChIP-exo peaks (**Figure B.9)**. Together, these enrichments support the notion that methylation establishment at hvCpGs may occur in early development and, at some loci, be linked to epigenetic silencing of transposable elements and regions of PofOm. To gain further insight into the timing and mechanism of DNAm establishment at hvCpGs, we examined whether hvCpGs can be divided into subsets based on the features they show. We found no evidence of distinct hvCpG subsets (see **Figure B.11**).

4.4 Discussion

In this chapter, we have probed KZFP-mediated silencing of TEs as a mechanism influencing methylation variability in early development. We did this by examining the association between hvCpG methylation and genetic variation at KZFP genes *in trans*, and the proximity of hvCpGs to retrotransposons, KAP1 binding sites and KZFP binding sites. Our findings suggest that methylation at some hvCpGs may be linked to KZFP-mediated silencing of retrotransposons and regions of parent-of-origin specific methylation in the early embryo.

The enrichment for KZFP mQTL associations amongst hvCpGs supports the notion that hvCpG methylation at a minority of hvCpGs can be influenced by KZFP expression, protein function and binding affinity. These mQTLs might also influence mRNA splicing, as splice isoforms that have altered KAP1 recruitment abilities have been found for several KZFPs[72–75]. Whilst mQTLs were identified in blood, this mechanism is likely to be common across tissues because hvCpGs show tissue-independent methylation variation and many show evidence of SIV. Supporting this, 35% of associated hvCpGs were near a binding site of the respective KZFP identified in HEK293T cells, and many KZFP mQTLs were eQTLs in other tissue types. Although only a small proportion (3.3%) of hvCpGs were associated with a KZFP mQTL, the mQTL meta-analysis we used is estimated to be 22.5-fold underpowered to detect mQTLs[63], meaning that many more hvCpGs may be associated with KZFP mQTL.

Importantly, the association between hvCpGs and KZFP mQTLs does not imply that KZFPs contribute to methylation variation at hvCpGs solely via genetic effects. At murine MEs, methylation variation between isogenic mice is attributed to variable binding affinity of KZFPs, resulting in stochastic methylation establishment at target IAP elements in early development that is stably maintained throughout cell division[47]. Given that hvCpGs are enriched for loci that show discordance between MZ twins and for loci that show sensitivity to the periconceptional environment (Chapter 3), it is possible that KZFPs may similarly contribute to environmental and stochastic DNAm establishment in early human development. Loss of KZFP binding at murine MEs is postulated to allow ZF-CxxC containing proteins such as TET1 to subsequently bind and protect the target region from further epigenetic silencing[48]. Notably, TET activity in ESCs is upregulated by vitamin C[76,77] and is also associated with increased

availability of alpha-ketoglutarate after glucose-injection in mice[78]. In this manner, KZFP binding site occupancy may influence site-specific susceptibility to environmental perturbation. Furthermore, KZFPs can interact with a number of cofactors beyond those required for epigenetic silencing[79] , raising the possibility that KZFP activity may be influenced by co-factors that are themselves environmentally labile. We speculate that G x E effects could also be at play in these scenarios, where a certain genotype might reduce KZFP DNA binding affinity or alter KZFP binding partner interactions.

We found that hvCpGs were enriched for proximity to KAP1 binding sites identified in H1-ESCs compared to array background CpGs, in agreement with previous findings at putative MEs using HEK293T cells[50]. The enrichment for H3K9me3 modifications in H1-ESCs further supports the notion that hvCpGs are proximal to regions targeted by KZFP-KAP1 epigenetic silencing machinery in the early embryo, as H3K9me3 is deposited by KAP1-recruited SETDB1[10]. However, despite KZFPs being the main recruiters of KAP1 in the early embryo, we did not observe an enrichment for proximal KZFP binding sites in HEK293T cells. This is likely due to the fact that the overexpression of KZFPs in HEK293T cells does not represent a physiologically relevant model. To our knowledge there is no publicly available dataset of KZFP ChIP-exo/ChIP-seq in human ESCs. We also note that KZFP expression is highly dynamic during embryogenesis[80,81] meaning that KAP1 binding sites obtained in H1-ESCs cells (derived from the inner cell mass), only give insight into epigenetic silencing at a relatively late stage of the preimplantation embryo. Therefore, whilst 10% of hvCpGs are within 3kb of a KAP1 binding site, it is possible that other hvCpGs may be targeted by KZFPs earlier in development.

KZFPs are known to epigenetically silence TEs in the early embryo by depositing DNAm and H3K9me3 that induce chromatin condensation[3,70,71]. Accordingly, hvCpGs were enriched for proximity to retrotransposon elements bound by KAP1 in H1-ESCs compared to those that were not, indicating that methylation variability at some hvCpGs may be linked to KZFP-mediated silencing of retrotransposons. Methylation states established by KZFP/KAP1 at TEs have been reported to be stably maintained into post-natal tissues[26] suggesting that this mechanism could play a role in establishing SIV that is maintained throughout tissue differentiation at hvCpGs. Whilst we found no evidence that CpG density of TEs is associated with methylation variability at hvCpGs, the underlying TE sequence may also infer a greater susceptibility to being variably methylated and requires further investigation[48].

A minority of hvCpGs were proximal to an LTR12C element, a primate-specific retrotransposon that has recently been linked to regions of SIV[48]. LTR12C loci do not undergo complete reprogramming in primordial germ cells as they are protected by KZFPs[82], meaning that it is possible that methylation states at associated hvCpGs could be meiotically heritable, making them interesting candidates for the study of transgenerational epigenetic inheritance. Aberrant LTR12C activation can also result in promoter activation in human tumour cells[23,83], which is notable given the associations between putative human MEs and cancer[84].

The enrichment for hvCpGs amongst regions of PofOm (Chapter 3) as well as their proximity to ZFP57 and ZNF445 binding sites, raises the possibility that methylation variability at some hvCpGs may be linked to KZFP-mediated maintenance of parent-of-origin specific methylation in the early embryo. The region upstream of ZFP57 shows evidence of being putative human ME[49] and has been associated with periconceptional and gestational environmental exposures[85,86]. This suggests that ZFP57 may mediate environmental/stochastic effects at multiple hvCpGs across the genome[85].

The association with ZNF445 binding sites is particularly notable because it is expressed during the first few cleavage divisions immediately after fertilisation, whilst ZFP57 is expressed after zygotic genome activation[71,75]. This could be a potential mechanism by which methylation variability at some hvCpGs is established before MZ twinning events in the cleavage-stage embryo[84].

We also highlighted ZNF2 as a potential novel regulator of PofOm. Many of the PofOm regions bound by ZNF2 were newly identified in the screen by Zink *et al.*[61], perhaps explaining why ZNF2 may not have been highlighted in a similar previous screen[71]. However, the timing of expression of ZNF2 requires further investigation as it is not expressed before zygotic genome activation[75]. Ultimate validation of ZNF2 as a putative novel regulator of PofOm would require *in vivo* knock-out experiments in mice.

A major limitation of this work is that the Illumina450K has very poor retrotransposon coverage, meaning that we were unable to directly interrogate methylation levels at hvCpG-associated retrotransposons. Secondly, the extent to which hvCpGs are targeted by KZFPs is unclear given that the mQTL analysis was underpowered to detect *trans* mQTLs, and that H1-

ESCs used to identify KAP1 binding sites only represent a late pre-implantation embryonic stage. Thirdly, we note that the overlap between mQTLs and eQTLs may be attributed linkage disequilibrium between two separate causal variants rather than be indicative of a causal relationship between methylation and gene expression. The application of multiple trait colocalization[87] will be used in future analyses to determine whether KZFP mQTLs that influence hvCpG methylation are also those that influence KZFP gene expression. Fourthly, whilst our analysis of KZFP mQTLs indicates that KZFPs can influence methylation at hvCpGs, the role of KZFP-mediated silencing of TEs in driving stochastic and/or environmental effects on methylation variation at hvCpGs as a result of variable KZFP binding remains an open question.

In summary, we have investigated KZFP-mediated silencing of TEs as a possible mechanism that may influence methylation variation at some hvCpGs in early development. This may have significant implications for deepening our understanding of how variable DNAm patterns are shaped in early development and their potential contribution to trait variation.

4.5 References

1. Huntley S, Baggott DM, Hamilton AT, et al. A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res*. 2006;16(5):669-677.

2. Emerson RO, Thomas JH. Adaptive evolution in zinc finger transcription factors. *PLoS Genet*. 2009;5(1):e1000325.

3. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*. 2017;543(7646):550-554.

4. Najafabadi HS, Garton M, Weirauch MT, et al. Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. *Genome Biol*. 2017;18(1):1-15.

5. Wolfe SA, Nekludova L, Pabo CO. DNA recognition by (Cys2His2) zinc finger proteins. *Annu Rev Biophys Biomol Struct*. 2000;29:183.

6. Witzgall R, O'Leary E, Leaf A, Onaldi D, Bonventre J v. The Krüppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. *Proceedings of the National Academy of Sciences*. 1994;91(10):4514-4518.

7. Friedman JR, Fredericks WJ, Jensen DE, et al. KAP-1, a novel corepressor for the highly conserved KRAB repression domain. *Genes Dev*. 1996;10(16):2067-2078.

8. Kim SS, Chen YM, O'Leary E, Witzgall R, Vidal M, Bonventre J v. A novel member of the RING finger family, KRIP-1, associates with the KRAB-A transcriptional repressor domain of zinc finger proteins. *Proceedings of the National Academy of Sciences*. 1996;93(26):15299-15304.

9. Moosmann P, Georgiev O, le Douarin B, Bourquin JP, Schaffner W. Transcriptional repression by RING finger protein TIF1β that interacts with the KRAB repressor domain of KOX1. *Nucleic Acids Res*. 1996;24(24):4859-4867.

10. Schultz DC, Ayyanathan K, Negorev D, Maul GG, Rauscher FJ. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev*. 2002;16(8):919-932.

11. Ryan RF, Schultz DC, Ayyanathan K, et al. KAP-1 corepressor protein interacts and colocalizes with heterochromatic and euchromatic HP1 proteins: a potential role for Kruppel-associated box–zinc finger proteins in heterochromatin-mediated gene silencing. *Mol Cell Biol*. 1999;19(6):4366-4378.

12. Schultz DC, Friedman JR, Rauscher FJ. Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative

unit that recruits a novel isoform of the Mi-2α subunit of NuRD. *Genes Dev.* 2001;15(4):428-443.

13. Quenneville S, Verde G, Corsinotti A, et al. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol Cell.* 2011;44(3):361-372.

14. Turelli P, Castro-Diaz N, Marzetta F, et al. Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements. *Genome Res.* 2014;24(8):1260-1270.

15. Najafabadi HS, Mnaimneh S, Schmitges FW, et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol.* 2015;33(5):555-562.

16. Chinwalla A, Cook LL, Delehaunty KD, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature.* 2002;420(6915):520-562.

17. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860-921. doi:10.1038/35057062

18. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. *Mob DNA.* 2016;7(1):9.

19. Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. *Annu Rev Genet.* 2012;46:651.

20. Kazazian HH. Mobile DNA transposition in somatic cells. *BMC Biol.* 2011;9(1):1-4.

21. Elbarbary RA, Lucas BA, Maquat LE. Retrotransposons as regulators of gene expression. *Science (1979).* 2016;351(6274):aac7247.

22. Daniel C, Behm M, Öhman M. The role of Alu elements in the cis-regulation of RNA processing. *Cellular and Molecular Life Sciences.* 2015;72(21):4063-4076.

23. Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. *Mob DNA.* 2016;7(1):1-21.

24. Kassiotis G, Stoye JP. Immune responses to endogenous retroelements: taking the bad with the good. *Nat Rev Immunol.* 2016;16(4):207-219.

25. Goodier JL. Restricting retrotransposons: a review. *Mob DNA.* 2016;7(1):1-30.

26. Quenneville S, Turelli P, Bojkowska K, et al. The KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development. *Cell Rep.* 2012;2(4):766-773.

27. Rowe HM, Friedli M, Offner S, et al. De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET. *Development.* 2013;140(3):519-529.

28. Wolf G, Yang P, Füchtbauer AC, et al. The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes Dev.* 2015;29(5):538-554.

29. Rowe HM, Jakobsson J, Mesnard D, et al. KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*. 2010;463(7278):237-240.

30. Wiznerowicz M, Jakobsson J, Szulc J, et al. The Krüppel-associated box repressor domain can trigger de novo promoter methylation during mouse early embryogenesis. *Journal of Biological Chemistry*. 2007;282(47):34535-34541.

31. Wolf D, Goff SP. Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature*. 2009;458(7242):1201-1204.

32. Wolf G, Greenberg D, Macfarlan TS. Spotting the enemy within: Targeted silencing of foreign DNA in mammalian genomes by the Krüppel-associated box zinc finger protein family. *Mob DNA*. 2015;6(1):1-20.

33. Ecco G, Cassano M, Kauzlaric A, et al. Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. *Dev Cell*. 2016;36(6):611-623.

34. Collins PL, Kyle KE, Egawa T, Shinkai Y, Oltz EM. The histone methyltransferase SETDB1 represses endogenous and exogenous retroviruses in B lymphocytes. *Proceedings of the National Academy of Sciences*. 2015;112(27):8367-8372.

35. Jacobs FMJ, Greenberg D, Nguyen N, et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*. 2014;516(7530):242-245.

36. Thomas JH, Schneider S. Coevolution of retroelements and tandem zinc finger genes. *Genome Res*. 2011;21(11):1800-1812.

37. Bruno M, Mahgoub M, Macfarlan TS. The Arms Race Between KRAB–Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. *Annu Rev Genet*. 2019;53(1):393-416. doi:

38. Fernandes JD, Haeussler M, Armstrong J, et al. KRAB Zinc Finger Proteins coordinate across evolutionary time scales to battle retroelements. *bioRxiv* (2018).

39. Bourque G, Leong B, Vega VB, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*. 2008;18(11):1752-1762.

39. Pehrsson EC, Choudhary MNK, Sundaram V, Wang T. The epigenomic landscape of transposable elements across normal human development and anatomy. *Nature Communications 2019 10:1*. 2019;10(1):1-16. doi:10.1038/s41467-019-13555-x

41. Sundaram V, Wysocka J. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society B*. 2020;375(1795):20190347.

42. Ecco G, Imbeault M, Trono D. KRAB zinc finger proteins. *Development (Cambridge)*. 2017;144(15):2719-2729.

43. Pontis J, Planet E, Offner S, et al. Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell*. 2019;24(5):724-735.

44. Iouranova A, Grun D, Rossy T, et al. KRAB zinc finger protein ZNF676 controls the transcriptional influence of LTR12-related endogenous retrovirus sequences. *Mob DNA*. 2022;13(1):4.

45. Elmer JL, Hay AD, Kessler NJ, Bertozzi TM, Ainscough EAC, Ferguson-Smith AC. Genomic properties of variably methylated retrotransposons in mouse. *Mob DNA*. 2021;12(1).

46. Kazachenka A, Bertozzi TM, Sjoberg-Herrera MK, et al. Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell*. 2018;175(5):1259-1271.e13.

47. Bertozzi TM, Elmer JL, Macfarlan TS, Ferguson-Smith AC. KRAB zinc finger protein diversification drives mammalian interindividual methylation variability. *Proceedings of the National Academy of Sciences*. 2020;117(49):31290-31300.

48. Costello KR, Leung A, Trac C, et al. Sequence features of retrotransposons allow for epigenetic variability. *Elife*. 2021;10:e71104.

49. Silver MJ, Kessler NJ, Hennig BJ, et al. Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome Biol*. 2015;16(1):1-14.

50. Kessler NJ, Waterland RA, Prentice AM, Silver MJ. Establishment of environmentally sensitive DNA methylation states in the very early human embryo. *Sci Adv*. 2018;4(7):eaat2624.

51. Kuehnen P, Mischke M, Wiegand S, et al. An Alu Element–Associated Hypermethylation Variant of the POMC Gene Is Associated with Childhood Obesity. *PLoS Genet*. 2012;8(3):e1002543.

52. Hop PJ, Luijk R, Daxinger L, et al. Genome-wide identification of genes regulating DNA methylation using genetic anchors for causal inference. *Genome Biol*. 2020;21(1).

53. Bretherick AD, Zeng Y, Walker RM, et al. Genetic control of KRAB-ZFP genes explains distal CpG-site methylation which associates with human disease phenotypes. *bioRxiv* (2022).

54. Derakhshan M, Kessler NJ, Ishida M, et al. Tissue-and ethnicity-independent hypervariable DNA methylation states show evidence of establishment in the early human embryo. *Nucleic Acids Res*. 2022;50(12):6735-6752.

55. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA*. 2021;12(1):2.

56. Hubley R, Finn RD, Clements J, et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res*. 2016;44(D1):D81-D89.

57. Wheeler TJ, Clements J, Eddy SR, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res*. 2012;41(D1):D70-D82.

58. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*. 2007;23(14):1846-1847.

59. Rhee HS, Pugh BF. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol*. 2012;100(1):21-24.

60. Vazin T, Freed WJ. Human embryonic stem cells: derivation, culture, and differentiation: a review. *Restor Neurol Neurosci*. 2010;28(4):589-603.

61. Zink F, Magnusdottir DN, Magnusson OT, et al. Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nat Genet*. 2018;50(11):1542-1552.

62. Wang J, Lunyak V v, Jordan IK. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. *Bioinformatics*. 2013;29(4):492-493.

63. Min JL, Moffitt T, Caspi A, Arseneault L, Relton C. Genomic and phenomic insights from an atlas of genetic effects on DNA methylation. *Nat Genet*. 2021;53(9):1311-1321.

64. Shen P, Xu A, Hou Y, et al. Conserved paradoxical relationships among the evolutionary, structural and expressional features of KRAB zinc-finger proteins reveal their special functional characteristics. *BMC Molecular and Cell Biology 2021 22:1*. 2021;22(1):1-15.

65. EnsDb RJ. Hsapiens. v75: Ensembl based annotation package. *R package version*. 2017;2(0).

66. Hahne F, Ivanek R. Visualizing genomic data using Gviz and bioconductor. In: *Statistical Genomics*. Springer; 2016:335-351.

67. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40(D1):D930-D934.

68. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288-295.

69. Wagner EJ, Carpenter PB. Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol*. 2012;13(2):115-126. doi:10.1038/nrm3274

70. Takahashi N, Gray D, Strogantsev R, et al. ZFP57 and the targeted maintenance of postfertilization genomic imprints. In: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol 80. Cold Spring Harbor Laboratory Press; 2016:177-187.

71. Takahashi N, Coluccio A, Thorball CW, et al. ZNF445 is a primary regulator of genomic imprinting. *Genes Dev*. 2019;33(1-2):49-54.

72. Shao H, Zhu C, Zhao Z, et al. KRAB-containing zinc finger gene ZNF268 encodes multiple alternatively spliced isoforms that contain transcription regulatory domains. *Int J Mol Med*. 2006;18(3):457-463.

73. Côté F, Boisvert FM, Grondin B, et al. Alternative promoter usage and splicing of ZNF74 multifinger gene produce protein isoforms with a different repressor activity and nuclear partitioning. *DNA Cell Biol*. 2001;20(3):159-173.

74. Sun L, Gu S, Li N, et al. A novel zinc finger gene ZNF468 with two co-expressional splice variants, ZNF468. 1 and ZNF468. 2. *Biochem Genet*. 2005;43(5):271-286.

75. Monteagudo-Sánchez A, Hernandez Mora JR, Simon C, et al. The role of ZFP57 and additional KRAB-zinc finger proteins in the maintenance of human imprinted methylation and multi-locus imprinting disturbances. *Nucleic Acids Res*. 2020;48(20):11394-11407.

76. Blaschke K, Ebata KT, Karimi MM, et al. Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature*. 2013;500(7461):222-226.

77. Minor EA, Young JI, Wang G. Ascorbate Induces Ten-Eleven Translocation (Tet) Methylcytosine Dioxygenase-mediated Generation of 5-Hydroxymethylcytosine*♦. *Journal of Biological Chemistry*. 2013;288(19):13669-13674.

78. Yang H, Lin H, Xu H, et al. TET-catalyzed 5-methylcytosine hydroxylation is dynamically regulated by metabolites. *Cell Res*. 2014;24(8):1017-1020.

79. Helleboid PY, Heusel M, Duc J, et al. The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. *EMBO J*. 2019;38(18):e101220.

80. Corsinotti A, Kapopoulou A, Gubelmann C, et al. Global and stage specific patterns of Krüppel-associated-box zinc finger protein gene expression in murine early embryonic cells. *PLoS One*. 2013;8(2):e56721.

81. Theunissen TW, Friedli M, He Y, et al. Molecular criteria for defining the naive human pluripotent state. *Cell Stem Cell*. 2016;19(4):502-515.

82. Dietmann S, Keogh MJ, Tang W, et al. Transposable elements resistant to epigenetic resetting in the human germline are epigenetic hotspots for development and disease. *bioRxiv* (2020).

83. Krönung SK, Beyer U, Chiaramonte ML, Dolfini D, Mantovani R, Dobbelstein M. LTR12 promoter activation in a broad range of human tumor cells by HDAC inhibition. *Oncotarget*. 2016;7(23):33484.

84. van Baak TE, Coarfa C, Dugué PA, et al. Epigenetic supersimilarity of monozygotic twin pairs. *Genome Biol*. 2018;19(1):2.

85. Silver MJ, Saffari A, Kessler NJ, et al. Environmentally sensitive hotspots in the methylome of the early human embryo. Deelen J, ed. *Elife*. 2022;11:e72031.

86. Amarasekera M, Martino D, Ashley S, et al. Genome-wide DNA methylation profiling identifies a folate-sensitive region of differential methylation upstream of ZFP57-imprinting regulator in humans. *The FASEB Journal*. 2014;28(9):4068-4076.

87. Giambartolomei C, Zhenli Liu J, Zhang W, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*. 2018;34(15):2538-2545.

Chapter 5

5. A hypervariably methylated region upstream of *C8orf31* is linked to ZNF808-mediated silencing of MER11C and has a causal effect on insulin dynamics

Rationale of chapter

This chapter follows on from Chapter 4, where we found that hvCpGs are enriched for KZFP mQTL associations relative to CpGs covered on the Illumina450K array (Figure 4.4C). A subsequent discussion with Dr Michael Imbeault (University of Cambridge), drew our attention to the KZFP *ZNF808*, which has recently been found to be required for cell fate specification during pancreas development by silencing MER11C retrotransposons in early development. Interestingly, we noted that most hvCpGs associated with *ZNF808* mQTLs are located in a differentially methylated region (DMR) that has been recently associated with fasting insulin levels measured in Gambian children by our group, upstream of the uncharacterised gene *C8orf31*. These observations motivated us to investigate the phenotypic consequences of DNAm modulation at *C8orf31* via ZNF808-mediated silencing of MER11C in early development.

Summary of chapter

The differentially methylated region (DMR) upstream of *C8orf31* is associated with fasting insulin levels and first-phase insulin secretion in Gambian children. This DMR shows evidence of causal effect on insulin secretion rate, acute insulin response and T1 diabetes in Europeans in a Two-sample Mendelian randomisation analysis. The DMR is associated with genetic variation at *ZNF808 in trans* and with genetic variation at an intragenic MER11C retrotransposon element *in cis* that is bound by ZNF808. The DMR overlaps previously identified metastable epiallele CpGs and is associated with *C8orf31* gene expression in thyroid. Together, these findings position ZNF808-mediated silencing of MER11C as a mechanism influencing *C8orf31* methylation in early development, with potential consequences for insulin dynamics in later life.

Notes

The Supplementary Figures and Tables for this chapter can be found in Appendix C.

5.1 Background

KRAB zinc finger proteins (KZFPs) are abundant in the human genome and, together with their TE targets, are hypothesised to play a role in the evolution of gene regulatory networks[1]. Recently, epigenetic silencing of the retrotransposon MER11C (belonging to the ERVK family of LTR-retrotransposons) by ZNF808 was found to be important for pancreas development[2]. Homozygous loss of function mutations in *ZNF808* were found to cause pancreatic agenesis[2], a rare phenotype characterised by insulin-dependent neonatal diabetes and exocrine pancreatic insufficiency[3]. *In vitro* inactivation of ZNF808 in embryonic stem cells was found to lead to aberrant activation of MER11C and the upregulation of genes linked to a liver rather than pancreatic identity[2]. Both ZNF808 and MER11C are primate-specific, providing an example of a KZFP and its TE target that is involved in wiring species-specific regulatory pathways[2].

The downstream genes involved in the pathway in which ZNF808-mediated silencing of MER11C influences pancreas development/function are yet to be characterised. Here, we focus on a differentially methylated region (DMR) upstream of *C8orf31* that was previously associated with fasting insulin levels in Gambian children by Antoun *et al*[4]. This DMR comprises hvCpGs (identified in Chapter 3) that are associated with ZNF808 mQTLs (Chapter 4), positioning *C8orf31* as an example of a gene at which KZFP-mediated silencing of TEs may influence downstream phenotypes.

C8orf31 is predicted to be a long non-coding RNA (alias LINC02904) and has been previously linked to glucose metabolism: genetic variation at *C8orf31* is associated with hyperglycaemia in non-diabetic individuals[5] and one CpG annotated to *C8orf31* on the Illumina27K array is differentially methylated in subcutaneous adipose tissue between MZ twins discordant for T2 diabetes[6]. *C8orf31* has additionally been found to have an oncogenic role in triple negative breast cancer[7] and to be upregulated in bladder cancer lines[8] and in early-stage abdominal Wilms tumours in children[9]. There is also evidence that *C8orf31* is sensitive to the environment, positioning it as a possible mediator of environmental risk on phenotype: *C8orf31* expression measured in cord blood is positively associated with prenatal maternal blood plasma concentrations of synthetic perfluoroalkyl substances (PFAS)[10], and *C8orf31* methylation

measured in whole blood in adults is associated with levels of polychlorinated bisphenols in serum[11].

In this chapter, we: i) explore whether ZNF808 expression/activity can influence methylation at the *C8orf31* DMR by silencing MER11C, ii) probe the causal pathways between *C8orf31* DMR methylation and phenotypes linked to diabetes pathogenesis and glucose metabolism using a Two-sample Mendelian randomisation framework, and iii) investigate the influence of the early developmental environment on methylation at this region. We additionally characterise *C8orf31* by examining its expression across tissues and by identifying genes with sequence homology.

Our analyses suggest that the DMR upstream of *C8orf31* has a causal effect on beta cell function and is influenced by ZNF808-mediated silencing of MER11C. This provides an example of how KZFP-mediated silencing of TEs in early development may have phenotypic consequences in later life.

5.2 Methods

5.2.1 Study cohorts

This study analyses genotype, methylation, and phenotype data from two cohorts of children born in rural West Kiang, The Gambia (**Figure 5.1**). Detailed descriptions of the ENID (Early Nutrition and Immune Development, ISRCTN49285450) trial and the EMPHASIS (Epigenetic Mechanisms linking Periconceptional nutrition and Health Assessed in India and sub-Saharan Africa, ISRCTN14266771) study are published elsewhere[12,13]. Briefly, for the ENID cohort, DNA was extracted from blood samples from 600 children aged 2 years that were born between 2011 and 2013. For the EMPHASIS cohort, DNA was extracted from blood samples from 289 Gambian children aged 7-9 years that were born between 2006 and 2008.

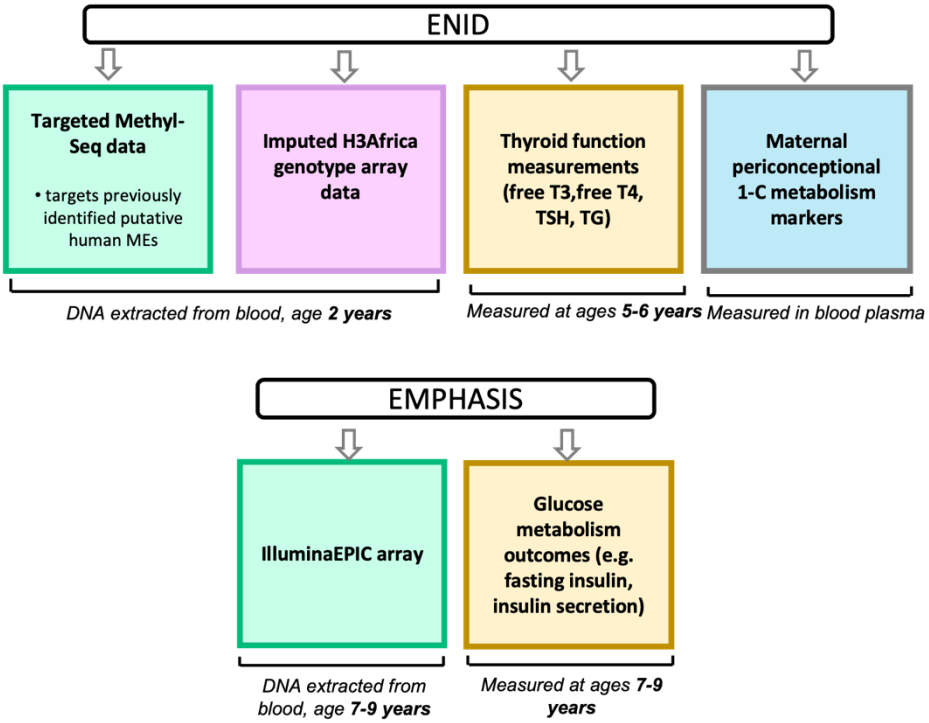


Figure 5.1. Overview of data available for ENID and EMPHASIS cohorts of Gambian children. The main cohort we used was the ENID cohort for which we had methylation (green) and genotype (purple) data available, generated from DNA extracted from blood at age 2years. We also had follow-up thyroid function measurements from a subset of the same children at ages 5-6 years (yellow), and measures of maternal periconceptional 1-C metabolism markers (blue). The EMPHASIS cohort was used to examine the association between concurrent methylation and traits linked to glucose metabolism (see 5.2.7) at ages 7-9 years.

5.2.2 Methylation data

EMPHASIS methylation data had been previously generated on the IlluminaEPIC array as previously described[14]. Briefly, this involved removing samples with sex mismatches and outlier samples, and removing low detection p value-, sex chromosome- and cross-reactive probes. We used this data to examine methylation at 9 IlluminaEPIC CpGs located inside the fasting insulin differentially methylated region (DMR) identified by Antoun et al. (see 'Previously identified CpG sets' below) (**Figure 5.2**).

ENID methylation data had been generated using the Agilent SureSelect Methyl-Seq targeted capture system as previously described[15]. This dataset captures putative ME regions identified in previous studies[16–18] and their flanking regions. This dataset was generated using hg38 coordinates, but for consistency with the rest of our analyses, we converted hg38 coordinates into hg19 coordinates using the *liftOver()* function in the rtracklayer R package[19]. We examined CpG methylation across the fasting insulin DMR and flanking regions (1kb on either side) (chr8: 144119106 – 144121681, hg19 coordinates). Of the 508 CpGs falling in this region, we selected those CpGs with a minimum read depth of 30x in at least 300 individuals, yielding 17 CpGs covered in a median of 480 individuals (IQR = [452, 501]). The average read depth was a median of 56x (IQR = [43,62]) across CpG-individual pairs. Note, 12 of the 17 CpGs fall into the DMR, and the remaining 5 overlap the 1st exon of *C8orf31* (**Figure 5.2**). These 5 CpGs are not covered on the IlluminaEPIC array but were retained in our analyses. The 17 CpGs fell into 4 correlated 'CpG blocks', each with an average inter-CpG Pearson r > 0.7 (**Figure C.1**).

5.2.3 Genotype data

Genotype data for 445 ENID individuals were generated for this study using the Human Heredity and Health in Africa ('H3Africa' or 'H3A') array[20], which is designed to have dense coverage of common genetic variants in multiple African populations. We called genotypes for 2,263,335 SNPs by applying the Linux-compatible 'gtc2vcf' pipeline available in GitHub (https://github.com/freeseek/gtc2vcf). This pipeline runs Illumina's proprietary GenCall algorithm implemented by Illumina Array Analysis Platform (IAAP) on IDAT files using manifest and cluster files (which we downloaded from the H3Africa website). We then removed low-quality SNPs (GenTrain score < 0.7) and samples with missingness in > 5% of SNPs,

leaving 429 ENID samples. These genotype data were then imputed by collaborator Noah Kessler using SHAPEIT v4.2.2, IMPUTE5 v1.1.5 and 1000 Genomes (1000G) Phase 3 reference panel. SNPs with an IMPUTE5 'info' metric < 0.9 were excluded to ensure a high imputation quality.

5.2.4 CpG sets

Hypervariable CpGs (hvCpGs)

4,143 Illumina450K CpGs identified in Chapter 3 that have top 5% methylation variance in many tissues and ethnicities, and which show several features linking them to the early embryo[21].

Fasting insulin DMR

Antoun *et al.*[4] previously identified a DMR associated with fasting insulin levels in the EMPHASIS cohort of Gambian children using the DMRcate method[22] and IlluminaEPIC array data. The DMR (orange region, **Figure 5.2**) spans 576 bp (chr8:144,120,106 - 144,120,681, hg19 coordinates) and is located immediately upstream of *C8orf31*.

Putative metastable epiallele (ME) CpGs

Systemic interindividual variation (SIV) CpGs

CpGs that show covariation across tissues derived from different germ layers, indicative of methylation establishment before germ layer differentiation. We used a set of 4,351 SIV-CpGs collated from four different studies which were covered on the Illumina450K array [16,18,23,24]. Further details on these studies are given in Supplementary Table 6 of Derakhshan *et al.*[21] and in Appendix A, Supplementary Table 6 of this thesis.

Epigenetic supersimilarity (ESS) CpGs

A set of 1,580 CpGs on the Illumina450K array identified by Van Baak *et al.*[17] that show greater-than-expected similarity between monozygotic (MZ) twins than dizygotic (DZ) twins, indicating methylation establishment before MZ cleavage[17].

Season-of-conception ('SoC') CpGs

A set of 768 CpGs associated with season-of-conception ('SoC') in the ENID cohort of Gambian 2-year-olds (FDR < 5%) identified by Silver *et al.*[25]. SoC effects at these CpGs are suggestive of sensitivity to the periconceptional environment [25].

5.2.5 ZNF808 binding sites and MER11C locations

The hg19 coordinates of MER11C elements were downloaded from RepeatMasker as described in Chapter 4. Chromatin Immunoprecipitation-exonuclease (ChIP-exo) peaks for ZNF808 and KAP1 were obtained in HEK293T cells and embryonic stem cells (ESCs) respectively by Imbeault *et al.*[26] and downloaded from GEO (GSE78099) as described in Chapter 4. Note, we used the HEK293T dataset for ZNF808 ChIP-exo peaks as it is the only publicly available KZFP ChIP-exo dataset.

5.2.6 Methylation quantitative trait loci (mQTL) associated with hvCpGs

European blood mQTLs

We used a recent publicly available meta-analysis of blood mQTLs identified in European cohorts by the Genetics of DNA Methylation Consortium (GoDMC)[27] to obtain mQTL summary-level data for *cis* and *trans* mQTLs at CpGs covered on the Illumina450K array. We applied significance thresholds of $p < 1 \times 10^{-8}$ and $p < 1 \times 10^{-14}$ for *cis* and *trans* mQTLs respectively[27].

Gambian blood mQTLs

We used H3A genotype and targeted Methyl-Seq data from 375 ENID individuals (**Figure 5.1**) to examine the association between methylation at the *C8orf31* DMR and i) *trans* mQTLs at *ZNF808* and its promoter region (chr19: 144,120,106 – 144,120,681, hg19 coordinates), referred to as 'ZNF808 mQTLs', and ii) *cis* mQTLs at the intragenic ZNF808 ChIP-exo peak (**Figure 5.2,** red region). SNPs were filtered using plink (v1.90b6.21) based on missingness across individuals (< 5%), Hardy Weinberg equilibrium (HWE) (exact test p-value < 0.001) and minor allele frequency (>1%) and pruned using -indep-pairwise (window = 50 kb, step size = 5 kb, r2 threshold = 0.5) to remove variants in strong linkage disequilibrium (LD). This left 78 *ZNF808*

SNPs and 2 SNPs in the ZNF808 ChIP-exo peak. mQTL analyses were performed using an additive linear model, with SNP genotype coded as allelic dosage as the predictor variable and CpG Beta values as the dependent variable, adjusted for sex. Note, we did not include age as a covariate as all ENID samples were collected at exactly 2 yrs. We did not adjust for any additional confounders such as smoking as incidence of smoking is close to zero in this population. Reported p-values are unadjusted for multiple tests.

5.2.7 Association between the *C8orf31* DMR and glucose metabolism outcomes

The EMPHASIS cohort (**Figure 5.1**) was used to examine the association between methylation at 9 IlluminaEPIC CpGs in the *C8orf31* DMR (**Figure 5.2**) and 7 phenotypes linked to insulin resistance and glucose tolerance collected previously by Antoun *et al*[4]. These phenotypes include insulin and glucose levels measured after fasting ('ins0', 'glu0') and at 30 minutes after oral glucose intake ('ins30', 'glu30'), glucose levels measured 120 minutes after glucose intake ('glu120'), first-phase insulin secretion (measured as the insulinogenic index[28] or 'IGI'), and insulin sensitivity (measured using the HOMA2-S index[29]). Note, ins0 and ins30 insulin measures had been natural log transformed (see Antoun *et al.*[4] for further details). All available measurements had been pre-adjusted for infant age, sex, body mass index (BMI) and height where these associations were significant by Antoun *et al*[4]. For our analysis, the available residuals from this adjustment step were used as outcome variables and CpG Beta values were used as predictor variables. Infant sex, age, sentrix plate and sentrix row were included as covariates in this model, following the approach of Antoun et al [4].

5.2.8 Two-sample Mendelian randomisation (2SMR)

Mendelian Randomisation uses genetic variants as instruments to assess the causal effect of a given exposure on an outcome/phenotype in the presence of unmeasured confounding[30]. Whilst one-sample MR relies on genetic, exposure and outcome data generated from the same sample-set, two-sample MR (2SMR) uses genotype-exposure and genotype-outcome summary statistics obtained in two non-overlapping datasets respectively[30]. We therefore used 2SMR to assess the causal relationship between methylation at the *C8orf31* DMR and phenotypes linked to glucose metabolism. (**Figure 5.6**). 2SMR was performed using the TwoSampleMR R package

(v. 0.5.6) which allows users to implement a variety of 2SMR methods and to access GWAS summary level data curated by MR-Base[31,32]. We used the 7 Illumina450K CpGs located in the *C8orf31* DMR as 'exposures' and mQTLs associated with these CpGs ($p < 1 \times 10^{-8}$ and $p < 1 \times 10^{-14}$ for *cis* and *trans* mQTLs respectively) in the GoDMC meta-GWAS[27] as 'genetic instruments' (**Figure 5.6**).

Summary level data (effect estimates, standard errors, and p-values) for genetic instruments were extracted from GoDMC ('exposure dataset') and from several publicly available European GWAS ('outcome datasets'). In total, we considered 10 traits linked to insulin dynamics and glucose metabolism and selected the GWAS with the largest sample size for each trait: Type 1 diabetes[33], Type 2 diabetes[34], fasting insulin[35], fasting glucose[36], acute insulin response (AIR)[37], insulin secretion rate[37], proinsulin levels[38], homeostasis model assessment of insulin resistance (HOMA-IR)[39], homeostasis model assessment of beta cell function (HOMA-B)[39] and glycated haemoglobin (Hb1Ac)[40] (**Supplementary Table C.1**). At this stage, any genetic instruments missing from these GWAS were replaced by proxy variants in high linkage disequilibrium ($r > 0.8$) in European phase3 1000G reference panel where possible. Exposure and outcome datasets were then harmonised to ensure that the effect alleles in both datasets were aligned. For each exposure-outcome pair, the effect size of exposure (CpG methylation) on the outcome (insulin/type 2 diabetes) was calculated using the Wald Ratio:

beta_outcome/beta_exposure,

where *beta_exposure* is the % increase in methylation per copy of the effect allele, and *beta_outcome* is the unit increase in the outcome per copy of the effect allele. Because each CpG was associated with multiple conditionally independent genetic instruments, the Wald ratios were weighted by the inverse variance of their association with the outcome and combined in fixed-effects meta-analysis using the 'fixed-effects inverse variance weighted' method in the TwoSampleMR package.

We used the MR Steiger method to confirm the direction of the causal relationship between exposure and outcome[41]. We tested for horizontal pleiotropy at genetic instruments using two approaches. One approach implemented the *mr_pleiotropy_test()* function which tests whether the MR-Egger intercept significantly deviates from zero[42] and the other implemented the *mr_heterogeneity()* function which tests whether the individual Wald ratios for each CpG-SNP pair show greater heterogeneity than expected by chance[42,43], indicative of horizontal pleiotropy.

5.2.9 Phenome-wide Two-Sample Mendelian randomisation

To assess whether the *C8orf31* DMR is causally associated with other traits not linked to glucose metabolism, we performed phenome-wide Two-Sample MR using summary level data from GWAS curated by MR-Base[44] and accessed using TwoSampleMR package. We required GWAS studies to have a minimum sample size of 10,000, a minimum of 1x 10^6 SNPs covered in the GWAS and to have been carried out in European populations. Where multiple GWAS studies were available for the same phenotype, we selected the study with the largest sample size. In total, we considered 4,469 traits in which the GoDMC genetic instruments described in **5.2.8** were covered in the GWAS. We performed 2SMR across all exposure-outcome pairs, using the inverse variance weighted fixed-effects method. A Bonferroni-adjusted p-value threshold ($p < 1.12 \times 10^{-5}$), adjusted for the number of unique phenotypes, was used to define significant associations.

5.2.10 Association between the *C8orf31* DMR and thyroid function measurements

Because *C8orf31* is most highly expressed in thyroid, we sought to examine whether methylation at the *C8orf31* DMR is linked to thyroid function. Thyroid function measurements were available for 118 ENID individuals at ages 5-8 years [15] (**Figure 5.1**), allowing us to assess the association between DMR methylation at age 2 and thyroid function at a later time point). Thyroid function measurements included: free T3 (free trio-iodothyronine) and free T4 (free thyroxine), the two main hormones secreted by the thyroid; TSH (pituitary-derived thyrotropin) which is secreted by the pituitary gland and stimulates the thyroid to secrete T4; and TG (thyroglobulin) which is required for the synthesis of free T3 and T4. These data had been collected by group member Dr Toby Candler to test the association between methylation at the putative ME *PAX8* and thyroid function, as described elsewhere[15]. In our study, the 17 *C8orf31* CpGs covered in the ENID targeted Methyl-Seq dataset (**Figure 5.2**) were used as predictors of thyroid function in a linear model that was adjusted for age, sex, and urinary iodine (because iodine concentrations can influence thyroid function[15]). Note, we log-transformed TSH and TG variables as these were non-normally distributed[15].

5.2.11 Association between the *C8orf31* DMR and 1C-metabolism maternal biomarkers

Maternal biomarker data for 350 women were available for the ENID cohort (**Figure 5.1**) and were collected as described elsewhere[15]. These biomarker data included 10 core nutritional biomarkers involved in 1-C metabolism (homocysteine, methionine, cysteine, choline, betaine, dimethylglycine (DMG), vitamin B12, folate, PLP and vitamin B2), and 18 other amino acid biomarkers linked to metabolic pathways that may also influence 1-C metabolism (alpha1-C acid glycoprotein (AGP), alanine, aspartate, arginine, glutamate, glycine, histidine, leucine, isoleucine, lysine, proline, phenylalanine, serine, threonine, tyrosine, uracil, uridine, valine). Briefly, all maternal biomarkers had been preadjusted for maternal BMI, maternal age, inflammation, and gestational age and then back extrapolated to the date of conception using previously described methods[45]. We studied associations between maternal biomarkers and methylation at 345 ENID children covered in the targeted Methyl-Seq dataset. For a given marker, we regressed each CpG beta value against the log-transformed maternal biomarker and adjusted the model for offspring sex.

5.2.12 *C8orf31* gene expression and methylation in thyroid tissue

The GTEx portal (www.gtexportal.org/home/) was used to examine the expression level of *C8orf31* across a panel of 56 post-natal tissues. Because *C8orf31* was most highly expressed in thyroid, we examined the relationship between DMR methylation and gene expression across 50 non-tumour thyroid samples using methylation and RNA-seq data downloaded from the 'The Cancer Genome Atlas Wanderer' web tool (http://maplab.imppc.org/wanderer/) [46]. These data were available for 3 CpGs located in the *C8orf31* DMR.

5.2.13 BLAST analysis

The Basic Local Alignment Search Tool (BLAST) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to search for homologous nucleotide sequences to *C8orf31* with the following parameters: query ID = BC073830.1, programme selection = 'Blastn', word size = 11 nucleotides.

5.2.14 Characterisation of mQTLs

To probe evidence that the ZNF808 mQTLs associated with *C8orf31* DMR methylation may influence the expression of ZNF808, we used the haploR R package[47] to query HaploReg v4.1[48] to explore: i) eQTL data collated from various sources including GTEx pilot analysis v6 and GEUOVADIS[49]; ii) ChromHMM promoter and enhancer states obtained from the 15-state and 25-state model using data from the Roadmap Epigenomics project[50], and iii) regulatory motif alterations obtained using position weight matrices from the ENCODE transcription factor binding experiments[51], at mQTLs and LD-linked SNPs (r2 > 0.8 in 1000 Genomes Phase 1 European ('EUR')). Full details on the data included in HaploReg v4.1 are described elsewhere[48].

RegulomeDB v2.0.3[52] was used as an additional tool to interpret the likelihood that a given variant has regulatory function. RegulomeDB uses many of the same ENCODE-based sources as HaploReg and integrates them with scores from DeepSEA (which uses a machine learning algorithm to predict chromatin effects of a genetic variant[53]) into a random forest model that generates a probability score between 0 and 1. SNPs with RegulomeDB scores closer to 1 are more likely to be regulatory variants[54].

5.3 Results

5.3.1 A DMR upstream of *C8orf31* shows high methylation variation and is associated with fasting insulin levels and insulin secretion in Gambian children

In Chapter 4, we identified 5 hvCpGs that were associated with ZNF808 mQTLs in GoDMC (**Figure 4.4C**). 4 of these hvCpGs fall into a single cluster located in immediately upstream of *C8orf31* (**Figure 5.2**). This hvCpG cluster overlaps a differentially methylated region (DMR) associated with fasting insulin levels in the EMPHASIS cohort of Gambian children (8-9 yrs) previously identified by Antoun *et al.*[4] (**Figure 5.2**).

There are 4 other Illumina450K CpGs located in the DMR (**Figure 5.2**). These CpGs had narrowly missed our thresholds for defining hvCpGs in Chapter 3, but all are amongst the top 5% of variable probes in at least 9 of the datasets we used in Chapter 3 (**Supplementary Table C.3**), suggesting that the DMR is consistently variable across tissue types. Notably, one hvCpG located in the DMR is a previously identified systemic interindividual variation ('SIV') CpG (cg07770222)[17] and another immediately downstream of the DMR is an epigenetic supersimilarity ('ESS') CpG (cg04612566)[17] (**Figure 5.2**). This supports the notion that methylation variation at the DMR is linked to the pre-gastrulation embryo.

We re-examined the same cohort of Gambian children used by Antoun *et al.*[4] ('EMPHASIS', see Methods) to determine the direction of the association between DMR methylation and fasting insulin levels. A 1 % increase in methylation was associated with a median increase in fasting blood insulin levels of 0.08 pmol/l (IQR = [0.19, 0.05]) (p-values < 8.43×10^{-3}, **Figure 5.3A**, **Supplementary Table C.2**).

We also tested the association between DMR methylation and 6 other risk factors relating to glucose tolerance and insulin resistance (**Figure C.2**). There was evidence of a previously unreported association insulinogenic index (IGI), a measure of first-phase insulin secretion in response to glucose[28]. A 1% increase in methylation was associated with a median unit decrease in IGI of 0.037 (IQR = [0.030,0.046]) (p-values < 9.05×10^{-3}, **Figure 5.3B, Supplementary Table C.2**).

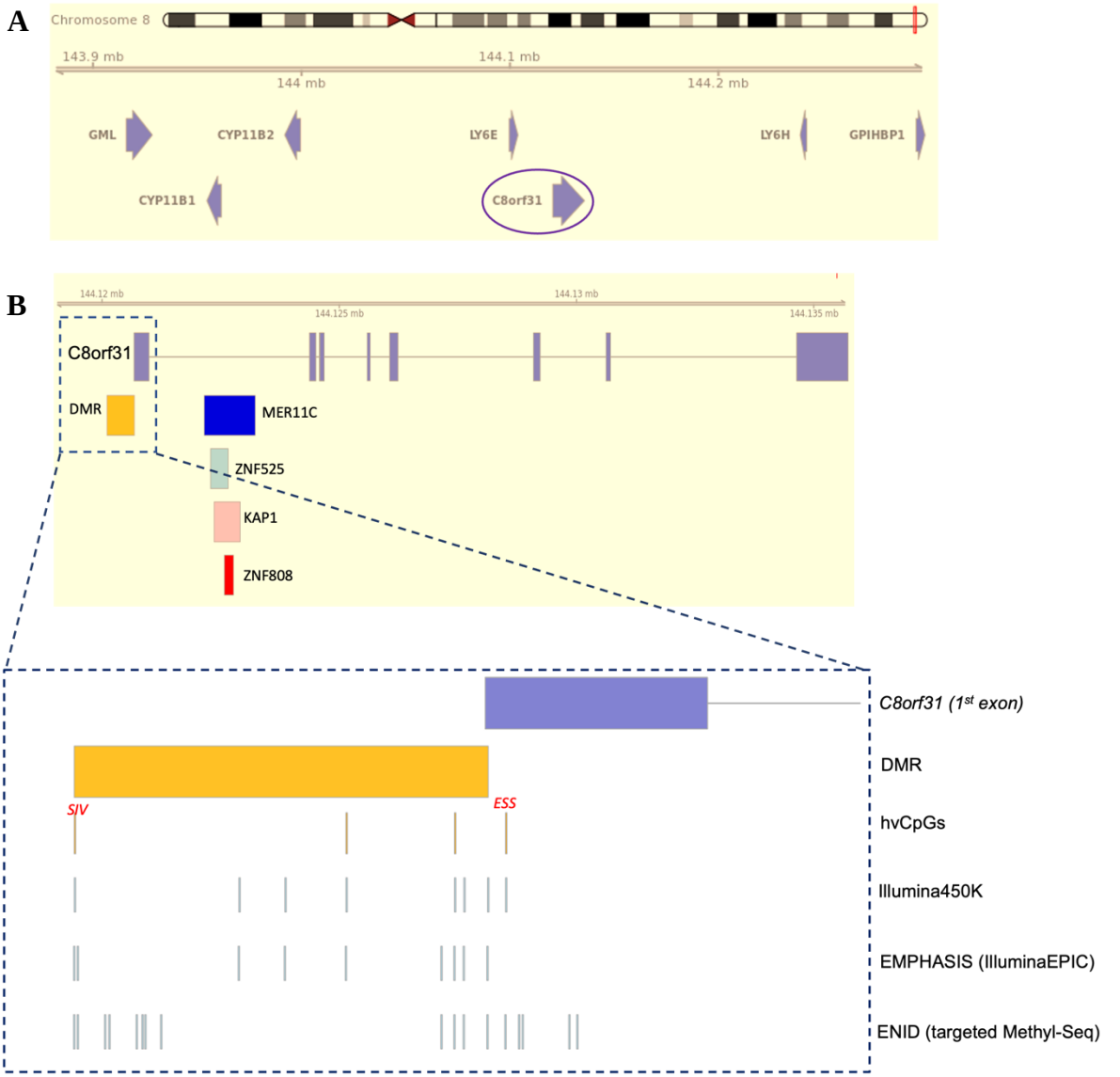


Figure 5.2. Schematic of *C8orf31* gene region and key features. A) Genomic location of *C8orf31* (chr8:144120679-144135720, hg19 coordinates) and surrounding genes. **B)** Top panel shows the *C8orf31* gene region with exons highlighted in purple and the fasting insulin DMR identified by Antoun *et al.* (chr8:144120106 – 144120681, hg19 coordinates) highlighted in orange. Also shown is the intragenic MER11C element (navy) which overlaps ChIP-exo peaks identified by Imbeault *et al.*[26] for ZNF808 (red), ZNF525 (blue) and KAP1 (pink). Bottom panel is zoomed-in view of the region indicated by the dashed rectangle in top panel, showing the positions of the hvCpGs and the Illumina450K array background relative to the DMR. Bottom two tracks show the CpGs covered by the IlluminaEPIC dataset (EMPHASIS cohort) of the targeted Methyl-Seq dataset (ENID cohort) (see Figure 5.1). Red text indicates the hvCpGs that are putative ME CpGs: one systemic interindividual variation (SIV) CpG (cg077702222), and one epigenetic supersimilarity (ESS) CpG (cg04612566).

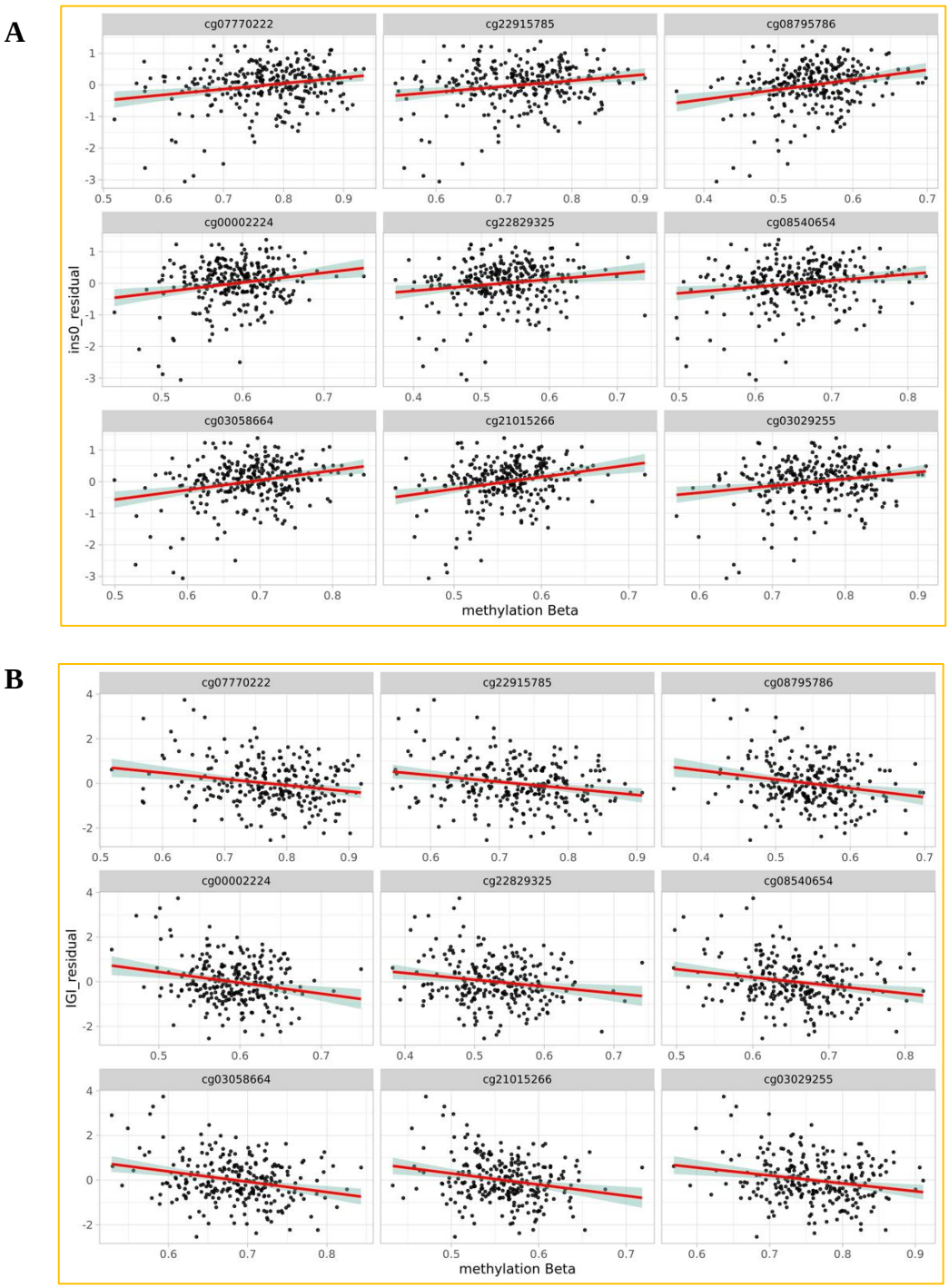


Figure 5.3. The DMR upstream of *C8orf31* is associated with fasting insulin and insulin secretion in the EMPHASIS cohort. A) Scatter plot showing the relationship between fasting insulin levels ('ins_0', y-axis) and methylation at each IlluminaEPIC CpG within the DMR, measured in the EMPHASIS cohort of Gambian children (n = 286). **B)** Scatter plot showing the relationship between insulinogenic index ('IGI') and methylation at each IlluminaEPIC CpG within the DMR, measured in the EMPHASIS cohort of Gambian children (n = 256_). Note, the y-axis gives the residuals from a linear regression model in which fasting insulin levels and IGI were adjusted for age, sex and BMI (see 'Methods'). The regression line (y ~ x) is in red and blue shaded regions are 95% confidence intervals.

5.3.2 The *C8orf31* DMR is associated with ZNF808 genetic variation in Gambian children

ZNF808 silences MER11C elements during pancreas development[2]. Accordingly, *C8orf31* contains a MER11C element that is located < 1500 bp downstream of the DMR (**Figure 5.2,** blue region). This MER11C element contains a ChIP-exo peak for ZNF808 identified in HEK293T cells and a peak for KAP1 identified in ESCs (**Figure 5.2,** red region). To probe ZNF808-mediated silencing of MER11C as a mechanism influencing methylation at the *C8orf31* DMR, we examined the association between the DMR and genetic variation at: i) the *ZNF808* gene region (including its promoter) *in trans* and ii) the ZNF808 binding site inside the MER11C element *in cis* (**Figure 5.4**).

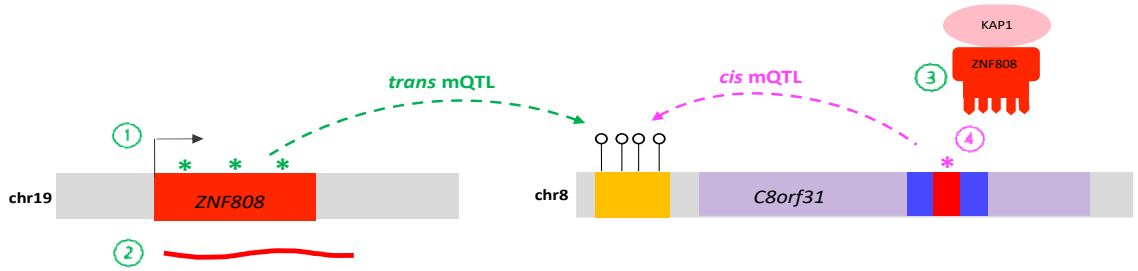


Figure 5.4. Rationale for testing the association between the *C8orf31* DMR and genetic variation at *ZNF808 in trans* and at the ZNF808 binding site *in cis* .ZNF808 trans mQTLs (green asterisks) may influence ZNF808 transcription (1), mRNA processing/stability/splicing (2) and/or protein function of ZNF808 (3). mQTLs located in the ZNF808 binding site in MER11C (pink asterisk) may influence the binding affinity of *ZNF808* (4) and consequently alter KAP1 recruitment. Because KAP1 recruits DNA methyltransferases, altered KAP1 recruitment may affect methylation at the DMR upstream of *C8orf31* (orange).

We explored these associations using the publicly available GoDMC mQTL meta-analysis in European blood[27] (see Methods) to maximise power to detect mQTL. We also explored these mQTLs using targeted Methyl-Seq data and imputed H3Africa array data from a cohort of Gambian 2-year-olds (ENID cohort, see 'Methods') because of the association between DMR methylation and insulin-related phenotypes observed Gambian children (**Figure 5.3**).

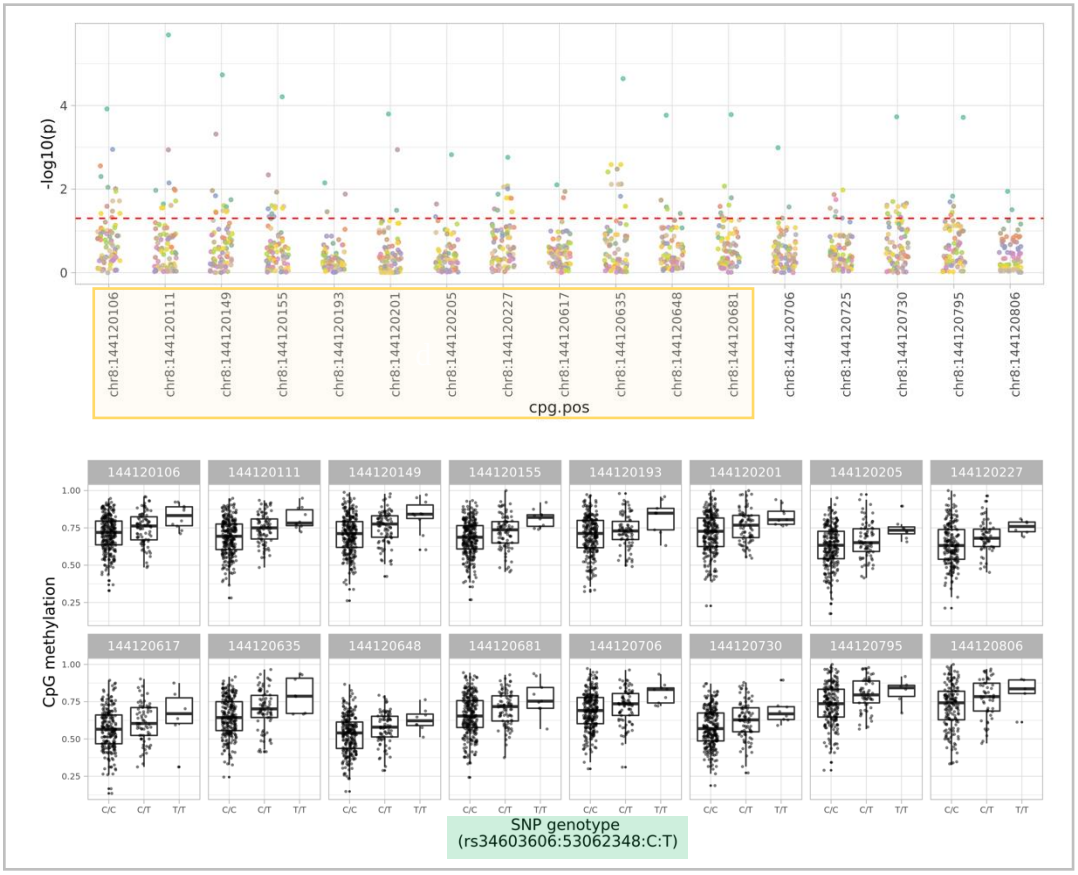
In the GoDMC meta-analysis, 5/7 Illumina450K CpGs located in the DMR were associated with a *ZNF808* mQTL *in trans* (**Supplementary Table C.4**). In the ENID cohort, we tested

the association between 78 *ZNF808* SNPs and 17 *C8orf31* CpGs (14 of which fall into the DMR). There were 128 nominal associations (p-value < 0.05) between 35 SNPs and all 17 CpGs. The top hit mQTL (rs34603606) showed evidence of a nominal association with 16/17 *C8orf31* CpGs. The median DNAm increase per copy of the effect allele (T) across the 16 CpGs was 5.0% (IQR = [4.6%, 5.3%]) (**Figure 5.5A**, bottom).

All 7 DMR CpGs were associated with at least one *cis* mQTL in the GoDMC meta-analysis (**Supplementary Table C.3**), but none of these variants were in the ZNF808 ChIP-exo peak within the MER11C element. In the ENID cohort, 2 SNPs were located in the ZNF808 ChIP-exo peak. We tested the association between these 2 SNPs and the 17 *C8orf31* CpGs. One SNP (rs6558371) showed evidence of an association with 16 CpG*s* ($p < 2.2 \times 10^{-5}$) (**Figure 5.5B**, top). The median DNAm decrease per copy of the effect allele (T) across the 16 CpGs was 7.9.% (IQR = [6.9%, 9.0%]) (**Figure 5.5B**, bottom). The other SNP showed evidence of an association ($p = 1.2 \times 10^{-12}$) with only one *C8orf31* CpG (**Figure 5.5B**, top).

Other KZFPs, such as ZNF578, ZNF525, ZNF440 and ZNF468, are also known to target MER11 (including MER11A, MER11B, MER11C) retrotransposons[26]. No DMR CpGs were associated with mQTLs located in these KZFP genes in the GoDMC meta-analysis, except for 2 CpGs that were associated with ZNF578 mQTLs (**Supplementary Table C.3**). In the ENID cohort, we had genotype data available for 115 *ZNF578* SNPs, 24 *ZNF525* SNPs, 27 *ZNF440* SNPs and 52 *ZNF568* SNPs. We tested the association between each of these SNPs and the 17 *C8orf31* CpGs, and observed several nominal SNP-CpG associations (p < 0.05) for each of the KZFP genes (*ZNF578*: 177 associations between 54 SNPs and 17 CpGs; *ZNF525*: 4 associations between 4 SNPs and 4 CpGs; *ZNF440*: 25 associations between 8 SNPs and 11 CpGs; *ZNF468*: 37 associations between 21 SNPs and 12 CpGs (see **Figure C.3**)). This raises the possibility that other KZFPs, in addition to ZNF808, may influence methylation at the *C8orf31* DMR.

A


-log10(p)
cpg.pos
chr8:144120106
chr8:144120111
chr8:144120149
chr8:144120155
chr8:144120193
chr8:144120201
chr8:144120205
chr8:144120227
chr8:144120617
chr8:144120635
chr8:144120648
chr8:144120681
chr8:144120706
chr8:144120725
chr8:144120730
chr8:144120795
chr8:144120806
CpG methylation
SNP genotype
(rs34603606:53062348:C:T)

B

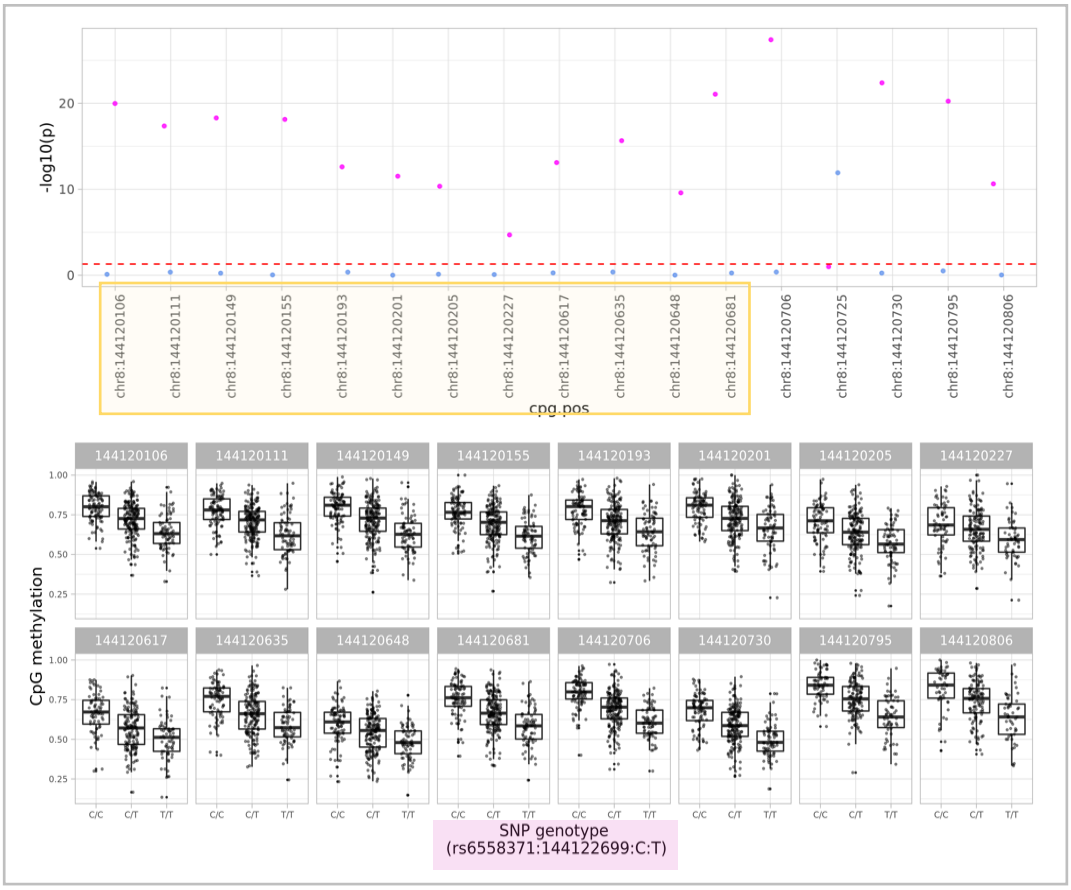

-log10(p)
cpg.pos
chr8:144120106
chr8:144120111
chr8:144120149
chr8:144120155
chr8:144120193
chr8:144120201
chr8:144120205
chr8:144120227
chr8:144120617
chr8:144120635
chr8:144120648
chr8:144120681
chr8:144120706
chr8:144120725
chr8:144120730
chr8:144120795
chr8:144120806
CpG methylation
SNP genotype
(rs6558371:144122699:C:T)

Figure 5.5. Association between *C8orf31* DMR and genotype at ZNF808 gene and at ZNF808 binding ChIP-exo peak in MER11C. A) Top: Manhattan plot of p-values from a linear model testing the association between methylation at 17 CpGs covered in the targeted Methyl-Seq dataset and genotype at 78 *ZNF808* SNPs covered in the imputed H3Africa array data across 375 ENID individuals. **Bottom:** Boxplots showing the distribution of methylation values across individuals stratified by genotype at rs434603606 (the top hit mQTL in the top panel) for significantly associated CpGs. **B) Top**: Manhattan plot of p-values from a linear model testing the association between methylation at 17 *C8orf31* CpGs and genotype at 2 SNPs located in the ZNF808 ChIP-exo peak across 375 ENID individuals. **Bottom:** Boxplots showing the distribution of methylation values across individuals stratified by genotype at rs6558371 (the top hit mQTL in the top panel) for nominally associated CpGs. In the top panels of A and B, the x-axis gives CpG position, y-axis gives p-value and point colour gives the SNP. The red line indicates a p-value threshold < 0.05. CpGs in yellow box are located in the DMR.

5.3.3 Characterising ZNF808 mQTLs and the ZNF808 binding site mQTL

To probe whether ZNF80*8* mQTLs influence methylation at the *C8orf31* DMR by altering expression and/or protein function of *ZNF808*, we functionally annotated the 5 GoDMC mQTLs and their LD-linked SNPs (r2 > 0.8 in EUR phase 1 reference) using HaploReg v. 4.1 [48] and RegulomeDB[54,55]. 93 SNPs were LD-linked to at least one mQTL in the HaploReg database (**Figure C.4A**), giving 98 variants in total, including the ZNF808 mQTLs we identified in the ENID cohort. These variants largely fall into non-transcribed regions of *ZNF808* (**Figure C.4B**). One SNP (rs61746334) is a missense variant (G>A, Ser437Asn) located in the C2H2-type zinc finger of ZNF808 but is not located in a known DNA binding residue (personal correspondence, Dr Michael Imbeault) so its effect on DNA binding affinity is difficult to interpret. 96/98 variants overlapped an eQTL for *ZNF808* in at least one GTEx tissue type, most commonly in tibial nerve (**Figure C.4C**), indicating that variants associated with DNA methylation at the DMR upstream of *C8orf31* in *trans* are also associated with RNA transcript levels of *ZNF808* in post-natal tissues. We also note that 18/98 variants are in predicted promoter regions in many cell types, including embryonic stem cells (**Figure C.5**), raising the possibility that they may also influence *ZNF80*8 expression in early development. Finally, 34% of variants had a RegulomeDB score above 0.5 (**Figure C.6**), supporting the notion that they have regulatory function.

The mQTL (rs6558371) located in the ZNF808 ChIP-exo peak within MER11C element is an eQTL for *C8orf31* in thyroid. However, its effect on the binding affinity for ZNF808 is unclear as it is located 10 bp downstream of the consensus binding motif for ZNF808 (personal

correspondence, Dr Michael Imbeault). This variant is however predicted to alter the binding affinity of Early-B cell factor (EBF) (**Figure C.7**), a transcription factor that regulates the specification of B lymphocytes[56].

5.3.4 The *C8orf31* DMR shows evidence of a causal effect on beta cell function

Given evidence that the DMR upstream of *C8orf31*: i) is associated with genetic variation at *ZNF808,* a gene that is required for pancreatic cell fate specification[2], ii) is associated with fasting insulin levels in Gambian children[4], and iii) includes a CpG (cg04612566) that is differentially methylated between MZ twins discordant for T2 diabetes[57], we explored the hypothesis that methylation at DMR causally influences phenotypes linked to diabetes pathogenesis using a Two-Sample Mendelian Randomization (MR) framework (**Figure 5.6**). This analysis was further motivated by the observation that GWAS traits linked to *C8orf31* include total cholesterol, low density lipoprotein cholesterol, body weight, body mass index (BMI), and hyperglycaemia[58] (**Supplementary Table C.5, Figure 5.6**). We used *cis* and *trans* GoDMC mQTLs as genetic instruments for 7 Illumina450K CpGs located in the DMR (**Supplementary Table C.6**) and extracted summary level data for these SNPs in GWAS for T1 diabetes, T2 diabetes, and 8 other glycemic traits (**Figure 5.6**, **Supplementary Table C.1**). MR assumes that these genetic instruments: i) are strongly associated with the exposure, ii) are not associated with confounders and iii) do not show horizontal pleiotropy, i.e., they are not causally associated with the outcome via an alternative exposure (**Figure 5.6**).

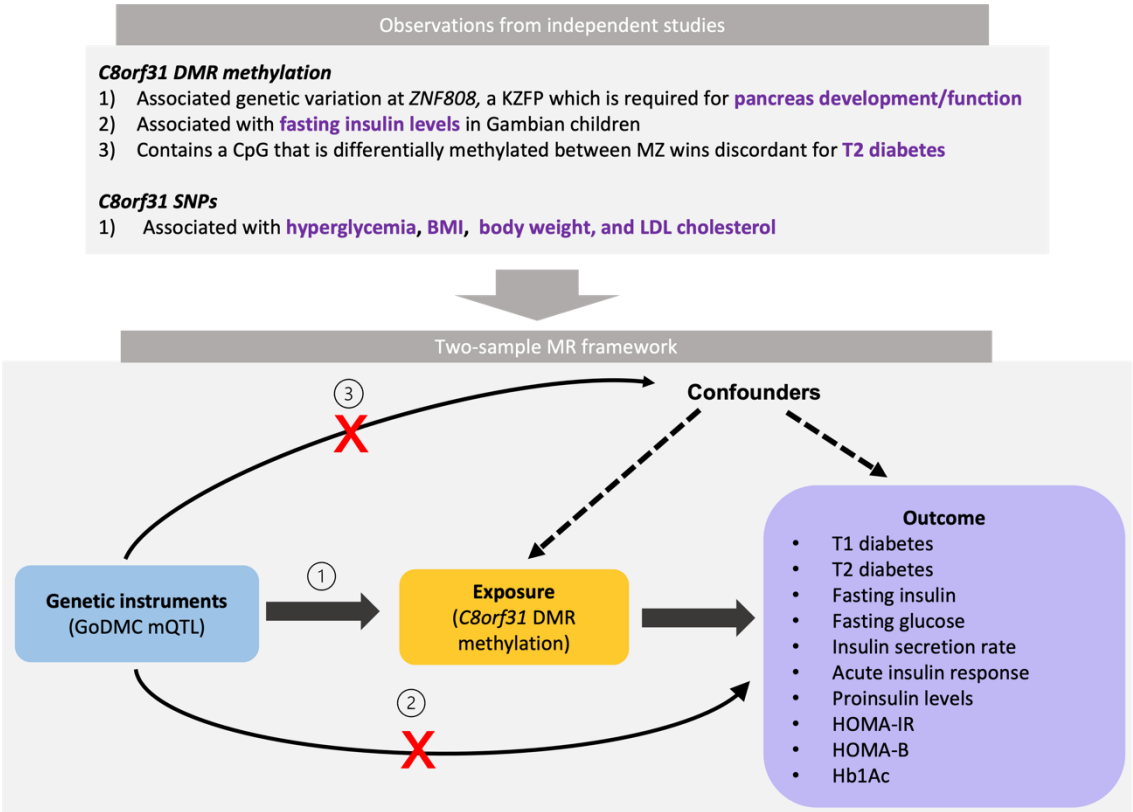


Figure 5.6. Two-sample Mendelian randomisation (2SMR) framework exploring the causal link between DMR methylation and phenotypes linked to glucose metabolism. Top panel highlights the observations from previous studies that motivated our 2SMR analysis. Bottom panel shows 2SMR framework. Illumina450k CpGs located in the DMR were used as 'exposures' and proxied using GoDMC mQTLs ('genetic instruments', Supplementary Table C.7). 10 phenotypes linked to diabetes pathogenesis were used as 'outcomes' (Supplementary Table C.1). 2SMR assumes that genetic instruments are 1) robustly associated with the exposure, 2) are only associated with the outcome via the exposure, and 3) are not associated with confounders.

We found evidence of a causal link ($p < 0.05$) between DMR methylation at all 7 CpGs and odds of Type 1 diabetes, 6 CpGs and insulin secretion rate (ISR), and 3 CpGs and acute insulin response (AIR) (**Figure 5.7**, **Supplementary Table C.7**). Evidence for a causal link remained when additionally considering GWAS traits where ISR had been adjusted for BMI and AIR had been adjusted for BMI and/or insulin sensitivity (**Figure 5.7**). We did not adjust these p-values for multiple-testing given that these GWAS traits are not independent.

An MR Steiger test, which compares the variance in the outcome explained by the genetic instruments to the variance in the exposure explained by genetic instruments[41], confirmed that the direction of the causal association was from methylation to the outcomes (**Supplementary Table C.8**). We found no evidence of a causal association with the other traits (**Figure 5.7**).

We probed the MR assumption that genetic instruments do not show horizontal pleiotropy (**Figure 5.6**) by examining MR-Egger intercept terms[59], where the intercept is interpreted as the pleiotropic effect of a genetic variant[59]. We found no evidence that intercepts significantly deviated from zero for any exposure-outcome pairs that we were able to test (**Supplementary Table C.7**). Heterogeneity in MR effect sizes between genetic instruments may also be indicative of horizontal pleiotropy[42,43]. We found no evidence of Wald ratio heterogeneity between genetic instruments proxying exposures associated with ISR and AIR but did find nominal evidence of heterogeneity at genetic instruments that proxied 2/7 DMR CpGs associated with T1 diabetes (Cochran's Q p-value = 0.01 and 0.038 for cg22829325 and cg03058664 respectively) (**Supplementary Table C.7**).

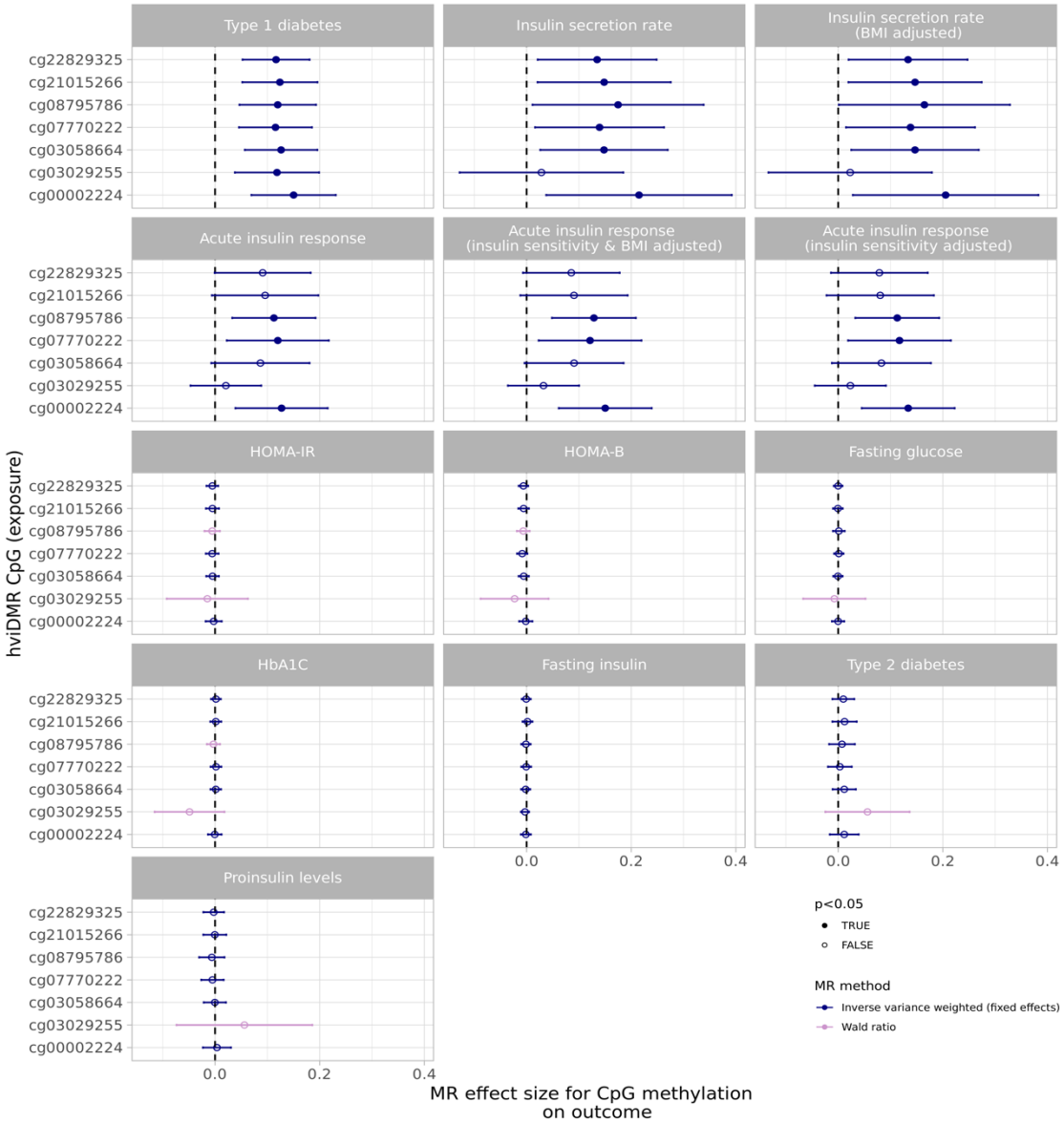


Figure 5.7. Forest plot showing Two-Sample MR effect ('2SMR') estimates for CpGs in the *C8orf31* DMR on diabetes-related outcomes. 7 Illumina450K CpGs located in the DMR were used as exposures (y-axis) and proxied by GoDMC *cis* and *trans* mQTLs to examine causal effect on 10 phenotypes ('outcomes'). The mean causal effect of each CpG on a given outcome is shown by round points, and horizontal lines are 95% confidence intervals. Filled points indicate p-value < 0.05 and are coloured by the 2SMR method. The Wald ratio method (purple) was used when only one genetic instrument was associated with the exposure, otherwise the inverse variance weighted fixed effects method (blue) was used.

5.3.5 Phenome-wide Two-Sample MR with *C8orf31* DMR CpGs

To probe other phenotypes causally associated with *C8orf31* DMR methylation without having made any *a priori* hypotheses about links to diabetes and glucose metabolism, we performed a phenome-wide 2SMR analysis that considered 4,469 traits covered in at least one GWAS study curated by MR-Base[44] (see 'Methods'). At a Bonferroni-corrected p-value threshold $p < 9.66 \times 10^{-6}$ (correcting for the number of unique traits) and requiring all 7 CpGs in the DMR to meet this threshold, we observed evidence of a causal relationship between DMR methylation and asthma onset (**Figure C.8**). However, there was strong evidence of heterogeneity in genetic instruments (with a TwoSample MR Cochran's p-value output of 0 for all exposures) indicating that these genetic instruments show horizontal pleiotropy, therefore violating a key MR assumption (**Figure 5.6**). This analysis was subject to a high multiple testing burden meaning that we cannot rule out a causal effect of *C8orf31* methylation on other traits.

5.3.6 Association between offspring *C8orf31* DMR methylation and periconceptional levels of maternal 1C metabolism biomarkers

Previous studies by our group have found that methylation states at several putative MEs measured in offspring are associated with season-of-conception (SoC) and with maternal periconceptional 1-C metabolism markers in a rural Gambian population[16,25,45], indicating sensitivity to the early developmental environment. We found no overlap between DMR CpGs and a previously identified set of 768 SoC CpGs using the IlluminaEPIC array[25]. We next examined the association between *C8orf31* DMR methylation (17 CpGs) in ENID individuals and a panel of 10 'core' and 18 'extended' 1-C biomarkers (see 'Methods') measured in maternal blood for 345 mother-child pairs. 13 biomarkers were nominally associated ($p < 0.05$) with at least 1 CpG (**Figure 5.8**).

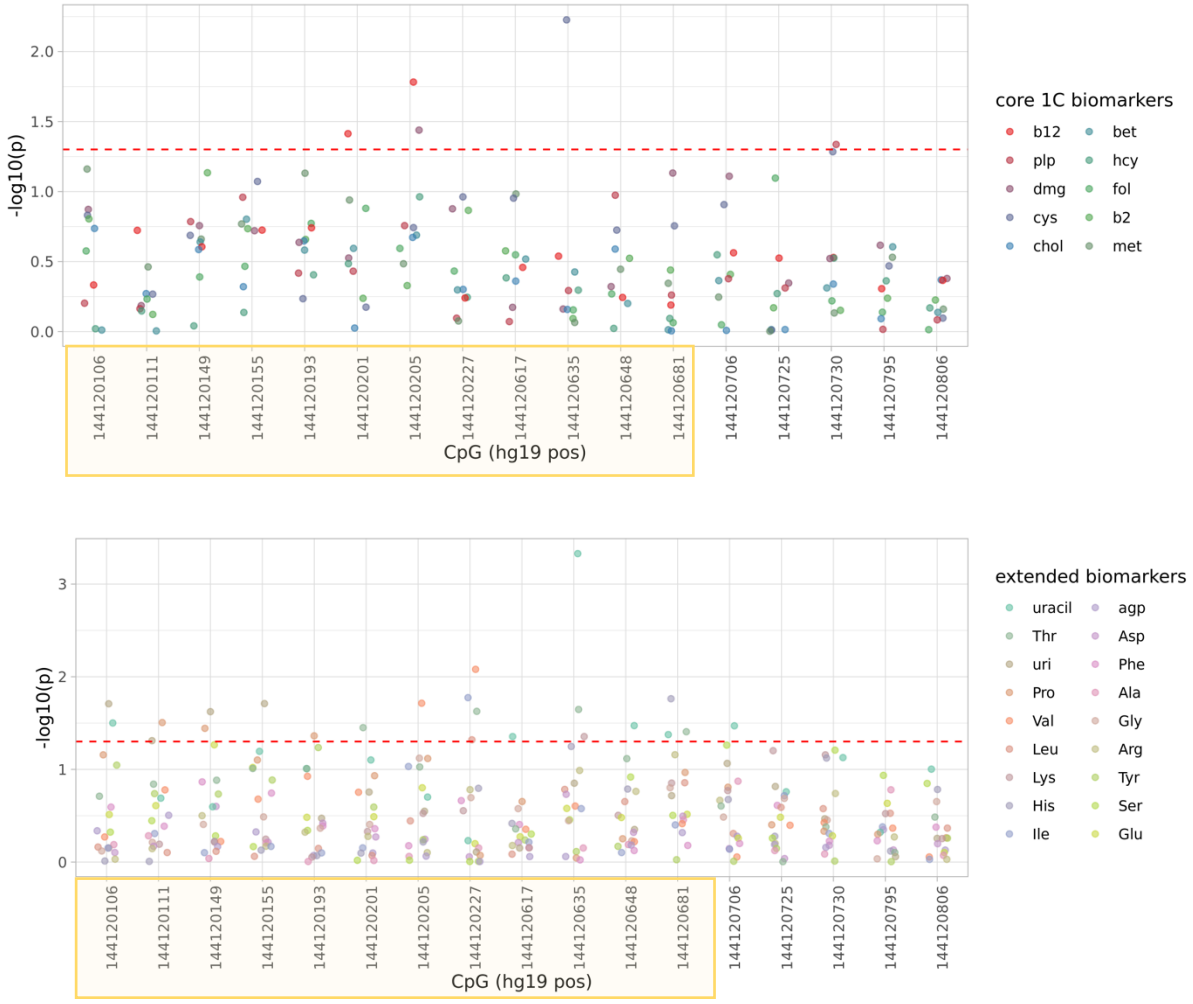


Figure 5.7. Testing the association between the *C8orf31* DMR and maternal 1C metabolism biomarkers in the ENID cohort. Scatter plot showing the strength of the association between each CpG covered in the targeted Methyl-Seq dataset and a panel of core 1-C metabolism biomarkers (top) and amino acids ('extended biomarkers', bottom). Horizontal dashed line indicates nominal associations with p-value < 0.05. The red line indicates a p-value threshold < 0.05. CpGs located in the DMR are highlighted in yellow box.

5.3.7 *C8orf31* expression is highest in thyroid and influenced by methylation

To further characterise *C8orf31,* we examined *C8orf31* expression levels in 54 tissues with RNA-seq data available from the GTEx portal (https://www.gtexportal.org/home/)[60]. *C8orf31* was broadly expressed across many tissues, with expression highest in thyroid and lowest in pancreas and whole blood (**Figure 5.9A**). We also examined gene expression data available for 1019 cancer cell lines in the Cancer Cell Line Encyclopedia (CCLE)[61]. *C8orf31* had highest expression levels in BT-20 (invasive ductal carcinoma), HCC1937 (breast ductal adenocarcinoma) and HCC1187 (breast ductal adenocarcinoma) cell lines (**Figure C.9**).

Given that *C8orf31* is expressed in thyroid and that thyroid function has been linked to glucose homeostasis[62], we next analysed methylation and RNA-seq data available for 56 non-cancerous thyroid samples in the TCGA database[63] for 3 DMR CpGs covered in this dataset. CpG methylation was negatively correlated with gene expression for 1 CpG (cg077702222) (Spearman's rho = -0.35, p-value = 0.014) (**Figure 5.9B**), in agreement with the canonical model in which promoter DNA methylation represses gene expression. This CpG is the SIV CpG (**Figure 5.2**), supporting the notion that methylation states established in early development can influence gene expression in post-natal tissues, on the assumption that methylation is stable in thyroid.

We next explored the association between *C8orf31* methylation and measures of thyroid function (TSH, TG, free T4 and free T3). These had been measured in children from the ENID cohort in a previous study[15] (see 'Methods'). We found no evidence of an association between *C8orf31* methylation and any of these measures.

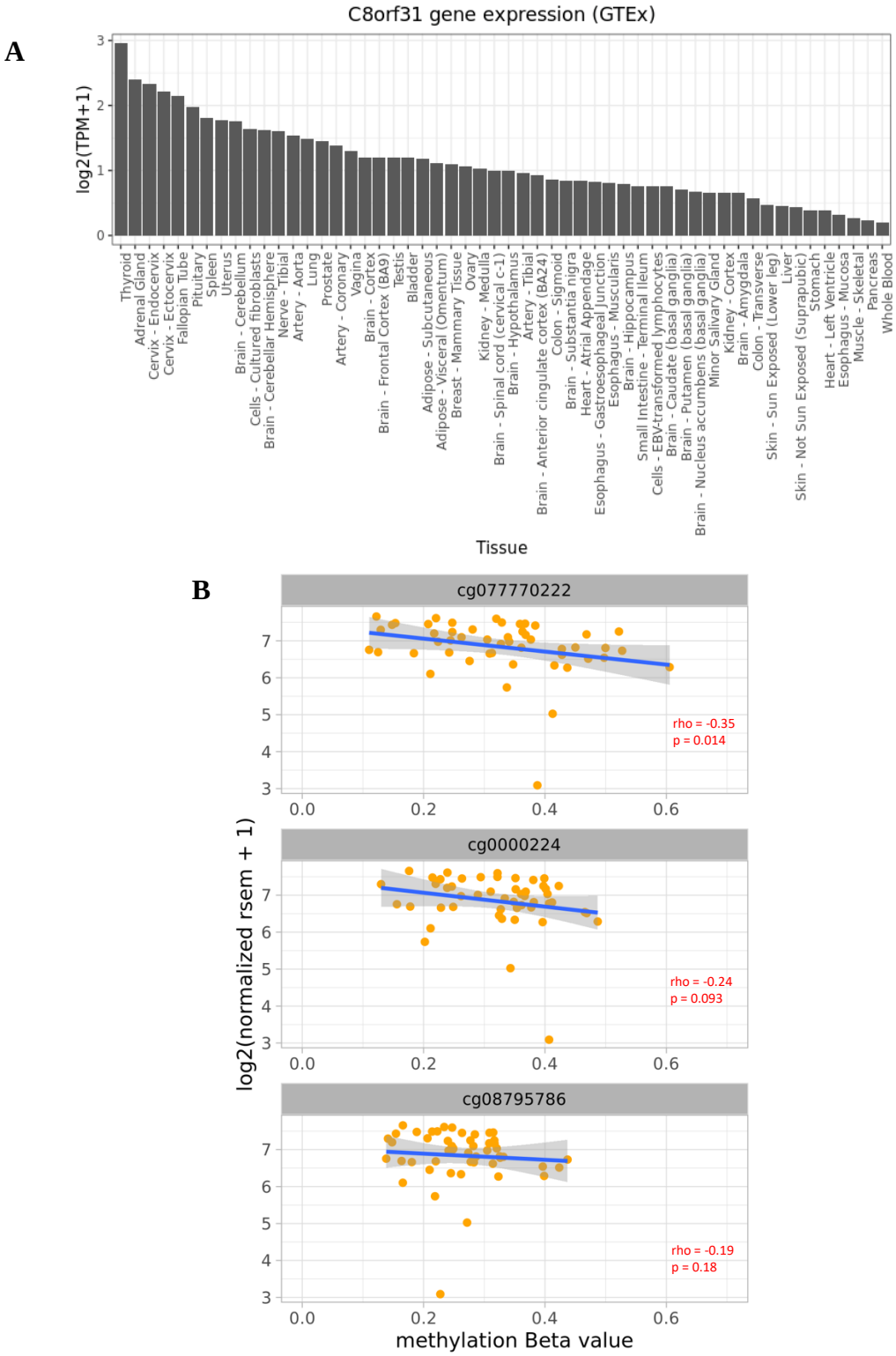


Figure 5.8. Expression and methylation of *C8orf31* in thyroid. A) Gene expression of *C8orf31* in GTEx tissues. TPM = transcripts per million. **B)** Scatter plot of methylation at 3 CpGs located in the *C8orf31* DMR against gene expression across 56 non-cancerous thyroid samples from The Cancer Genome Atlas (TCGA). Plots were generated in the TCGA wanderer web tool.

5.3.9 BLASTn analysis of *C8orf31*

To further probe the function of *C8orf31* we performed a BLASTn analysis to search for genes with similar nucleotide sequences with known function. The top hits were for uncharacterised *C8orf31* homologs in gorilla and chimpanzee (with % sequence identity > 97% and E value = 0.00), and all hits were found in primates (**Figure C.10**), indicating that *C8orf31* is a primate-specific gene. Notably, *C8orf31* shares ~83% sequence identity (E value = 9 x 10^{-77}, sequence coverage = 35%) with the lymphocyte antigen 6 member S (Ly6S), which is transcribed from the complementary strand of *C8orf31* in the opposite direction[64] (**Figure 5.10**). The two genes have overlapping exons. *Ly6S* is a newly identified gene which, at the time of writing is not yet annotated in the UCSC database. Notably, *Ly6S* encodes a membrane-linked protein whose expression is inducible by type 1 interferon (IFN) and is involved in inflammatory and immune-related responses as well as viral resistance[64.], which is notable given the link between *C8orf31* methylation and T1 diabetes.

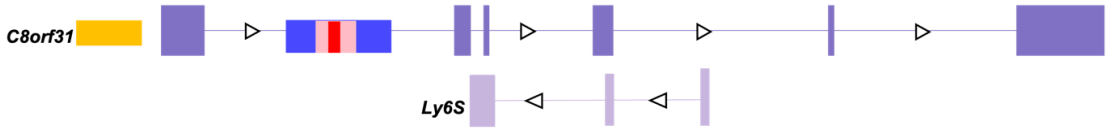


Figure 5.9. Schematic of *C8orf31* and *Ly6S* genes. *Lys6S* is located on the complementary strand to *C8orf31* and is transcribed in the opposite direction. Exons are in purple, the DMR is in orange, and positions of the MER11C element, KAP1 binding site and ZNF808 binding site are in blue, pink, and red respectively. This diagram is based on information given in supplementary materials of Shmerling *et al.*[62].

5.4 Discussion

We have identified a DMR upstream of *C8orf31* that shows evidence of establishment in the early embryo and is linked to ZNF808 targeting of an intragenic MER11C retrotransposon. The DMR is associated with markers of beta cell function and T1 diabetes, positioning KZFP-mediated silencing of TEs in early development as a mechanism that may have phenotypic consequences in later life (**Figure 5.11**).

Several lines of evidence indicate that the DMR is linked to ZNF808-mediated silencing of the retrotransposon MER11C. Firstly, *C8orf31* includes an intragenic MER11C element that overlaps a ZNF808 binding site and a KAP1 binding site. The KAP1 ChIP-exo peak was obtained in embryonic stem cells[26], supporting the notion that this MER11C element is targeted by epigenetic silencing machinery in early development. Secondly, the DMR is robustly associated with ZNF808 genetic variation *in trans* in both European and Gambian populations. These mQTLs overlap eQTLs for *ZNF808* and are in LD with variants located in predicted *ZNF808* promoter regions in many cell types, including embryonic stem cells, suggesting that these mQTLs may have a regulatory function. Thirdly, one mQTL located in the ZNF808 ChIP-exo peak nested in MER11C showed a strong negative association with DMR methylation. Although this mQTL was not located in the consensus ZNF808 binding site, the effect allele was predicted to increase binding affinity for Early B Cell Factor (EBF), raising the possibility that ZNF808 and EBF competitively bind at this region.

De Franco *et al.*[2] found that ZNF808-mediated silencing of MER11C is critical for pancreas development/function. Notably, *C8orf31* was amongst the genes that were upregulated during pancreatic differentiation upon functional inactivation of *ZNF808* in embryonic stem cells (personal correspondence, Dr Michael Imbeault). This supports the notion that ZNF808 may epigenetically silence *C8orf31* by binding to its intragenic MER11C element and may explain our observation that *C8orf31* shows low expression levels in the pancreas.

Methylation states at TEs have been reported to spread to flanking regions[65] . Whilst we were unable to interrogate the methylation status at the MER11C element and flanking regions due to lack of coverage in our Illumina array data (450K and EPIC) and target capture MethylSeq data, the DMR is located < 2kb upstream of the MER11C element, meaning that methylation

spreading across this region is plausible. Examination of the DNAm states in MER11C using whole genome bisulphite sequencing data or nanopore sequencing (which has better coverage of repetitive regions[66]) would allow us to examine methylation states across the whole *C8orf31* gene and promoter region.

An increase in methylation at the DMR was found to be positively associated with fasting insulin levels and negatively associated with insulinogenic index (IGI) (a measure of first-phase insulin secretion in response to glucose) in Gambian children. The latter association was not reported by Antoun *et al*[4]., which we speculate is likely due to multiple testing thresholds applied in their study which examined genome-wide associations. High fasting insulin levels are a marker of insulin resistance[67], and a low insulinogenic index is linked to reduced beta cell function[68], both of which contribute to T2 diabetes pathogenesis[69]. Although we found no evidence of a causal association with T2 diabetes in our 2SMR analysis in European adults, we note that hypermethylation of one CpG in the DMR (cg04612566) measured in adipose tissue has been previously associated with T2 diabetes in an MZ twin study[6]. Interestingly, genetic variation at a SNP intronic to *ZNF808* (rs146957339) is associated with T2 diabetes in African Americans[70]. This SNP was not covered in our H3A genotype data, so we were unable to examine its association with methylation at the *C8orf31* DMR.

Our 2SMR analysis revealed evidence of a casual effect of increased methylation on acute insulin response (AIR) and insulin secretion rate (ISR) in response to glucose in Europeans. The direction of this association seemingly contradicts the negative association between methylation and IGI found in our Gambian cohort. We note that AIR and ISR were measured using the intravenous glucose tolerance test whilst ISR was measured using the oral glucose tolerance test, meaning that these measurements may not be directly comparable. Alternatively, an increase in ISR and AIR observed in adults might reflect a compensatory increase in insulin secretion in response to chronic insulin resistance[71].

We also found evidence of a causal association between several DMR CpGs and T1 diabetes risk, suggesting that methylation at this region may play a role in the autoimmune destruction of pancreatic beta cells. The complementary strand of *C8orf31* encodes a novel protein, Ly6S, whose expression is linked to inflammatory and immune response as well as virus resistance[64]. Amongst the genes regulated by Ly6S expression is NF-kappa-B[64], which is known to play a

critical role in T1 diabetes pathogenesis by activating autoreactive T cells[72]. Further work is needed to examine the relationship between *C8orf31* and *LY6S* expression, and the functional roles of these two genes in relation to insulin dynamics and diabetes pathogenesis. *LY6S* may be downregulated by transcriptional silencing mediated by the *C8orf31* long non-coding RNA[73]. In this scenario, ZNF808-mediated silencing of MER11C would also influence LY6S expression. *In* vitro experiments in which *C8orf31* and Ly6S are inactivated followed by RNA sequencing will improve our understanding of the function of these genes.

The overlap between the DMR CpGs and putative MEs supports the notion that methylation variation at *C8orf31* is established in the early embryo before gastrulation. We found no evidence of an association with maternal biomarkers linked to 1C metabolism pathways, which may be linked to lack of power. The same DMR has been previously associated with prenatal polychlorinated biphenyl exposure (PBE) in Europeans[11]. Interestingly, PBE exposure has been linked to repressed immune function in childhood[11] and to several metabolic traits, including antigalactic acid decarboxylase antibodies (thyroid antibodies) that are a marker of beta cell function[74]. Taken together with the results from our 2SMR analysis, this indicates that *C8orf31/LY6S* are promising candidates for exploring how exposures in early development can influence insulin dynamics in later life. It remains to be determined whether the environmentally induced methylation changes at the DMR are linked to epigenetic regulation of MER11C and if so, whether this reflects environmental perturbation of the ZNF808 protein itself, a binding partner of ZNF808, or other transcription factors that act independently of ZNF808.

DNA methylation states that are sensitive to the early development environment and that are also associated with genetic variation may allow for rapid adaption to changing environments[75]. Interestingly, ZNF808 mQTLs associated with *C8orf31* methylation have a higher MAF in Gambian and African populations than in Europeans, raising the possibility that these SNPs may have undergone selection in African populations. Whilst our observed causal association with T1 diabetes is tentative, evidence suggests that SNPs associated with T1 diabetes risk are under global positive selection[76] and one hypothesised benefit of this selection is in mediating immune responses against enteroviruses[76,77]. Future work examining evidence of selection at ZNF808 mQTLs would help understand whether increased methylation at the DMR is evolutionarily advantageous in some populations.

This work has several limitations. Firstly, we used ZNF808 ChIP-exo peaks obtained in the HEK293T cell line, which does not represent a physiologically relevant model for investigating DNA binding dynamics in the early embryo. Secondly, whilst the overlap between ZNF808 mQTLs and eQTLs suggests that genetic variation at ZNF808 may influence methylation at the *C8orf31* DMR via altering ZNF808 expression, further methods such as multiple trait colocalization[78] are needed to confirm this. Thirdly, the 2SMR results support but do not prove a causal relationship between *C8orf31* DMR methylation and glucose metabolism outcomes, because MR can be biased by weak genetic instruments, horizontal pleiotropy, and reverse causation. We took steps to reduce these biases, but our results require replication in additional large sample-sized cohorts and validation using *in vitro* cell line models. Furthermore, although 2SMR allows the use of summary statistics obtained from large GWAS (such as the GoDMC mQTL analysis in European blood samples), tissue-specific effects have not been considered, and the lack of individual level data limits the ability to stratify analyses by additional factors such as sex or age. Fourthly, 2SMR was performed using European adult samples, whilst the association between *C8orf31* DMR methylation and IGI/fasting insulin levels was performed using samples from Gambian children. Since DNAm is influenced by ancestry and age[79–81] , and phenotypes linked to glucose metabolism may also vary more in adults than in children (due to the presence of comorbidities in the former), further work in needed to confirm the extent to which the reported associations are population- and age-specific. Analysis of genotype, methylation, and phenotype data from cohorts with larger sample-sizes would improve the power to detect these associations.

In summary, we have investigated an example of how KZFP-mediated silencing of TEs in early development may influence phenotype in later life. In Chapter 4, we identified many other KZFPs that were associated with hvCpG methylation, raising the possibility that this may be the first of many examples.

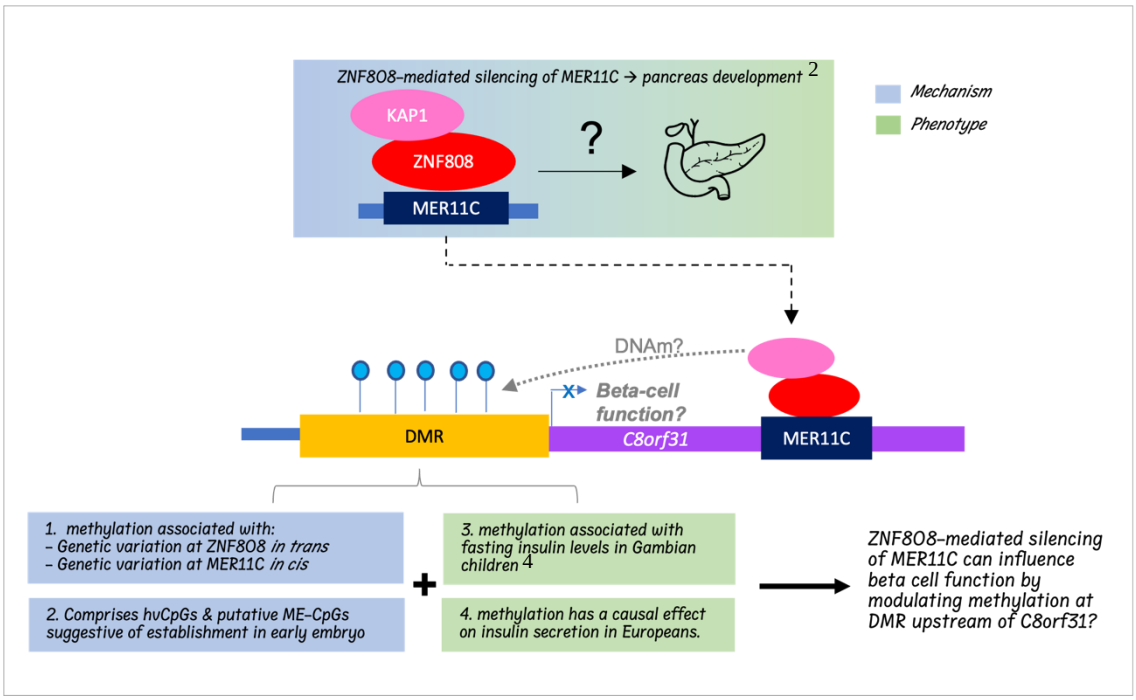


Figure 5.10. Summary of findings. ZNF808-mediated silencing of MER11C in early development has been previously found to be important for pancreas development[2]. In this chapter, several findings (1-4) suggest that *C8orf31* may be a downstream gene involved in this pathway. We hypothesise that ZNF808 binds to the MER11C retrotransposon located in the *C8orf31* gene region and recruits KAP1. In turn, we speculate that KAP1 recruits epigenetic silencing machinery that deposits DNA methylation at *C8orf31* which spreads to the DMR. Our 2SMR analysis indicates that a consequent increase in methylation at hypervariable region may have a casual effect on insulin secretion and T1 diabetes.

5.5 References

1. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*. 2017;543(7646):550-554.
2. de Franco E, Owens NDL, Montaser H, et al. Primate-specific ZNF808 is essential for pancreatic development in humans. *med*R*xiv* (2021).
3. Allen HL, Flanagan SE, Shaw-Smith C, et al. GATA6 haploinsufficiency causes pancreatic agenesis in humans. *Nat Genet*. 2012;44(1):20-22.
4. Antoun E, Issarapu P, di Gravio C, et al. DNA methylation signatures associated with cardiometabolic risk factors in children from India and The Gambia: results from the EMPHASIS study. *Clin Epigenetics*. 2022;14(1):6.
5. Liu CT, Merino J, Rybin D, et al. Genome-wide Association Study of Change in Fasting Glucose over time in 13,807 non-diabetic European Ancestry Individuals. *Sci Rep*. 2019;9(1):9439.
6. Ribel-Madsen R, Fraga MF, Jacobsen S, et al. Genome-wide analysis of DNA methylation differences in muscle and fat from monozygotic twins discordant for type 2 diabetes. *PLoS One*. 2012;7(12):e51302.
7. Yang R, Xing L, Wang M, Chi H, Zhang L, Chen J. Comprehensive Analysis of Differentially Expressed Profiles of lncRNAs/mRNAs and miRNAs with Associated ceRNA Networks in Triple-Negative Breast Cancer. *Cellular Physiology and Biochemistry*. 2018;50(2):473-488.
8. Matsuda R, Enokida H, Chiyomaru T, et al. LY6K is a novel molecular target in bladder cancer on basis of integrate genome-wide profiling. *Br J Cancer*. 2011;104(2):376-386.
9. Wang Z, Cheng H, Qi L, Sui D. Comprehensive analysis of long non-coding RNA using an associated competitive endogenous RNA network in Wilms tumor. *Mol Med Rep*. 2020;22(1):105-116.
10. Pennings JLA, Jennen DGJ, Nygaard UC, et al. Cord blood gene expression supports that prenatal exposure to perfluoroalkyl substances causes depressed immune functionality in early childhood. *J Immunotoxicol*. 2016;13(2):173-180.
11. Pittman GS, Wang X, Campbell MR, et al. Polychlorinated biphenyl exposure and DNA methylation in the Anniston Community Health Survey. *Epigenetics*. 2020;15(4):337-357.
12. Moore SE, Fulford AJC, Darboe MK, Jobarteh ML, Jarjou LM, Prentice AM. A randomized trial to investigate the effects of pre-natal and infant nutritional supplementation on infant immune development in rural Gambia: The ENID trial: Early Nutrition and Immune Development. *BMC Pregnancy Childbirth*. 2012;12(1):1-8.

13. Chandak GR, Silver MJ, Saffari A, et al. Protocol for the EMPHASIS study; epigenetic mechanisms linking maternal pre-conceptional nutrition and children's health in India and Sub-Saharan Africa. *BMC Nutr*. 2017;3(1):81.

14. Saffari A, Shrestha S, Issarapu P, et al. Effect of maternal preconceptional and pregnancy micronutrient interventions on children's DNA methylation: Findings from the EMPHASIS study. *Am J Clin Nutr*. 2020;112(4):1099-1113.

15. Candler T, Kessler N, Gunasekara C, et al. DNA methylation at a nutritionally sensitive region of the PAX8 gene is associated with thyroid volume and function in Gambian children. *Sci Adv*. 2021;7(45):eabj1561.

16. Kessler NJ, Waterland RA, Prentice AM, Silver MJ. Establishment of environmentally sensitive DNA methylation states in the very early human embryo. *Sci Adv*. 2018;4(7):eaat2624.

17. van Baak TE, Coarfa C, Dugué PA, et al. Epigenetic supersimilarity of monozygotic twin pairs. *Genome Biol*. 2018;19(1):2.

18. Harris RA, Nagy-Szakal D, Kellermayer R. Human metastable epiallele candidates link to common disorders. *Epigenetics*. 2013;8(2):157-163. doi:10.4161/epi.23438

19. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*. 2009;25(14):1841-1842.

20. Ramsay M, Sankoh O, Consortium as members of the AG study and the H. African partnerships through the H3Africa Consortium bring a genomic dimension to longitudinal population studies on the continent. *Int J Epidemiol*. 2016;45(2):305-308.

21. Derakhshan M, Kessler NJ, Ishida M, et al. Tissue-and ethnicity-independent hypervariable DNA methylation states show evidence of establishment in the early human embryo. *Nucleic Acids Res*. 2022;50(12):6735-6752.

22. Peters TJ, Buckley MJ, Statham AL, et al. De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin*. 2015;8(1):1-16.

23. Gunasekara CJ, Scott CA, Laritsky E, et al. A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol*. 2019;20(1):105.

24. Baak TE van, Coarfa C, Dugué PA, et al. Epigenetic supersimilarity of monozygotic twin pairs. *Genome Biol*. 2018;19(1):2.

25. Silver MJ, Saffari A, Kessler NJ, et al. Environmentally sensitive hotspots in the methylome of the early human embryo. Deelen J, ed. *Elife*. 2022;11:e72031. doi:10.7554/eLife.72031

26. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*. 2017;543(7646):550-554.

27. Min JL, Hemani G, Hannon E, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet*. 2021;53(9):1311-1321.

28. Phillips DIW, Clark PM, Hales CN, Osmond C. Understanding oral glucose tolerance: comparison of glucose or insulin measurements during the oral glucose tolerance test with specific measurements of insulin resistance and insulin secretion. *Diabetic Medicine*. 1994;11(3):286-292.

29. Wallace TM, Levy JC, Matthews DR. Use and abuse of HOMA modeling. *Diabetes Care*. 2004;27(6):1487-1495.

30. Davies NM, Holmes M v, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ*. 2018;362:k601-k601.

31. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. *Elife*. 2018;7.

32. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet*. 2017;13(11):e1007081.

33. Forgetta V, Manousaki D, Istomine R, et al. Rare genetic variants of large effect influence risk of type 1 diabetes. *Diabetes*. 2020;69(4):784-795.

34. Xue A, Wu Y, Zhu Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun*. 2018;9(1):1-14.

35. Chen J, Spracklen CN, Marenne G, et al. The trans-ancestral genomic architecture of glycemic traits. *Nat Genet*. 2021;53(6):840-860.

36. Manning AK, Hivert MF, Scott RA, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet*. 2012;44(6):659-669.

37. Wood AR, Jonsson A, Jackson AU, et al. A genome-wide association study of IVGTT-based measures of first-phase insulin secretion refines the underlying physiology of type 2 diabetes variants. *Diabetes*. 2017;66(8):2296-2309.

38. Strawbridge RJ, Dupuis J, Prokopenko I, et al. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes*. 2011;60(10):2624-2634.

39. Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*. 2010;42(2):105-116.

40. Soranzo N, Sanna S, Wheeler E, et al. Common variants at 10 genomic loci influence hemoglobin A1C levels via glycemic and nonglycemic pathways. *Diabetes*. 2010;59(12):3229-3239.

41. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet*. 2017;13(11):e100708.

42. Greco M F del, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat Med*. 2015;34(21):2926-2940.

43. Bowden J, del Greco M F, Minelli C, et al. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int J Epidemiol*. 2019;48(3):728-742.

44. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. Loos R, ed. *Elife*. 2018;7:e34408. doi:10.7554/eLife.34408

45. Dominguez-Salas P, Moore SE, Baker MS, et al. Maternal nutrition at conception modulates DNA methylation of human metastable epialleles. *Nat Commun*. 2014;5(1):3746.

46. Díez-Villanueva A, Mallona I, Peinado MA. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenetics Chromatin*. 2015;8(1):1-8.

47. Zhbannikov IY, Arbeev K, Ukraintseva S, Yashin AI. haploR: an R package for querying web-based annotation tools. *F1000Res*. 2017;6.

48. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40(D1):D930-D934.

49. Lappalainen T, Sammeth M, Friedländer MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501(7468):506-511.

50. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-330.

51. Dunham I. An integrated encyclopedia of DNA elements in the human An integrated encyclopedia of DNA elements in the human genome genome Repository Citation Repository Citation. *Nature*. 2012;489(7414):57.

52. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22(9):1790-1797.

53. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods*. 2015;12(10):931-934.

54. Dong S, Boyle AP. Predicting functional variants in enhancer and promoter elements using RegulomeDB. *Hum Mutat*. 2019;40(9):1292-1298.

55. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22(9):1790-1797.

56. Lukin K, Fields S, Hartley J, Hagman J. Early B cell factor: Regulator of B lineage specification and commitment. In: *Seminars in Immunology*. Vol 20. Elsevier; 2008:221-227.

57. Grundberg E, Meduri E, Sandling JK, et al. Global Analysis of DNA Methylation Variation in Adipose Tissue from Twins Reveals Links to Disease-Associated Variants in Distal Regulatory Elements. *The American Journal of Human Genetics*. 2013;93(5):876-890.

58. Liu CT, Merino J, Rybin D, et al. Genome-wide Association Study of Change in Fasting Glucose over time in 13,807 non-diabetic European Ancestry Individuals. *Sci Rep*. 2019;9(1):9439.

59. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015;44(2):512-525.

60. Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585.

61. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603-607.

62. Eom YS, Wilson JR, Bernet VJ. Links between thyroid disorders and glucose homeostasis. *Diabetes Metab J*. 2022;46(2):239-256.

63. Weinstein JN, Collisson EA, Mills GB, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. 2013;45(10):1113-1120.

64. Shmerling M, Chalik M, Smorodinsky NI, et al. LY6S, a New IFN-Inducible Human Member of the Ly6a Subfamily Expressed by Spleen Cells and Associated with Inflammation and Viral Resistance. *Immunohorizons*. 2022;6(4):253-272.

65. Rebollo R, Miceli-Royer K, Zhang Y, Farivar S, Gagnier L, Mager DL. Epigenetic interplay between mouse endogenous retroviruses and host genes. *Genome Biol*. 2012;13(10):R89. doi:10.1186/gb-2012-13-10-r89

66. Ewing AD, Smits N, Sanchez-Luque FJ, et al. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Mol Cell*. 2020;80(5):915-928.e5.

67. ter Horst KW, Gilijamse PW, Koopman KE, et al. Insulin resistance in obesity can be reliably identified from fasting plasma insulin. *Int J Obes*. 2015;39(12):1703-1709.

68. Tura A, Kautzky-Willer A, Pacini G. Insulinogenic indices from insulin and C-peptide: comparison of beta-cell function from OGTT and IVGTT. *Diabetes Res Clin Pract*. 2006;72(3):298-301.

69. Association AD. Diagnosis and classification of diabetes mellitus. *Diabetes Care*. 2010;33(Supplement_1):S62-S69.

70. Guan M, Keaton JM, Dimitrov L, et al. Genome-wide association study identifies novel loci for type 2 diabetes-attributed end-stage kidney disease in African Americans. *Hum Genomics*. 2019;13(1):1-17.

71. Cobelli C, Toffolo GM, Man CD, et al. Assessment of β-cell function in humans, simultaneously with insulin sensitivity and hepatic extraction, from intravenous and oral glucose tests. *American Journal of Physiology-Endocrinology and Metabolism*. 2007;293(1):E1-E15.

72. Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. *The Lancet*. 2014;383(9911):69-82.

73. Malecová B, Morris K v. Transcriptional gene silencing mediated by non-coding RNAs. *Curr Opin Mol Ther*. 2010;12(2):214.

74. Langer P, Kočan A, Tajtáková M, et al. Multiple adverse thyroid and metabolic health signs in the population from the area heavily polluted by organochlorine cocktail (PCB, DDE, HCB, dioxin). *Thyroid Res*. 2009;2(1):1-7.

75. Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A*. 2010;107 Suppl(suppl 1):1757-1764.

76. Corona E, Dudley JT, Butte AJ. Extreme evolutionary disparities seen in positive selection across seven complex diseases. *PLoS One*. 2010;5(8):e12236.

77. Raj T, Kuchroo M, Replogle JM, Raychaudhuri S, Stranger BE, De Jager PL. Common Risk Alleles for Inflammatory Diseases Are Targets of Recent Positive Selection. *The American Journal of Human Genetics*. 2013;92(4):517-529.

78. Giambartolomei C, Zhenli Liu J, Zhang W, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*. 2018;34(15):2538-2545.

79. Moen EL, Zhang X, Mu W, et al. Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics*. 2013;194(4):987-996. doi:10.1534/genetics.113.151381

80. Rahmani E, Shenhav L, Schweiger R, et al. Genome-wide methylation data mirror ancestry information. *Epigenetics Chromatin*. 2017;10(1):1.

81. Day K, Waite LL, Thalacker-Mercer A, et al. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol*. 2013;14(9).

Chapter 6

6. Discussion

Summary of chapter

This chapter summarises the main findings from this thesis in relation to the key research questions introduced in Chapter 1. It also discusses the significance of these findings, limitations, and directions for future research.

6.1 Summary of findings

The aim of this thesis was to identify and characterise tissue- and ethnicity- independent hypervariable DNA methylation loci in humans, reasoning that studying these loci would give insights into fundamental biological mechanisms influencing methylation variation. This section summarises the main findings in relation to the three key research questions introduced in Chapter 1.

Q1. Do hypervariable CpGs ('hvCpGs') exist in the human genome?

By leveraging 30 mainly publicly-available Illumina450K datasets, comprising 19 unique tissue/cell types and 8 ethnicities, we identified CpGs that were in the top 5% of variable loci by methylation variance in at least 65% of datasets. This yielded ~4000 CpGs that were hypervariable in a median of 13 tissue types and 7 ethnicities. Our methodological approach was designed to be robust to dataset specific drivers of methylation variation, such as sex, age, and cell heterogeneity. Supporting this, hvCpGs were highly variable in single-cell, single-sex and infant datasets and were insensitive to epigenetic drift effects in blood. Although hvCpG variability is unlikely to be driven by technical artefacts that are shared across datasets, we had pre-adjusted each dataset for the first 10 PCs of methylation variation to reduce technical artefacts and maximise power to detect true biological variation. To our knowledge, this is the first study that investigates the extent to which interindividual variation is shared across diverse tissues and ethnicities in the human genome.

Q2. Do hvCpGs show evidence of establishment in the early human embryo?

Noting that tissue-independent methylation variation has been previously observed at putative human MEs, we explored the hypothesis that hvCpGs may similarly be established in early development. We collated several published CpG sets to explore this hypothesis and observed large overlaps with these: i) 23% of hvCpGs overlapped a CpG previously identified as showing systemic interindividual variation (SIV), suggestive of methylation establishment before gastrulation[1–4]; ii) 13% of hvCpGs overlapped an epigenetic supersimilarity locus, indicative of establishment before MZ twinning[1], and iii) 7% overlapped a CpG that was differentially

methylated between MZ and DZ twins[5], further linking hvCpGs to MZ twinning events. These overlaps represented strong enrichments compared to the Illumina450K array background and to controls with similar distributions of methylation values across individuals. To further confirm that these enrichments are not driven by distributional properties of hvCpGs, we note that we could have also used multiple, randomly selected sets of control CpGs. In total, 37% of hvCpGs overlapped at least one of these CpG sets associated with DNAm establishment in the early embryo.

Using an unpublished foetal multi-tissue dataset, we found that an additional 23% of hvCpGs showed methylation correlation across tissues derived from different germ layers, suggesting that more hvCpGs may be established in early development beyond those identified in the studies above, and that the identification of hvCpGs is a high-powered approach to detect putative human MEs. Importantly, whilst previous SIV screens have used largely adult tissues, the median gestational age for this dataset was 16 weeks, confirming the notion that methylation variation at hvCpGs arises in early life. Finally, hvCpGs were enriched for CpGs robustly associated with season-of-conception in a rural Gambian population of subsistence farmers[6], indicative of sensitivity to the periconceptional environment.

Taken together, these findings suggest that variable methylation states at hvCpGs may have been established in the very early embryo and persisted into a diverse range of post-natal tissues.

Q3. Can we identify molecular mechanisms underpinning the establishment of variable methylation states in the early embryo and can we link these to postnatal phenotype?

- **Finding (a): hvCpG methylation is influenced but not determined by genotype**

We used a high-powered meta-analysis of mQTLs identified in blood from European samples[7] to probe genetic effects at hvCpGs. We reasoned that 'universal' (multi-tissue, multi-ethnic) mQTLs that may be driving cross-dataset methylation variability at hvCpGs, would have been identified in this meta-analysis. The majority (90%) of hvCpGs were associated with at least one mQTL, indicating that hvCpGs are influenced by genetic variation. However, hvCpGs were on average 5-fold more variable than controls that were matched for similar mQTL effects. Supporting this, hvCpGs were enriched for CpGs that are equivalently variable between MZ

co-twins and unrelated individuals, indicative of methylation variability that is independent of genotype[8].

This suggests that many hvCpGs could be classed as putative human MEs since they show SIV that is influenced but not determined by genotype. Although we acknowledge that unmeasured genetic variation or epistasis effects may influence hvCpG variability, the enrichment for loci that are discordant between MZ twins and for loci that are sensitive to the periconceptional environment, supports the notion that a portion of hvCpG variability may be attributed to stochastic and/or environmental effects in early development.

- **Finding (b): KZFP-mediated silencing of TEs can influence hvCpG methylation**

Murine ME methylation variation has been linked to variable binding of KZFPs to IAP elements (a class of retrotransposon) in a sequence dependent manner, resulting in stochastic methylation establishment and methylation variability between isogenic mice[3,9]. Although putative human MEs have been similarly linked to transposable elements[3,10], the extent to which KZFP-mediated silencing of TEs influences methylation had not been explored. HvCpGs were enriched for associations with genetic variants in *trans* located in KZFP genes suggesting that KZFP expression and activity can influence methylation at hvCpGs. Furthermore, ~10% of hvCpGs were within 3kb of a KAP1 (a co-factor of KZFPs) binding site in H1-ESCs. HvCpGs were particularly enriched for certain classes of retrotransposon, and for retrotransposon elements that were bound by KAP1 in ESCs compared to those not bound by KAP1. These findings suggest that hvCpGs may be preferentially targeted by KZFPs in the early embryo compared to array background CpGs, and that methylation variability at some hvCpGs may be linked to KZFPs-mediated silencing of TEs in the early embryo.

- **Finding (c): KZFP-mediated silencing of TEs may link methylation patterns in the early embryo to life-long health outcomes**

Finally, we identified an example of how KZFP-mediated silencing of TEs may influence phenotype in later life. A DMR comprising hvCpGs upstream of *C8orf31* was associated with genetic variation at the ZNF808 gene *in trans* and at a ZNF808 binding site located in a MER11C retrotransposon *in cis*. ZNF808-mediated silencing of MER11C is critical for pancreas

development/function[11], and we found that the *C8orf31* DMR is causally associated with traits linked to pancreatic beta cell function in a two-sample mendelian randomisation (2SMR) analysis, raising the possibility that *C8orf31* may be involved in this pathway. Overall, this finding suggests that KZFP-mediated silencing of TEs in early development may have a casual effect on phenotypes in later life by modulating methylation at hvCpGs.

6.2 Significance of findings

Our findings: i) position hvCpGs as age-stable, tissue-independent biomarkers of early stochastic and/or environmental effects on DNA methylation, and ii) highlight KZFP-mediated silencing of TEs as a novel mechanism by which variable DNAm loci established in the early human embryo can influence phenotype in later life.

This section expands on the relevance of these findings to the field of epigenetic epidemiology, with a particular focus on the Developmental Origins of Health and Disease (DOHaD)[12], and to furthering understanding into fundamental mechanisms influencing methylation variation. More speculatively, it discusses how the existence of hvCpGs in the human genome may offer insights into the role of DNAm in human adaptation and evolution.

6.2.1 Relevance to epigenetic epidemiology

6.2.1.1 hvCpGs as diagnostic and prognostic disease biomarkers

DNAm is considered an ideal disease biomarker because it is easily measurable and stable[13,14]. However, in a clinical setting, methylation needs to be measured in easily accessible tissues using non-invasive methods, and tissues such as blood and saliva may not reflect pathologically relevant tissue. Furthermore, the cell-type specificity of DNAm means that methylation variation measured in bulk tissue may be driven by differences in cell type proportions between individuals caused by the disease itself[15].

To the extent that hvCpGs can be associated with disease, hvCpGs would be particularly useful disease biomarkers. Firstly, many hvCpGs show SIV, meaning that disease-associated methylation states can be measured in easily accessible tissues. Because methylation at SIV loci is consistent across cell types, their methylation status is unlikely to be driven by cell type proportion differences in bulk tissue. Secondly, hvCpGs show high methylation variation in infant datasets and are robust to epigenetic drift effects in blood, highlighting their potential to contribute to disease prognosis. Indeed, longitudinal datasets have shown that SIV loci measured at one point can predict disease later in life[1,16,17]. Recently, SIV-CpG methylation has been used to predict disease status using machine learning models that were able to distinguish between schizophrenia case-control samples, outcompeting polygenic risk scores[18]. Direct evidence for age stability of hvCpG methylation can in future be investigated by examining methylation correlation across the life-course using longitudinal cohorts such as ALSPAC (Avon Longitudinal Study of Parents and Children)[19,20] and Generation R[21].

6.2.1.2 Utility for epigenome-wide association studies (EWAS)

Whilst probing the association between hvCpGs and phenotype was not a main aim of this thesis, we envisage that hvCpGs are particularly useful for EWAS. Firstly, hvCpGs show consistently high variability across diverse datasets, indicating that variability at these loci is robust to dataset-specific drivers of methylation variation, and is therefore both biologically meaningful and stable. Secondly, their high degree of variability will improve the power to detect associations, compared to the majority of the Illumina450K array CpGs that are not variable. Thirdly, hvCpGs are likely established in the early embryo and stable with age, making them useful for inferring the direction of causality between DNAm and phenotype. We envisage that hvCpGs therefore have the potential to make a significant contribution to furthering understanding of the aetiology of non-heritable diseases and the drivers of complex trait variation for which genetic variation cannot explain all phenotypic variation.

HvCpGs may also be particularly useful for studying the role of DNAm in mediating the effect of the early developmental environment on disease risk in later life. Whilst we examined the enrichment for season-of-conception loci identified in Gambian individuals, it is possible that hvCpGs may be associated with other environmental exposures. As discussed in Chapter 2, several putative human MEs show evidence of mediating the effects of early development

environment (such as season-of-conception, famine exposure, exposure to endocrine disruptors) on phenotypes in later life[12].

6.2.1.3 A novel, high-powered method for identifying putative human MEs

Screens to identify putative human MEs have relied on rare multi-tissue datasets with small sample sizes. We have shown that the identification of hvCpGs may be an alternative method to detect putative MEs. Supporting this, hvCpGs were particularly enriched for 'high confidence' SIV loci that had been identified in multiple independent studies. Additionally, some hvCpGs that had not been previously found to show SIV, were correlated across germ layers in our foetal multi-tissue dataset.

This approach would have several advantages over putative ME screens to date. Firstly, our hvCpG screen uses datasets which each have a large number of individuals, improving power to detect interindividual variation compared to existing approaches. Indeed, our simulation analyses using a 4-indvidual multi-tissue dataset typical of previous screens confirmed that datasets with small numbers of individuals have limited power to detect SIV loci. Secondly, we use largely publicly-available data that are easily obtainable compared to rare multi- datasets that are difficult to acquire and costly. Thirdly, our datasets span a larger number of tissue types beyond those used in putative ME studies, meaning that we may be detecting CpGs that are correlated across a more diverse range of tissues. Fourthly, by using a large number of genetically diverse datasets and ethnicities our approach may identify putative MEs that are particularly robust to genetic effects, compared to existing SIV screens that have largely used Caucasian sample. Finally, our approach also minimises dataset-specific drivers of methylation variation, such as sex, age, cell heterogeneity, and batch effects, thereby increasing confidence that hvCpGs are driven by true biological variation.

6.2.2 Insights into fundamental biological mechanisms driving methylation variation

KZFPs are implicated in establishing methylation variation at murine MEs (also referred to as variably methylated IAPs ('VM-IAPs')). According to this model, stochastic methylation establishment is attributed to variable KZFP binding between genetically identical cells in early development. Our findings suggest that methylation variation at some hvCpGs may similarly

have arisen due to variable epigenetic silencing by KZFPs in the early embryo. Although we are yet to understand whether this mechanism can influence methylation variation at all hvCpGs, this finding is significant because it furthers knowledge of fundamental biological mechanisms driving interindividual methylation variation.

We have not conclusively shown whether this mechanism can drive stochastic and environmental effects on methylation in early development. It is possible that hvCpGs are enriched for specific sequences or TF binding sites that render them susceptible to being variably bound by KZFPs in a random manner, resulting in stochastic methylation establishment as has been postulated at murine MEs. As discussed in Chapter 4, there may also be a number of ways in which KZFPs could underpin environmental lability at their genomic targets. KZFPs may bind effectors whose activity is influenced by the intracellular environment or may influence binding site occupancy of other environmentally sensitive proteins involved in DNAm, such as TET1 which is upregulated by vitamin C[22]. KZFP transcription or protein structure may also be influenced by environmental effects. The TSS of ZFP57 (a key regulator of parent-of-origin specific methylation) is associated with periconceptional and gestational exposures[10,23], raising the possibility that transcription of ZFP57 is environmentally sensitive, thereby influencing methylation at ZFP57 targets in the early embryo.

Uncovering the molecular mechanisms driving epigenetic variation in early development also offers insights into how DNAm contributes to disease aetiology. For example, our finding that ZNF808 -mediated silencing of MER11C influences hvCpG methylation at the *C8orf31* DMR has implications on understanding the influence of DNAm variation in the early embryo on insulin dynamics in later life. In the long term, these insights could inform health interventions, such as nutrient supplementation during the periconception and gestational periods to improve offspring health outcomes[24–28]. Furthermore, uncovering these mechanisms may offer new perspectives on the evolutionary significance of DNAm variation and its potential role in human adaptation (discussed in the next section).

6.2.3 The role of DNAm in human adaptation and evolution

Understanding how species adapt to changing environments is a key aim of evolutionary biology. Whilst increases in the frequency of favourable genetic variants occur over many generations, epigenetic mechanisms such as DNAm could facilitate adaptive responses to rapidly changing

environments in a single generation[29]. Our hvCpGs had top 5% methylation variation across ethnicities, indicating that these variable loci may be a conserved feature of the human genome.

Methylation variability at some hvCpGs could reflect evolutionarily advantageous adaptive mechanisms. For example, stochastic methylation that drives phenotypic variation may increase population fitness by increasing the chances that a proportion of the population will be well-suited to the current environment[30,31]. Alternatively, methylation patterns that are directly influenced by the environment, may be a source of phenotypic plasticity that allows the developing foetus to adjust to its current environment and/or anticipate the post-natal environment[32–34] .

These adaptive mechanisms may have evolved if stochastic methylation variation is maintained in each generation by a heritable genetic variant[35] . We have not explored whether hvCpGs are associated with these so-called *methylation variance mQTLs* ('mvQTLs'), but previous studies have identified mvQTLs at the putative human MEs *PAX-8* and *ZFP57* [1,36]. These mvQTLs may cause methylation variation by altering binding sites for TFs that directly or indirectly affect methylation. In line with this, the mvQTL associated with *PAX8* is in LD with an INDEL polymorphism[36]. In the context of phenotypic plasticity, these mvQTL may confer greater sensitivity to the early developmental environment, thereby genetically hardwiring environmental responsiveness into the human genome[33].

Transposable elements are known to contribute to mammalian evolution[37,38]. The epigenetic regulation of TEs by KZFPs in the early embryo is hypothesised to allow the genome to tolerate these elements in order for their regulatory potential to be exploited[39]. Indeed, TEs have a large repertoire of TF binding sites that can be used by the host in a context-dependent manner[39,40]. Recent evidence also suggests that that *de novo* methylation of the TE can induce the generation of new regulatory elements within the host DNA itself due to the spontaneous deamination of cytosine to thymine that changes the underlying genotype[41].

If methylation variation at hvCpGs is linked to KZFP-mediated silencing of TEs, we speculate that this may be evolutionarily relevant. We hypothesise that leaving some evolutionarily young TEs (such as IAPs and LTR12C elements) variably methylated in diverse tissues may be a source of phenotypic variation, allowing the host to 'test-out' whether any methylation states are

advantageous before being co-opted. If advantageous to the host, this methylation state may eventually become fixed, perhaps in a tissue-specific manner. Supporting this, mathematical modelling has shown that stochastic epigenetic modification at retrotransposons can increase both the probability and the rate of genetic fixation at the transposable elements and associated genes[42]. If the mechanisms regulating retrotransposon methylation (such as KZFPs) are sensitive to environmental effects in early development, they may act a source of phenotypic plasticity by influencing gene expression[43].

6.3 Limitations

This section discusses key limitations of the work described in this thesis and the ways in which these could be overcome with new technologies.

Limitation 1. Coverage of the Illumina450K array

We analysed CpGs covered on the Illumina450K array because this is the platform most widely used by the research community and our study design required a large number of publicly available datasets. However, the array covers only 1.7% of the 28 million CpG sites in the human genome and was designed to target genes and promoters[44,45], whilst the most variable regions in the genome exist largely outside of promoter regions[2]. Therefore, whilst we have assessed the top 5% of variable probes on the array, we may be missing many more variable probes genome-wide. As whole-genome bisulphite sequencing (WGBS) data becomes increasingly publicly available, our method for identifying hvCpGs can be easily applied to these datasets.

Limitation 2: Genetic effects at hvCpGs

The publicly available mQTL meta-analysis is the most high-powered mQTL screen to date[7], making it an invaluable source to probe the influence of genotype on DNAm variation. However, this mQTL screen is underpowered to detect *trans* mQTLs[7], meaning that there may be unidentified long-range genetic effects that influence methylation at hvCpGs. Whilst hvCpGs are enriched for *trans* genetic effects at KFZP genes, the lack of power to detect *trans* mQTLs

meant that we were unable to confirm i) how many hvCpGs are influenced by KZFP expression/activity, and ii) the extent to which KZFP genetic variation contributes to methylation variation at a given hvCpG.

Additionally, the mQTL meta-analysis does not consider multi-locus effects that may be influence methylation variability at hvCpGs[46], and does not give insight into the influence of other sources of genetic variation (such as short tandem repeat lengths[47]) on hvCpGs methylation. This limits our understanding the relative influences of genetic, environmental, and stochastic effects at hvCpGs. We note that whilst MZ datasets may overcome these drawbacks by allowing the assessment of methylation variability in the absence of genetic variation, such an approach is not appropriate for hvCpGs showing epigenetic supersimilarity, a property attributed to methylation establishment *before* MZ twinning.

Limitation 3: Pinpointing the timing of methylation establishment

The datasets we used to investigate evidence for hvCpG establishment in the early human embryo indicate that the timing of methylation establishment varies across hvCpG loci. For example, whilst some hvCpGs show epigenetic supersimilarity, indicative of establishment before MZ cleavage[1], others show MZ discordance, indicative of establishment *after* MZ cleavage[8]. The youngest methylation dataset we used was from foetal tissues with a median gestational age of 16 weeks, long after the implantation events that occur at embryonic day 7 in humans. Examination of methylation variation at multiple developmental stages across a sufficient number of embryos would offer insights into the locus-specific timing of establishment of hvCpG methylation variation.

Limitation 4: hvCpG methylation status at the single-cell level

We have not explored how interindividual methylation variation is linked to methylation establishment at the cellular level. At murine MEs, stochastic methylation establishment in the early embryo causes intercellular variegation effects (where some cells are hypermethylated and others are hypomethylated). These methylation states are then inherited by cell progeny, meaning that methylation measured in post-natal tissue is reflective of the proportion of cells

that were methylated in early development. This proportion will vary between individuals, resulting in interindividual variation.

A recent study examined the transcription profiles of >1500 cells from 88 pre-implantation embryos, ranging from the 8-cell stage to the time immediately before implantation, using single-cell RNA-sequencing[48]. A similar atlas of pre-implantation methylation states would have two advantages. Firstly, if a sufficient number of embryos were available at each embryonic stage, the timepoint of the onset of interindividual methylation variation in the pre-implantation embryo could be determined (relevant for limitation 3). Secondly, if multiple cells were collected from each embryo, intercellular methylation variation in genetically identical cells cultured under the same conditions could be explored in order to determine if there is indeed a stochastic component to methylation establishment at hvCpGs in early development. Since CpG methylation can either be methylated or unmethylated at the read-level, intercellular methylation heterogeneity established in early development could be subsequently inherited through successive cell generations, resulting in intermediate methylation states measured in bulk tissue.

Additionally, DNA and RNA can be isolated from the same cells in order to couple methylation with gene expression[49], which would be particularly useful for examining the interplay between KZFP expression and hvCpG methylation.

In practice, obtaining human pre-implantation embryos is difficult due to ethical challenges. We note the recent progress in developing *in vitro* model systems for pre-implantation embryonic development[50,51], which may in future provide a tractable system for interrogating pre-implantation methylation and transcription dynamics without the need for human embryos. This system would also be useful for exploring how environmental exposures in early development (such as nutrition) influence methylation.

Limitation 5: Mapping transposable elements and examining their methylation

We analysed the proximity of hvCpGs to TEs annotated by RepeatMasker, a tool which uses a sequence homology approach to match repeats from curated databases such as Dfam[52] against a reference genome. We used the hg19 reference genome because this is the genome used by the Ilumina450K array platform. However, this genome assembly is known to have gaps at GC rich regions and repetitive sequences[53], due to difficulties in uniquely mapping short-read

sequencing data to repetitive regions. Furthermore, RepeatMasker does not annotate TEs *de novo,* meaning that nonreference TE insertions will not have been annotated[54].

TE assembly and annotation methods can be greatly improved by long-read sequencing data[55] such as Oxford Nanopore Technologies (ONT)[56] . The long reads (> 10kb in length) can cover the whole TE, thereby improving the ability to map unique TEs in the genome, even at a low coverage[55,57,58]. Furthermore, methylation and genetic data can be obtained at the same time for a given read. Unlike bisulphite conversion that can lead to methylation inaccuracies[59], ONT uses native untreated DNA and differentiates between methylated and unmethylated cytosines using hidden Markov models and machine learning approaches[60].

Although nanopore sequencing is currently costly, as tools that optimise processing efficiency and read quality are developed, this platform will be invaluable for exploring TE epigenomics. In relation to our research, the improved TE annotation capability will allow in-depth investigation into how hvCpG methylation is linked to transposable elements. In particular, obtaining methylation and genotype data in a single molecule-specific long read will give insight into the determinants of hvCpG methylation heterogeneity at the single-cell level.

6.4 Future research

The work presented in this thesis has opened several avenues for future research:

6.4.1 Functional characterisation of the DMR upstream of *C8orf31*

In Chapter 5, we presented evidence that the DMR upstream of *C8orf31* may be influenced by ZNF808-mediated silencing of an intragenic MER11C element. However, we were unable to examine the methylation status of the MER11C element due to lack of coverage in our targeted Methyl-Seq dataset. Using ONT[56] (currently being generated by our lab using DNA from Gambian children), we will be able to examine methylation status across the entire *C8orf31* gene and promoter region within a single read. If the *C8orf31 DMR* is linked to methylation at the intragenic MER11C element, this would confirm that ZNF808 not only influences methylation at its target retrotransposon but also at flanking regions.

Next, we plan to perform functional analyses with collaborators at the University of Cambridge (Dr Michael Imbeault and Professor Anne Ferguson-Smith). The clustered regulatory interspaced short palindromic repeat (CRISPR)-deactivated Cas9 (dCas9)-based editing system[61] is a particularly useful tool to examine functional consequences of modulating methylation. This system uses dCas9 fused to DNMT3a[62] or to TET proteins[63] to respectively methylate and demethylate the target region without DNA cleavage. *C8orf31* shows low expression levels in the pancreas, in line with the notion that ZNF808-mediated silencing of MER11C is required for pancreas development[11]. Upregulating *C8orf31* in the pancreatic beta cells by fusing dCas9 to a TET protein, followed by RNA-sequencing, will therefore allow us to determine the downstream genes that are regulated by *C8orf31* in the pancreas.

In Chapter 5, we noted that the complementary strand of *C8orf31* encodes the newly identified protein LY6S[64]. Further work is needed to explore the relationship between C*8orf31* and *LY6S* expression, and whether the long non-coding RNA transcribed from *C8orf31* can regulate *LY6S*.

6.4.2 Further probe KZFP-mediated silencing of TEs as a mechanism influencing methylation variation in early development

In Chapter 4, we examined the association between KZFP mQTLs and hvCpGs in order to investigate whether KZFP activity can influence methylation at hvCpGs. Inactivation of KZFPs in ESCs by genome editing tools such as CRISPR-Cas9 (where Cas9 is catalytically active)[65] , and/or examination of the association between hvCpG methylation and KZFP expression in ESCs[66], would confirm whether KZFPs can influence hvCpG methylation in early development[67–69] .

Furthermore, the extent to which methylation variation at hvCpGs is linked to sequence variation at nearby TEs genome-wide remains to be explored. In Chapter 5, we found that the DMR upstream of *C8orf31* was associated with genetic variation at a ZNF808 binding site in a nearby MER11C element. Because the TE targets of KZFPs are well established[40], this analysis could be extended to all hvCpGs that are proximal to a TE. The targeted Methyl-Seq dataset used in Chapter 5 would be useful for this as it covers many of our hvCpGs.

Whilst these analyses could support the notion that hvCpGs methylation is influenced by KZFP-mediated silencing of TEs, unpicking the relative influences of genetic, stochastic, and environmental effects is more challenging. The application of ONT would allow examination of whether methylation variability is linked to TE sequence features, such as variants that might reduce KZFP binding affinity, or binding sites for proteins that may have an antagonistic effect on methylation, as has been found in mouse[70]. If such features are identified, their effect on KZFP binding affinity could be directly examined using *in* vitro methods such as ChIP-seq[71] or ChIP-exo[72]. Culturing ESCs in varying nutrient conditions might also give greater insight into the extent to which KZFP activity and expression is influenced by environmental effects.

6.4.3 *C8orf31*: the first of many examples?

We focussed on the DMR upstream of *C8orf31* because it is associated with genetic variation at *ZNF808*, which has a known role in pancreatic cell fate specification[11]. This link to *ZNF808* helped us to select which GWAS traits to include in our two-sample mendelian randomisation analysis (2SMR) in order to investigate the causal effect of DMR methylation on phenotype[73].

Other hvCpGs were also associated with KZFP *trans* mQTL, meaning that there may be additional examples of how methylation states influenced by KZFP-mediated silencing of TEs in early development may influence phenotype in later life. The 'Open Targets' database[74] could be used to infer the possible phenotypes influenced by these KZFPs. These phenotypes could inform the GWAS used in 2SMR analyses, instead of using a phenome-wide 2SMR approach which would have a high multiple testing burden.

6.4.4 Beyond KZFP-mediated silencing of TEs?

The epigenetic reprogramming events that occur in early development are highly dynamic and other mechanisms besides KZFP-mediated silencing of TEs embryo may drive methylation variation at hvCpGs. Notably, hvCpGs were enriched for regions of parent-of-origin specific methylation (PofOm) and for proximity to binding sites for ZNF445 and ZFP57 that maintain PofOm post fertilisation. These findings suggest that mechanisms regulating PofOm may influence methylation variation at specific regions, possibly in an environmentally sensitive manner. Indeed, methylation at several imprinted genes has been linked to environmental exposures in early human development[10,75–78]. Most PofOm regions originate as germline differentially methylated regions (gDMRs) that are resistant to methylation erasure shortly after fertilisation and are maintained in the developing embryo. This could be investigated by examining the overlap between hvCpGs and gDMRs that have been identified using WGBS data[79]. Methylation variation at regions of PofOm may be particularly relevant to DOHaD because of the known role of imprinted genes in foetal growth and development[67–69].

6.4.5 hvCpGs on chromosome X

Chromosomes X and Y are often excluded from methylation studies because of the gender bias introduced by the random inactivation of one X chromosome that occurs during gestation in females[80]. This means that the signal in female samples is averaged across the active and inactive X chromosomes. We had similarly removed X and Y chromosomes from our analyses. However, the X chromosome contains many genes with important biological functions[81,82] and there is evidence that aberrant epigenetic patterning on the X chromosome may affect disease development [81,83]. Future work could therefore identify hvCpGs on the X chromosome by pre-processing this chromosome separately before stratifying the analyses by sex.

6.4.6 Screen for methylation variance QTLs 'mvQTLs'

Screening for the association between hvCpG methylation variation and genotype would give insights into the role of hvCpGs in human evolution and adaptation. The presence of 'methylation variance QTLs' (mvQTLs) would support the hypothesis that methylation variation is fixed by a heritable genetic variant[35]. In some cases, these mvQTLs may represent G x E effects, that confer phenotypic plasticity. Screens for mvQTLs associated with hvCpGs in our Gambian dataset and other large datasets with both genotype and methylation data available, followed by investigation into whether these variants are under positive selection in humans[84,85], may be indicative of potential adaptive advantage of hvCpGs. We would be particularly interested in identifying mvQTLs at KZFP genes and their TE targets, to probe whether methylation variation is genetically hardwired at features that are epigenetically regulated in early development and that can influence gene expression.

6.5 Conclusion

This thesis has added to existing knowledge of the fundamental properties of human DNAm variation. First, we have identified a set of CpGs that are highly variable in diverse tissues and ethnicities, suggesting that epigenetic variability is a fundamental feature of the genome. Second, we have linked these CpGs to the early embryo, indicating that mechanisms driving methylation variability in early development can be maintained in diverse tissues and ethnicities. Thirdly, we have highlighted KZFP-mediated silencing of transposable elements as a possible mechanism influencing methylation variability in the early embryo. Further characterisation of these CpGs, and exploration of their mechanisms of establishment, will give insights into the role of DNAm in mediating early stochastic/environmental effects on life-long health and disease, and in potentially facilitating adaptive responses to rapidly changing environments. Overall, we believe that hvCpGs have the potential to advance understanding of how variable methylation states may have evolved and how they contribute to phenotypic variation.

6.6 References

1. van Baak TE, Coarfa C, Dugué PA, et al. Epigenetic supersimilarity of monozygotic twin pairs. *Genome Biol.* 2018;19(1):2.
2. Gunasekara CJ, Scott CA, Laritsky E, et al. A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol.* 2019;20(1):105.
3. Kessler NJ, Waterland RA, Prentice AM, Silver MJ. Establishment of environmentally sensitive DNA methylation states in the very early human embryo. *Sci Adv.* 2018;4(7):eaat2624.
4. Harris RA, Nagy-Szakal D, Kellermayer R. Human metastable epiallele candidates link to common disorders. *Epigenetics.* 2013;8(2):157-163.
5. van Dongen J, Nivard MG, Willemsen G, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat Commun.* 2016;7.
6. Silver MJ, Saffari A, Kessler NJ, et al. Environmentally sensitive hotspots in the methylome of the early human embryo. Deelen J, ed. *Elife.* 2022;11:e72031.
7. Min JL, Moffitt T, Caspi A, Arseneault L, Relton C. Genomic and phenomic insights from an atlas of genetic effects on DNA methylation. *Nat Genet.* 2021;53(9):1311-1321.
8. Planterose Jiménez B, Liu F, Caliebe A, et al. Equivalent DNA methylation variation between monozygotic co-twins and unrelated individuals reveals universal epigenetic inter-individual dissimilarity. *Genome Biol.* 2021;22(1):18.
9. Bertozzi TM, Elmer JL, Macfarlan TS, Ferguson-Smith AC. KRAB zinc finger protein diversification drives mammalian interindividual methylation variability. *Proceedings of the National Academy of Sciences.* 2020;117(49):31290-31300.
10. Silver MJ, Kessler NJ, Hennig BJ, et al. Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome Biol.* Published online 2015.
11. de Franco E, Owens NDL, Montaser H, et al. Primate-specific ZNF808 is essential for pancreatic development in humans. *medRxiv* (2021).
12. Fleming TP, Watkins AJ, Velazquez MA, et al. Origins of lifetime health around the time of conception: causes and consequences. *The Lancet.* 2018;391(10132):1842-1852.
13. García-Giménez JL, Seco-Cervera M, Tollefsbol TO, et al. Epigenetic biomarkers: current strategies and future challenges for their use in the clinical laboratory. *Crit Rev Clin Lab Sci.* 2017;54(7-8):529-550.
14. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. *Nat Rev Genet.* 2019;20(2):109-127. -2
15. Qi L, Teschendorff AE. Cell-type heterogeneity: Why we should adjust for it in epigenome and biomarker studies. *Clin Epigenetics.* 2022;14(1):31. 3

16. Candler T, Kühnen P, Prentice AM, Silver M. Epigenetic regulation of POMC; implications for nutritional programming, obesity and metabolic disease. *Front Neuroendocrinol*. 2019;54.

17. van Dijk SJ, Tellam RL, Morrison JL, Muhlhausler BS, Molloy PL. Recent developments on the role of epigenetics in obesity and metabolic disease. *Clin Epigenetics*. 2015;7(1):1-13.

18. Gunasekara CJ, Hannon E, MacKay H, et al. A machine learning case–control classifier for schizophrenia based on DNA methylation in blood. *Transl Psychiatry*. 2021;11(1):412.

19. Boyd A, Golding J, Macleod J, et al. Cohort profile: the 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol*. 2013;42(1):111-127.

20. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol*. 2013;42(1):97-110.

21. Kooijman MN, Kruithof CJ, van Duijn CM, et al. The Generation R Study: design and cohort update 2017. *Eur J Epidemiol*. 2016;31:1243-1264.

22. Blaschke K, Ebata KT, Karimi MM, et al. Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature*. 2013;500(7461):222-226.

23. Amarasekera M, Martino D, Ashley S, et al. Genome-wide DNA methylation profiling identifies a folate-sensitive region of differential methylation upstream of ZFP57-imprinting regulator in humans. *The FASEB Journal*. 2014;28(9):4068-4076.

24. Garmendia ML, Corvalan C, Uauy R. Assessing the Public Health Impact of Developmental Origins of Health and Disease (DOHaD) Nutrition Interventions. *Ann Nutr Metab*. 2014;64(3-4):226-230.

25. Chu AHY, Tint MT, Chang HF, et al. High placental inositol content associated with suppressed pro-adipogenic effects of maternal glycaemia in offspring: the GUSTO cohort. *Int J Obes*. 2021;45(1):247-257.

26. Saffari A, Shrestha S, Issarapu P, et al. Effect of maternal preconceptional and pregnancy micronutrient interventions on children's DNA methylation: Findings from the EMPHASIS study. *Am J Clin Nutr*. 2020;112(4):1099-1113.

27. Keats EC, Haider BA, Tam E, Bhutta ZA. Multiple-micronutrient supplementation for women during pregnancy. *Cochrane Database of Systematic Reviews*. 2019;(3).

28. Godfrey KM, Cutfield W, Chan SY, Baker PN, Chong YS. Nutritional intervention preconception and during pregnancy to maintain healthy glucose metabolism and offspring health ("NiPPeR"): study protocol for a randomised controlled trial. *Trials*. 2017;18(1):1-12.

29. Yona AH, Frumkin I, Pilpel Y. A relay race on the evolutionary adaptation spectrum. *Cell*. 2015;163(3):549-559.

30. Herman JJ, Spencer HG, Donohue K, Sultan SE. How stable 'should' epigenetic modifications be? Insights from adaptive plasticity and bet hedging.. *Evolution (N Y)*. 2014;68(3):632-643.

31. Angers B, Perez M, Menicucci T, Leung C. Sources of epigenetic variation and their applications in natural populations. *Evol Appl*. 2020;13(6):1262-1278.

32. Low FM, Gluckman PD, Hanson MA. Developmental Plasticity, Epigenetics and Human Health. *Evol Biol*. 2012;39:650-665.

33. Gluckman PD, Hanson MA, Low FM. The role of developmental plasticity and epigenetics in human health. *Birth Defects Res C Embryo Today*. 2011;93(1):12-18.

34. Bateson P, Gluckman P, Hanson M. The biology of developmental plasticity and the Predictive Adaptive Response hypothesis. *J Physiol*. 2014;592(11):2357-2368.

35. Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A*. 2010;107 Suppl(suppl 1):1757-1764.

36. Candler T, Kessler N, Gunasekara C, et al. DNA methylation at a nutritionally sensitive region of the PAX8 gene is associated with thyroid volume and function in Gambian children. *Sci Adv*. 2021;7(45):eabj1561.

37. Colonna Romano N, Fanti L. Transposable Elements: Major Players in Shaping Genomic and Evolutionary Patterns. *Cells*. 2022;11(6).

38. Senft AD, Macfarlan TS. Transposable elements shape the evolution of mammalian development. *Nat Rev Genet*. 2021;22(11):691-711.

39. Ecco G, Imbeault M, Trono D. KRAB zinc finger proteins. *Development (Cambridge)*. 2017;144(15):2719-2729.

40. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*. 2017;543(7646):550-554. doi:10.1038/nature21683

41. Zhou W, Liang G, Molloy PL, Jones PA. DNA methylation enables transposable element-driven genome expansion. *Proc Natl Acad Sci U S A*. 2020;117(32):19359-19366.

42. Branciamore S, Rodin AS, Gogoshin G, Riggs AD. Epigenetics and Evolution: Transposons and the Stochastic Epigenetic Modification Model. 2(2):148-162.

43. Sharif J, Shinkai Y, Koseki H. Is there a role for endogenous retroviruses to mediate long-term adaptive phenotypic response upon environmental inputs? *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2013;368(1609):20110340.

44. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288-295.

45. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17(1).

46. Egli T, Vukojevic V, Sengstag T, et al. Exhaustive search for epistatic effects on the human methylome. *Sci Rep*. 2017;7(1).

47. Quilez J, Guilmatre A, Garg P, et al. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res*. 2016;44(8):3750-3762.

48. Petropoulos S, Edsgärd D, Reinius B, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell*. 2016;165(4):1012-1026.

49. Angermueller C, Clark SJ, Lee HJ, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods*. 2016;13(3):229-232.

50. Liu X, Tan JP, Schröder J, et al. Modelling human blastocysts by reprogramming fibroblasts into iBlastoids. *Nature*. 2021;591(7851):627-632.

51. Yu L, Wei Y, Duan J, et al. Blastocyst-like structures generated from human pluripotent stem cells. *Nature*. 2021;591(7851):620-626.

52. Hubley R, Finn RD, Clements J, et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res*. 2016;44(D1):D81-D89.

53. Lander ES. Initial impact of the sequencing of the human genome. *Nature*. 2011;470(7333):187-197.

54. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75-81.

55. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. *Nat Rev Genet*. 2018;19(11):688-704.

56. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018;36(4):338-345.

57. McDonald TL, Zhou W, Castro CP, et al. Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat Commun*. 2021;12(1):3586.

58. Chu C, Borges-Monroy R, Viswanadham V v, et al. Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun*. 2021;12(1):3836.

59. Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. *Biology (Basel)*. 2016;5(1):3.

60. Liu Y, Rosikiewicz W, Pan Z, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol*. 2021;22(1):295.

61. Urbano A, Smith J, Weeks RJ, Chatterjee A. Gene-specific targeting of DNA methylation in the mammalian genome. *Cancers (Basel)*. 2019;11(10):1515.

62. Huang YH, Su J, Lei Y, et al. DNA epigenome editing using CRISPR-Cas SunTag-directed DNMT3A. *Genome Biol*. 2017;18(1):1-11.

63. Kang JG, Park JS, Ko JH, Kim YS. Regulation of gene expression by altered promoter methylation using a CRISPR/Cas9-mediated epigenetic editing system. *Sci Rep*. 2019;9(1):1-12.

64. Shmerling M, Chalik M, Smorodinsky NI, et al. LY6S, a New IFN-Inducible Human Member of the Ly6a Subfamily Expressed by Spleen Cells and Associated with Inflammation and Viral Resistance. *Immunohorizons*. 2022;6(4):253-272.

65. Ran F, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*. 2013;8(11):2281-2308.

66. Lemma RB, Fleischer T, Martinsen E, et al. Pioneer transcription factors are associated with the modulation of DNA methylation patterns across cancers. *Epigenetics Chromatin*. 2022;15(1):13.

67. Piedrahita JA. The role of imprinted genes in fetal growth abnormalities. *Birth Defects Res A Clin Mol Teratol*. 2011;91(8):682-692.

68. Moore GE, Ishida M, Demetriou C, et al. The role and interaction of imprinted genes in human fetal growth. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2015;370(1663):20140074.

69. Reik W, Walter J. Genomic imprinting: Parental influence on the genome. *Nat Rev Genet*. 2001;2(1):21-32.

70. Costello KR, Leung A, Trac C, et al. Sequence features of retrotransposons allow for epigenetic variability. *Elife*. 2021;10:e71104.

71. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet*. 2009;10(10):669-680.

72. Rhee HS, Pugh BF. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol*. 2012;100(1):21-24.

73. Relton CL, Smith GD. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Injt J Epidemiol*. 2012;41(1):161-176.

74. Koscielny G, An P, Carvalho-Silva D, et al. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res*. 2017;45(D1):D985-D994.

75. Heijmans BT, Tobi EW, Stein AD, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A*. 2008;105(44):17046-17049.

76. Steegers-Theunissen RP, Obermann-Borst SA, Kremer D, et al. Periconceptional maternal folic acid use of 400 μg per day is related to increased methylation of the IGF2 gene in the very young child. *PLoS One*. 2009;4(11).

77. Tserga A, Binder AM, Michels KB. Impact of folic acid intake during pregnancy on genomic imprinting of IGF2/H19 and 1-carbon metabolism. *The FASEB Journal*. 2017;31(12):5149-5158.

78. Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nat Rev Genet*. 2019;20(4):235-248.

79. Sanchez-Delgado M, Court F, Vidal E, et al. Human Oocyte-Derived Methylation Differences Persist in the Placenta Revealing Widespread Transient Imprinting. Bartolomei MS, ed. *PLoS Genet*. 2016;12(11):e1006427-e1006427.

80. Lindsay S, Monk M, Holliday R, et al. Differences in methylation on the active and inactive human X chromosomes. *Ann Hum Genet*. 1985;49(2):115-127.

81. Deng X, Berletch JB, Nguyen DK, Disteche CM. X chromosome regulation: diverse patterns in development, tissues and disease. *Nat Rev Genet*. 2014;15(6):367-378.

82. Libert C, Dejager L, Pinheiro I. The X chromosome in immune functions: when a chromosome makes the difference. *Nat Rev Immunol*. 2010;10(8):594-604.

83. Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet*. 1999;23(2):185-188.

84. Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. *BMC Biol*. 2017;15(1):1-10.

85. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol*. 2006;4(3):e72-.

Appendix A

Supplementary Figures for Chapter 3

The supplementary material for this chapter is also available at:
https://academic.oup.com/nar/article/50/12/6735/6609816#supplementary-data

A

Variance threshold (i)

% datasets (j)	20%	19%	18%	17%	16%	15%	14%	13%	12%	11%	10%	9%	8%	7%	6%	5%	4%	3%	2%	1%
60	48376	44647	41087	37564	34259	30943	27767	24663	21748	18904	16269	13798	11497	9409	7344	5568	4013	2701	1592	645
65	40811	37627	34441	31370	28425	25546	22737	20165	17635	15301	13067	10998	9105	7350	5734	4330	3127	2117	1212	469
70	34763	31834	28954	26246	23610	21180	18753	16479	14411	12390	10517	8756	7167	5744	4490	3321	2436	1575	898	305
75	28218	25608	23222	20949	18756	16663	14621	12853	11054	9445	7953	6560	5340	4291	3289	2459	1735	1102	597	186
80	21957	19831	17835	15990	14147	12503	10913	9516	8131	6796	5646	4631	3737	2954	2230	1647	1134	695	353	101
85	15967	14383	12832	11346	9944	8651	7463	6409	5431	4522	3731	3039	2394	1857	1406	1005	690	395	163	34
90	10429	9261	8152	7105	6188	5281	4528	3804	3189	2594	2136	1691	1297	1002	717	471	289	163	58	9
95	4628	4035	3464	2986	2538	2131	1777	1476	1188	962	747	560	405	277	172	113	68	27	6	0
100	1681	1428	1196	1012	830	693	573	459	385	304	231	167	121	81	54	33	25	9	2	0

B

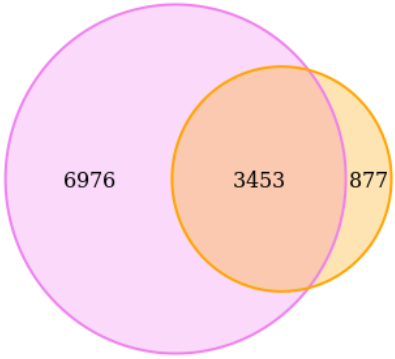


Figure S1. **A large proportion of hvCpGs identified using *i,j* = [5,65] overlap an alternative set obtained using *i,j* = [20,90]. A)** Heatmap showing the number of CpGs identified at *i,j* = [5,65] (orange) and *i,j* =[20,90] (purple). **B)** Venn diagram showing the overlap between these two CpG sets.

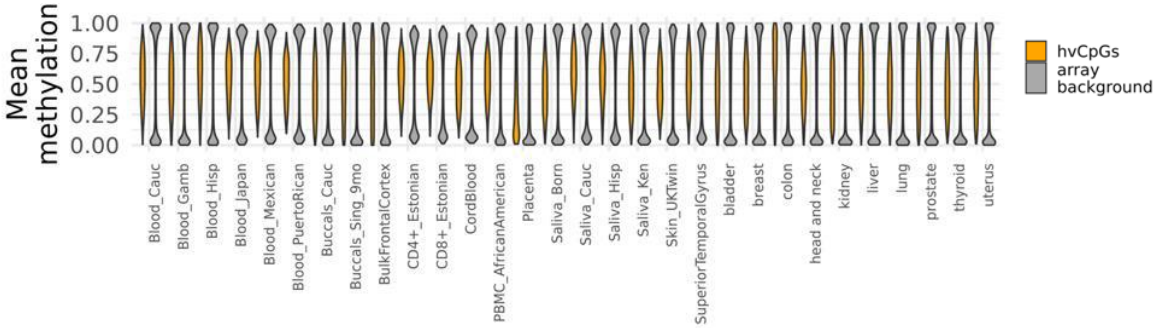


Figure S2. hvCpGs are enriched for intermediate methylation values. The distribution of mean methylation Beta values at hvCpGs (orange) and array background CpGs (grey).

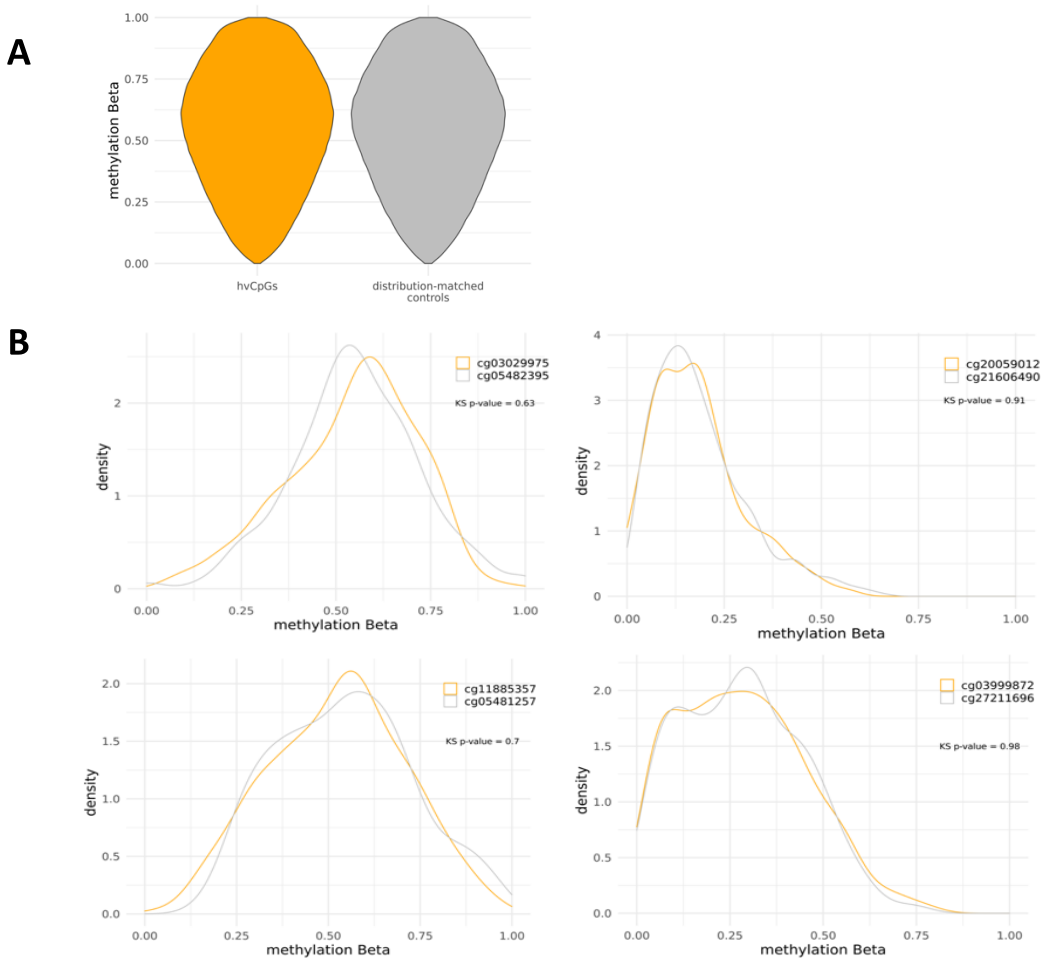


Figure S3. Matching each hvCpG to a control with a similar distribution of methylation Beta values in Caucasian blood. A) The distribution of methylation Beta values at 3,566 hvCpGs and their corresponding distribution-matched controls selected from the 'Blood_Cauc' dataset (Supplementary Table 1). **B)** Examples of the similar distribution of methylation Beta values at hvCpGs (orange) and their corresponding controls (grey). The median Kolmogorov-Smirnov p-value over all 3,566 hvCpG-control pairs was 0.91 (IQR = [0.63, 0.98]). See 'Methods' for further details.

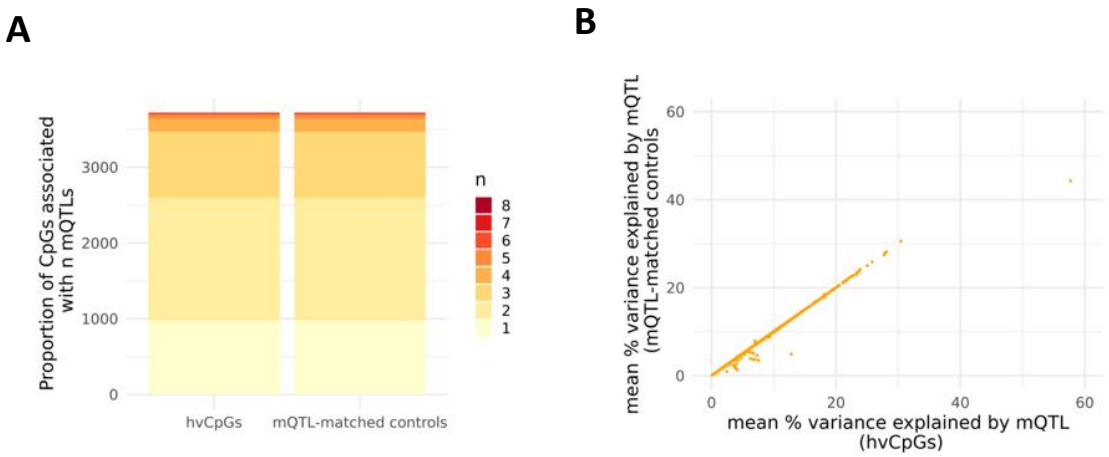


Figure S4. Matching each hvCpG to a CpG with similar mQTL effects using the GoDMC mQTL analysis. Each of the 3,722 hvCpGs reported in the GoDMC mQTL analysis (Min et al. 2021), was matched to a control CpG with **A)** the same number of mQTL associations, and **B)** similar mean % variance explained by mQTL, requiring each control CpG to be present in at least as many datasets as the hvCpG.

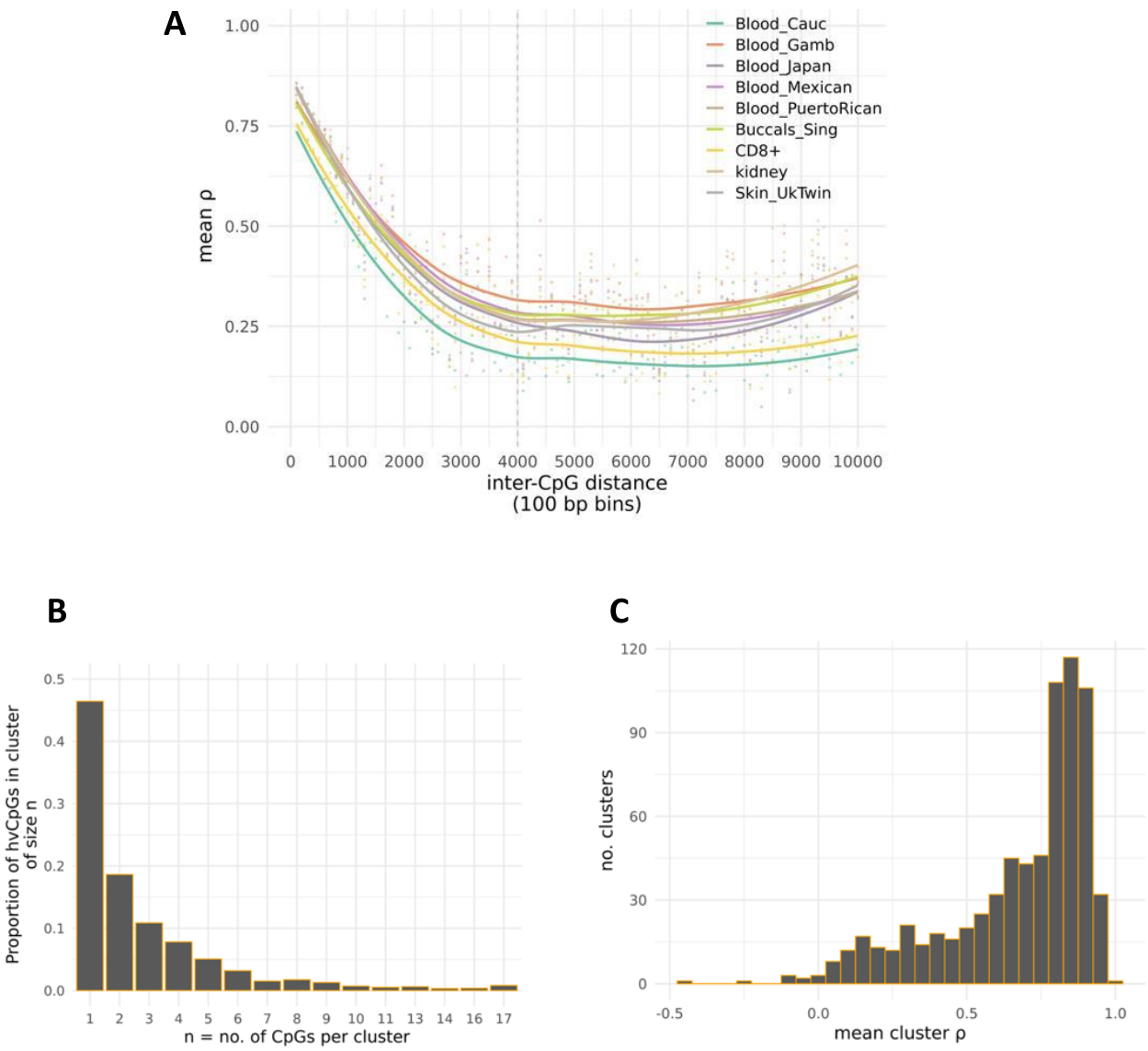


Figure S5. Clustering of hvCpGs. A) The decay of methylation correlation with distance at hvCpGs. Each point indicates the average pairwise Spearman rho across hvCpG pairs with inter-CpG distance falling within each 100 bp bin. Curves are loess lines of best fit for each dataset. Dashed vertical line at 4000 bp is the threshold we used for defining hvCpG clusters. **B)** The proportion of hvCpGs that fall into clusters comprising *n* CpGs, using an inter-CpG distance of 4000 bp to define clusters. **C**) The average Spearman rho across the 716 hvCpG clusters that comprise at least 2 CpGs. For each cluster, we calculated the average pairwise Spearman rho across hvCpG pairs for every dataset in which all CpGs in the cluster were covered, before taking the mean across these datasets. Note, in A and C we only considered datasets with N $\geq$ 100.

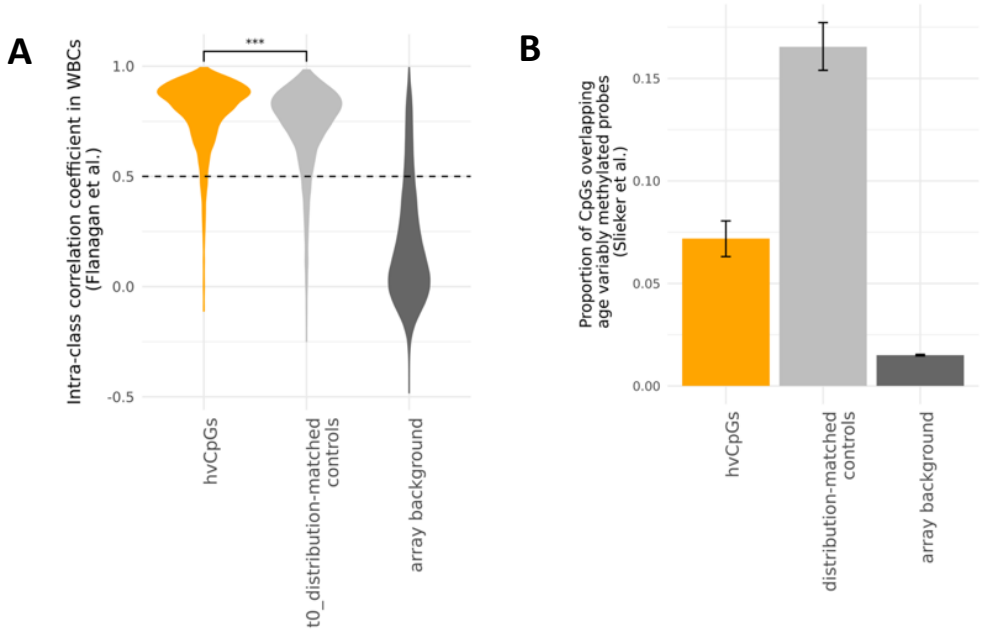


Figure S6. Methylation at hvCpGs is stable with age and is not influenced by sex. A) The distribution of temporal ICC scores at 3,564 hvCpGs and CpGs with similar distribution of methylation Beta values at the first time-point (t0) covered in the Flanagan et al. (2015) dataset (see Methods). The horizontal line at ICC = 0.5 indicates the threshold specified by Flanagan et al. (2015) above which probes are considered to be temporally stable. *** indicates Wilcoxon paired signed-rank test p-value < 0.001. **B)** The proportion of 3,566 hvCpGs and corresponding 'distribution-matched controls' (Table 1) covered in the 'Blood_Cauc' dataset that overlap probes showing increased variability with age in whole blood in a study of 18-to-88-year olds by Slieker et al. (2016).

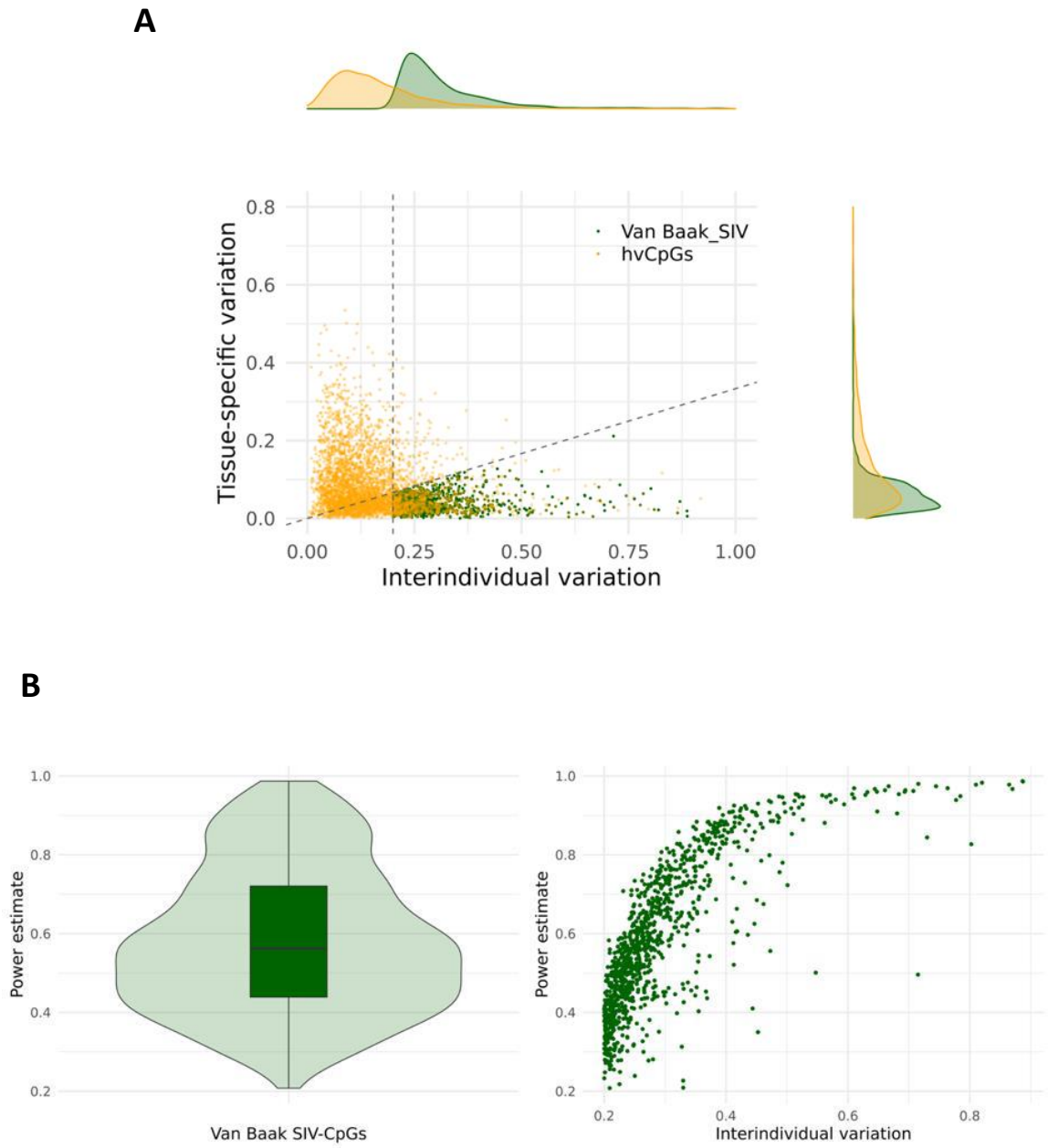


Figure S7. Estimating the power to detect SIV-CpGs reported by van Baak et al. (2018) using a 4-individual multi-tissue dataset (GSE50192). A) Replication of the screen by van Baak et al., in which 1,042 SIV-CpGs (green points) were identified, using the same methylation data from abdominal aorta, gall bladder and ischiatic nerve in 4 individuals. Orange points are hvCpGs. **B**) Estimating the power to detect each of the 1,042 SIV-CpGs reported by van Baak et al. (2018). See Methods for further details**. Left:** distribution of power estimates for each SIV-CpG across 1000 permutations**. Right:** confirmation that estimated power to detect SIV is correlated with interindividual variation.

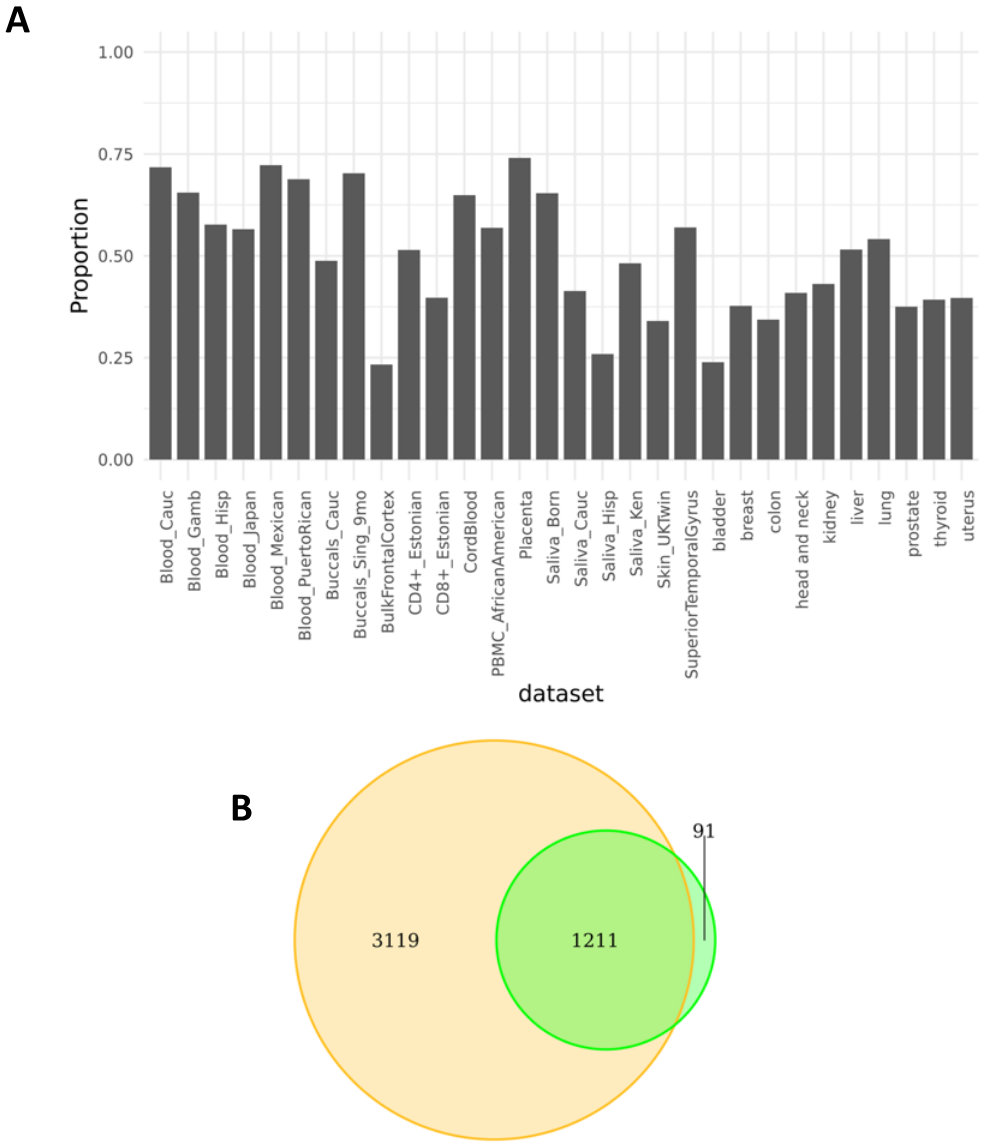


Figure S8. Adjustment for the first 10 PCs of methylation variation in each dataset increases the number of hvCpGs. A) For each dataset, we extracted the top 5% of CpGs by methylation Beta variance before and after adjusting methylation for the first 10 PCs. We then determined the proportionate overlap between PC-adjusted and PC-unadjusted variable CpGs. As expected, PC-adjustment influences the most variable CpGs within each dataset. **B)** The overlap between the set of 4,330 hvCpGs identified using PC-adjusted datasets (orange) and a set of 1,302 hvCpGs identified using unadjusted datasets (green). PC-adjusted: $lm(meth \sim 10\ PCs + age + sex)$, unadjusted: $lm(meth \sim age + sex)$.

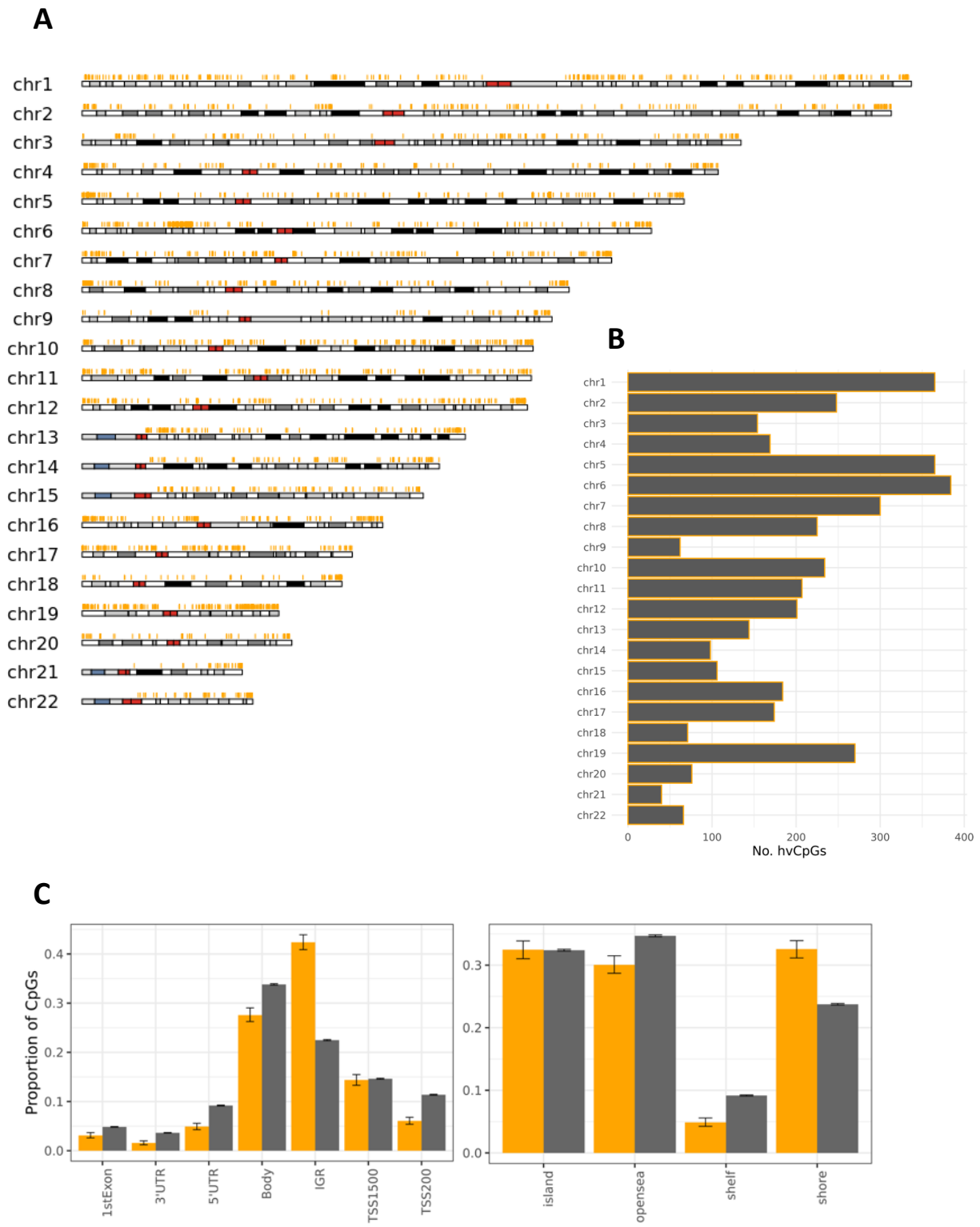


Figure S9. Genomic locations of hvCpGs. A) The positions of hvCpGs across autosomal hg19 chromosomes. **B)** The number of hvCpGs (orange) located on each chromosome. **C).** Genomic locations of hvCpGs (orange) and array background (grey) in relation to gene bodies (left) and CpG islands (right). UTR = untranslated region, IGR = intergenic region, TSS = transcription start site. Error bars are bootstrapped 95% confidence intervals.

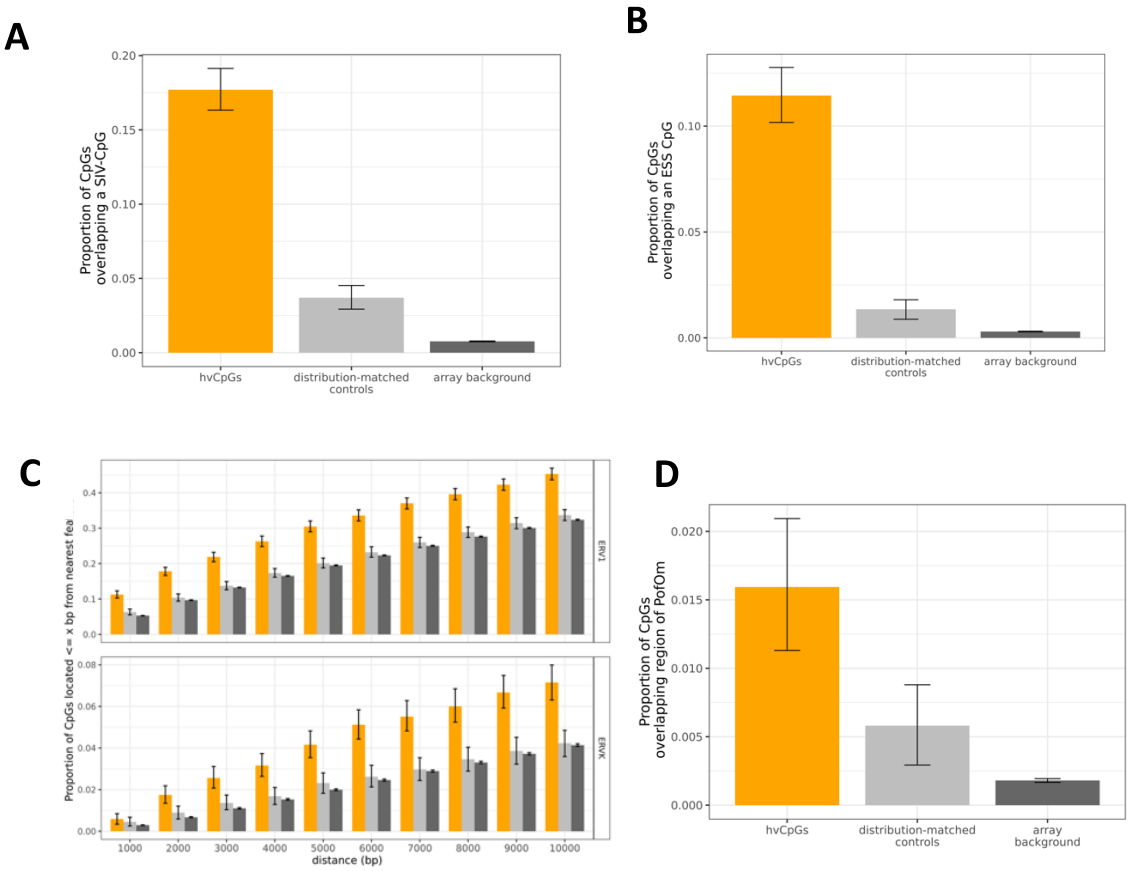


Figure S10. The association between hvCpGs and features suggestive of methylation establishment in the early embryo is maintained when using a de-clustered set of hvCpGs. The proportion of 2,178 de-clustered hvCpGs (in which no CpG is within 4 kb from another) and corresponding distribution-matched controls covered in the 'Blood_Cauc' dataset that **A)** overlap SIV-CpGs (4.8-fold enrichment relative to controls)**, B)** overlap ESS-CpGs (8.5-fold enrichment)**, C)** are located $\leq x$ bp from nearest ERV1 and ERVK transposable elements, and **D)** overlap regions of parent-of-origin specific methylation (PofOm) identified by Zink et al. (2018) (2.7-fold enrichment). See Supplementary Table 12 for further details.

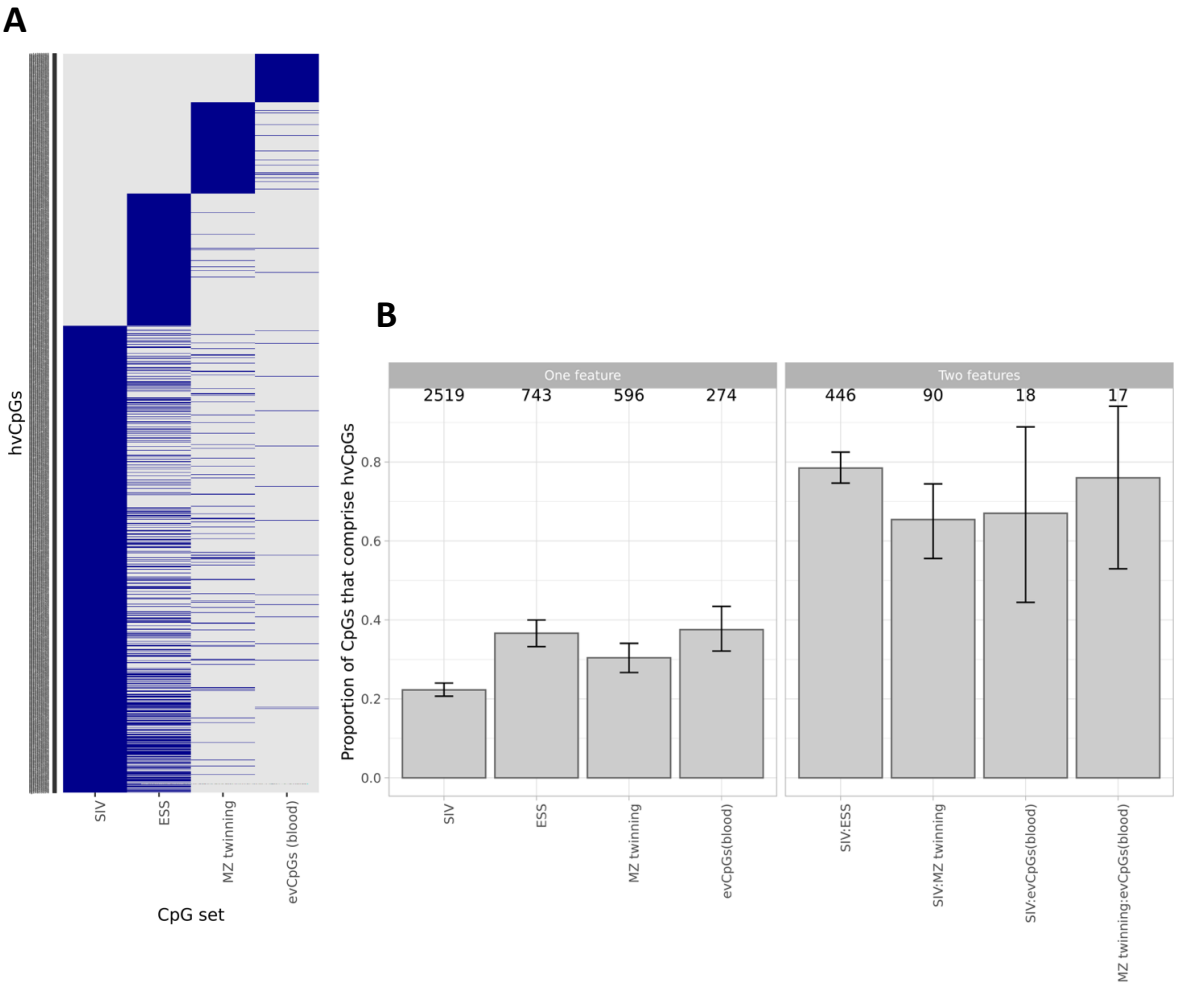


Figure S11. hvCpGs comprise a large proportion of loci that have been previously found to show more than one feature linked to variable methylation establishment in the early embryo. **A)** Heatmap showing the overlap of 1,574 (38%) hvCpGs with SIV, ESS, MZ-twinning CpGs and evCpGs. **B)** The proportion of previously identified CpGs showing one feature (left) or two features (right) linked to the early embryo that comprise hvCpGs. Number of CpGs overlapping the array background is given above each bar, requiring a minimum of 15 CpGs. Error bars are bootstrapped confidence intervals. SIV = systemic interindividual variation, ESS = epigenetic supersimilarity, evCpGs = equivalently variable CpGs.

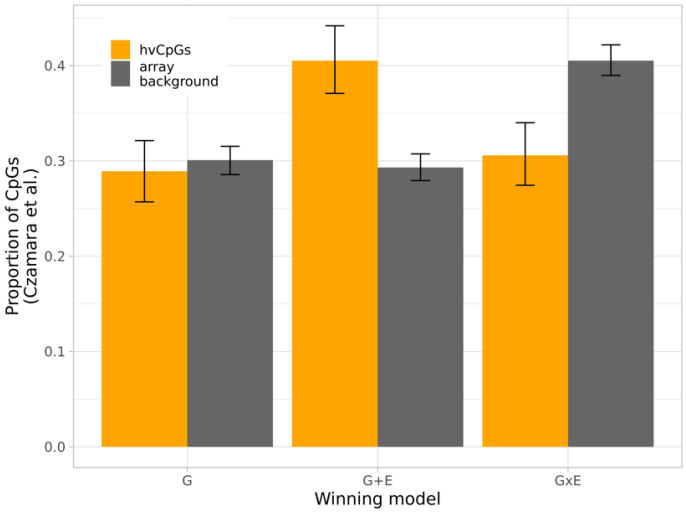


Figure S12. Methylation variability of hvCpGs in whole blood is best explained by G+E effects estimated by Czamara et al. (2019). The proportion of 747 hvCpGs and 3,644 array background CpGs covered by the variably methylated probes (VMPs) reported by Czamara et al. that are best explained by G, G + E and G x E effects. Error bars are bootstrapped 95% confidence intervals.

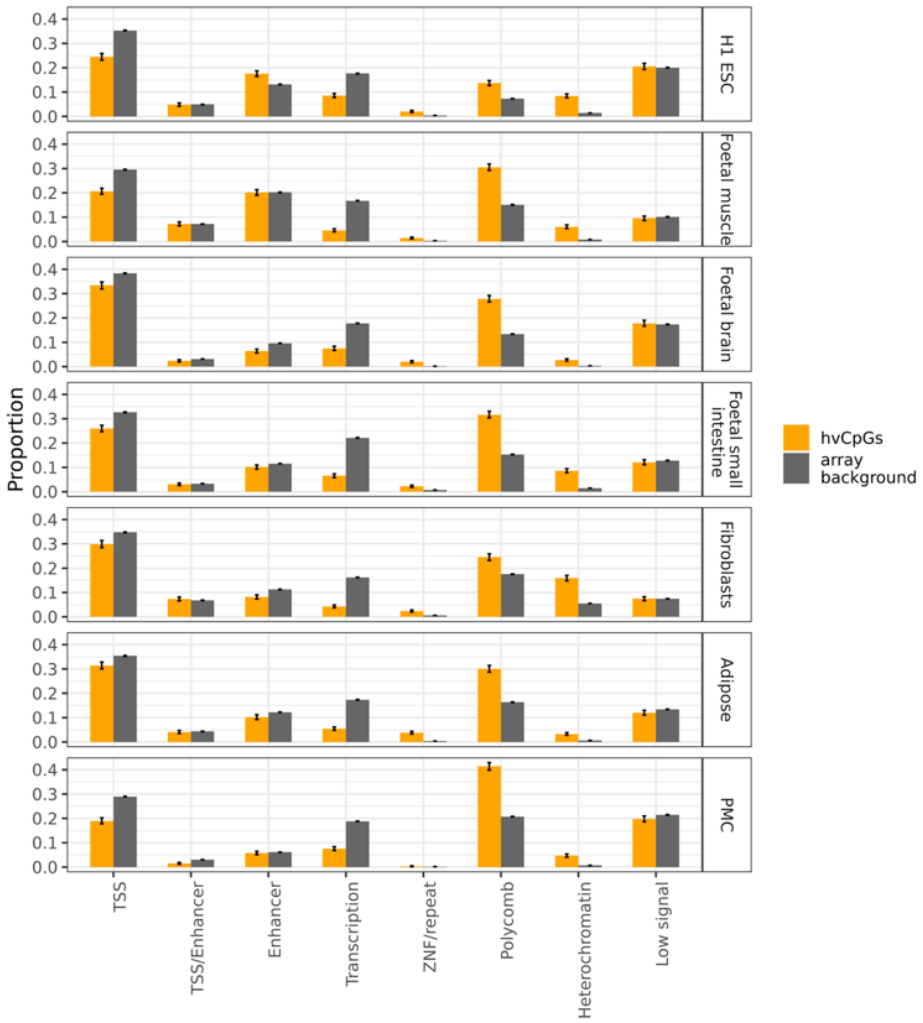


Figure S13. Functional annotation of hvCpGs using the ChromHMM model. The ChromHMM 15-state model using data from H1-hESC, foetal muscle, foetal brain, foetal small intestine, fibroblasts, adipose and primary mononucleocytes from the Roadmap Epigenomics Consortium. 15-states have been collapsed to 8 states for clarity (Supplementary Table 13). TSS = transcription start site. ZNF = zinc finger gene. Error bars indicate bootstrapped 95% confidence intervals.

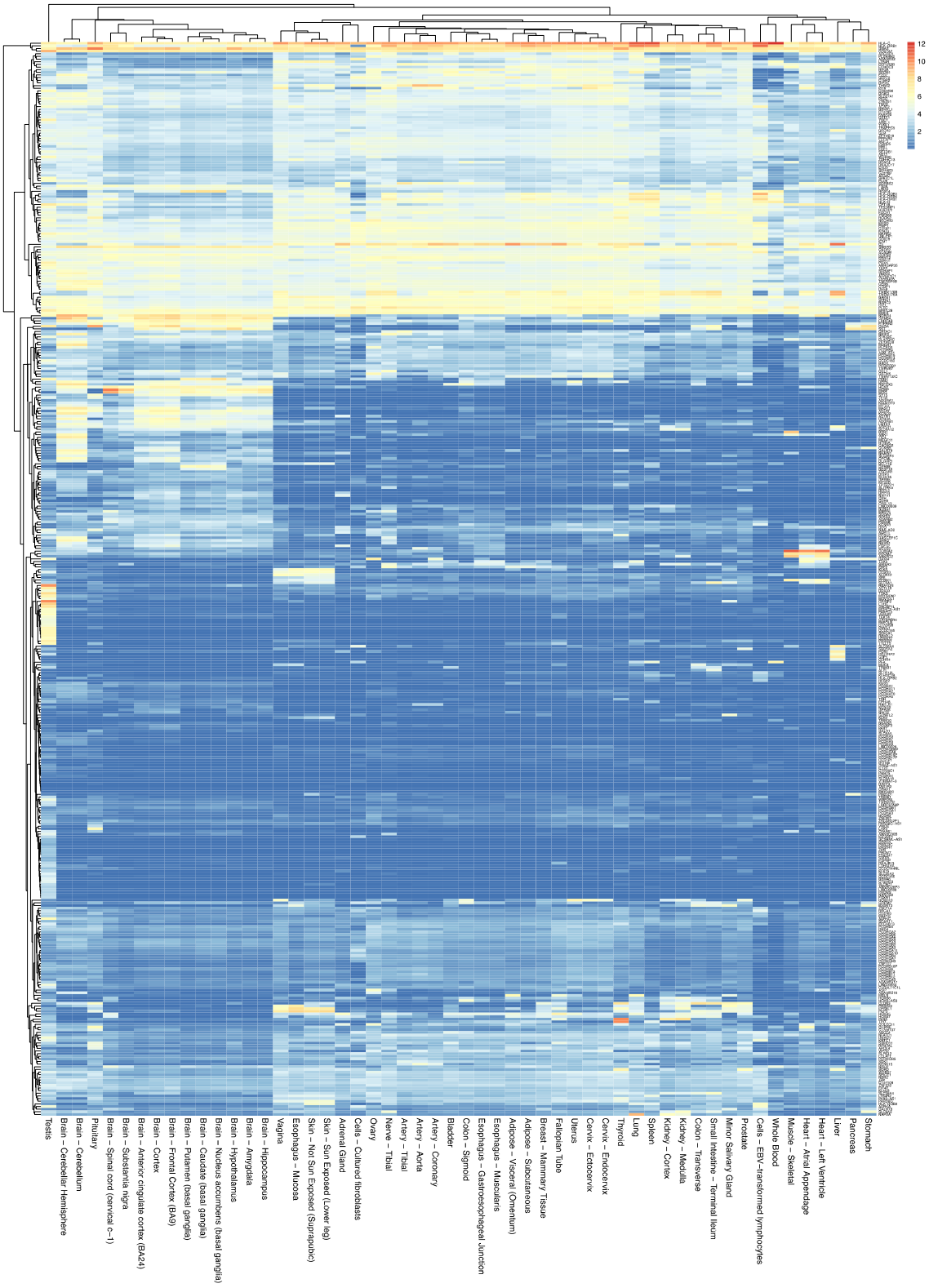


Figure S14. GTEx RNA expression of hvCpG-associated genes. Shown are the 416 genes that are annotated to hvCpG clusters (2 or more hvCpGs no more than 4kb apart). TPM = transcripts per million.

Appendix B

Supplementary Figures for Chapter 4

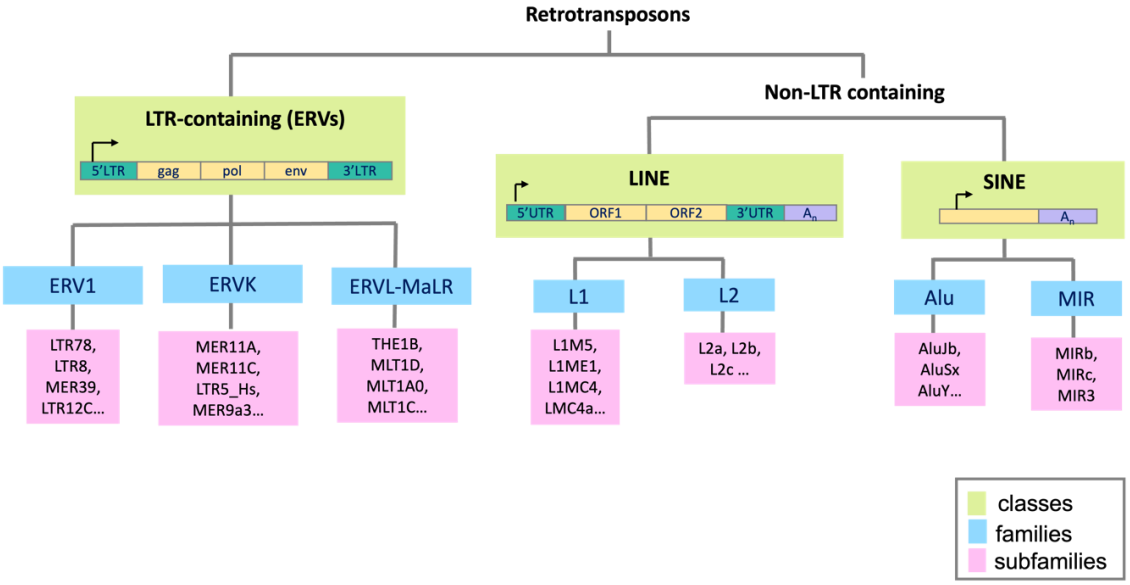


Figure B.1. Classification, nomenclature, and description of retrotransposons in the human genome. Retrotransposons transpose through an RNA intermediate that is reversed transcribed into DNA. Retrotransposons are split into Long Terminal Repeat (LTR)-containing retrotransposons and non-LTR-containing retrotransposons. In humans, LTR retrotransposons are largely represented by endogenous retroviruses (ERVs). A complete ERV element includes two flanking LTRs, a *gag* gene encoding the viral particle coat, a *pol* gene encoding a reverse transcriptase and an *env* gene encoding the envelope protein. The majority of LTR retrotransposons in the human genome exist as 'solo' LTRs, as the internal sequences can be lost due to homologous recombination between LTRs. The main ERV families in humans are ERV1, ERVK, ERVL, ERVL and ERVL-MaLR that are each related to different retrovirus groups. Non LTR-containing retrotransposons can be split into Long Interspersed Repeats (LINEs) and Short Interspersed Repeats (SINEs). LINEs have their own promoter, a 3' poly-A tail, and two open reading frames that encode an RNA-binding protein (ORF1) and an endonuclease with reverse transcriptase activity (ORF2). SINEs are 'non-autonomous', requiring the reverse transcriptase encoded by L1 elements in order to transpose. The main SINE families include Alu elements and Mammalian interspersed repeats (MIR). Whilst Alus are active, MIRs are an ancient family of inactive tRNA-derived SINEs. Each 'family' can be further split into subfamilies based on sequence similarity, as defined by RepeatMasker and DFAM. Examples of subfamilies are shown in the pink box. Note, for clarity, this schematic only includes the main ERV, LINE and SINE families.

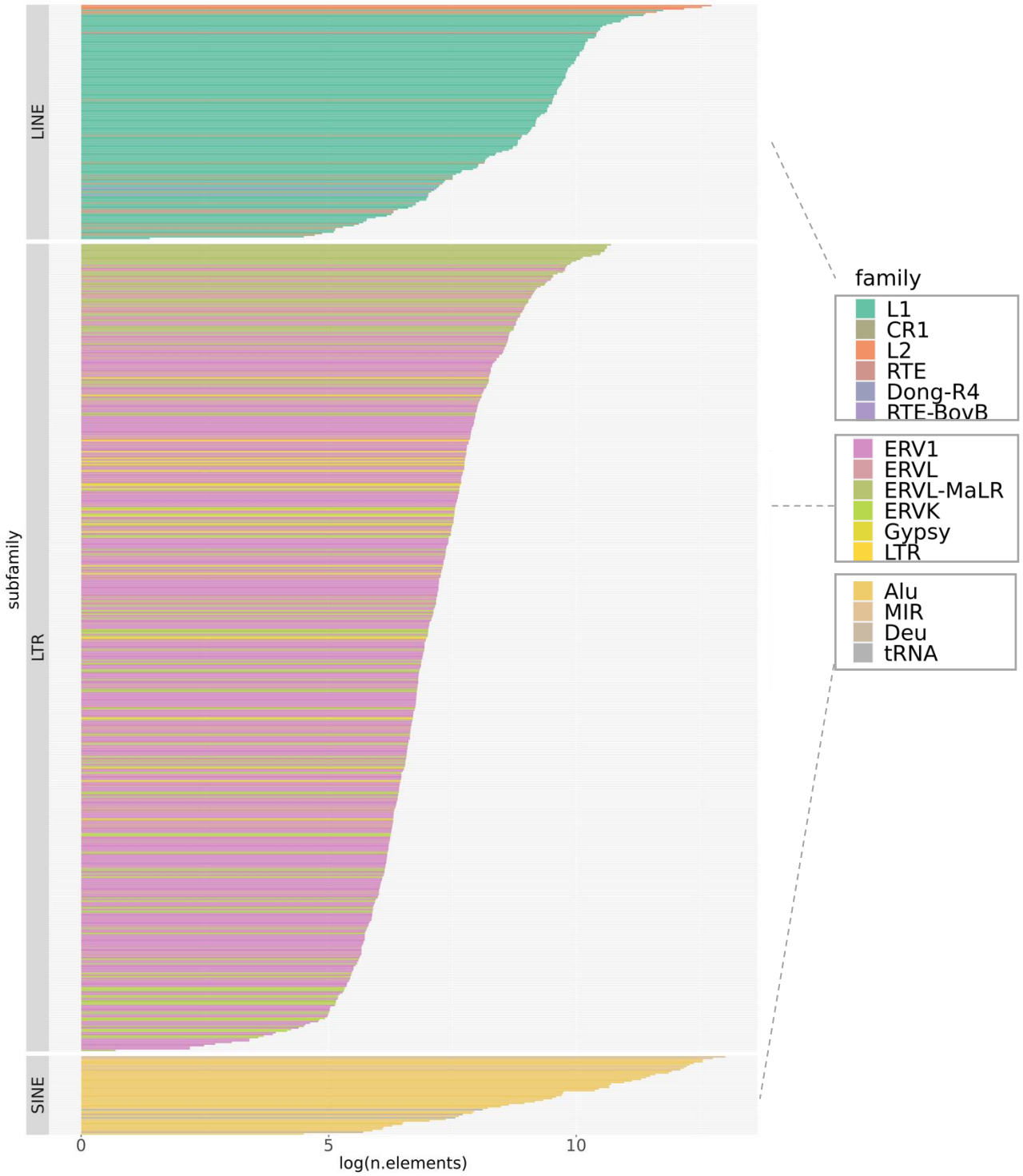


Figure B.2. Overview of retrotransposon elements downloaded from RepeatMasker. We downloaded hg19 coordinates for 698 subfamilies belonging to 16 'families'. X-axis gives the number of instances of each subfamily in the genome (referred to as 'elements'). Note this is a logarithmic scale. Each bar represents a subfamily and is coloured by the 'family' it belongs to (see legend). Facet labels indicate the retrotransposon class (Figure B.1).

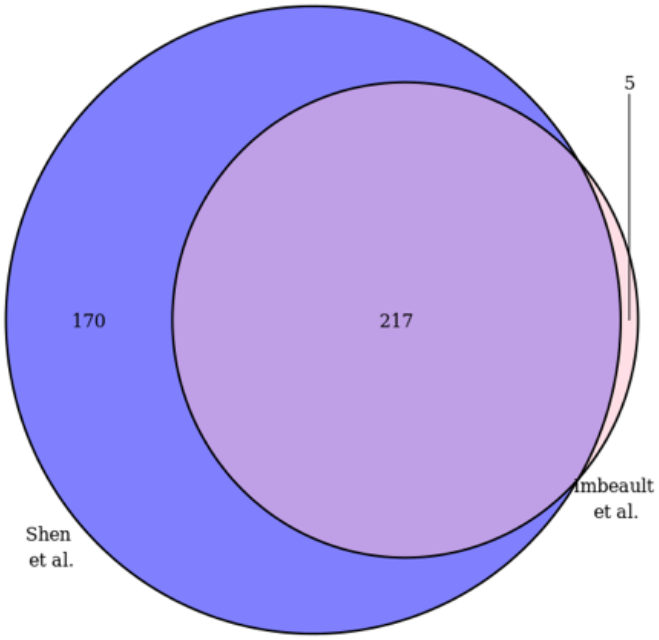


Figure B.3. Collation of KZFP genes used in this chapter. Venn diagram shows the overlap between 387 KZFP genes identified by Shen *et al.* 2021, and the 222 KZFPs with ChIP-exo data from Imbeault *et al.* (2017).

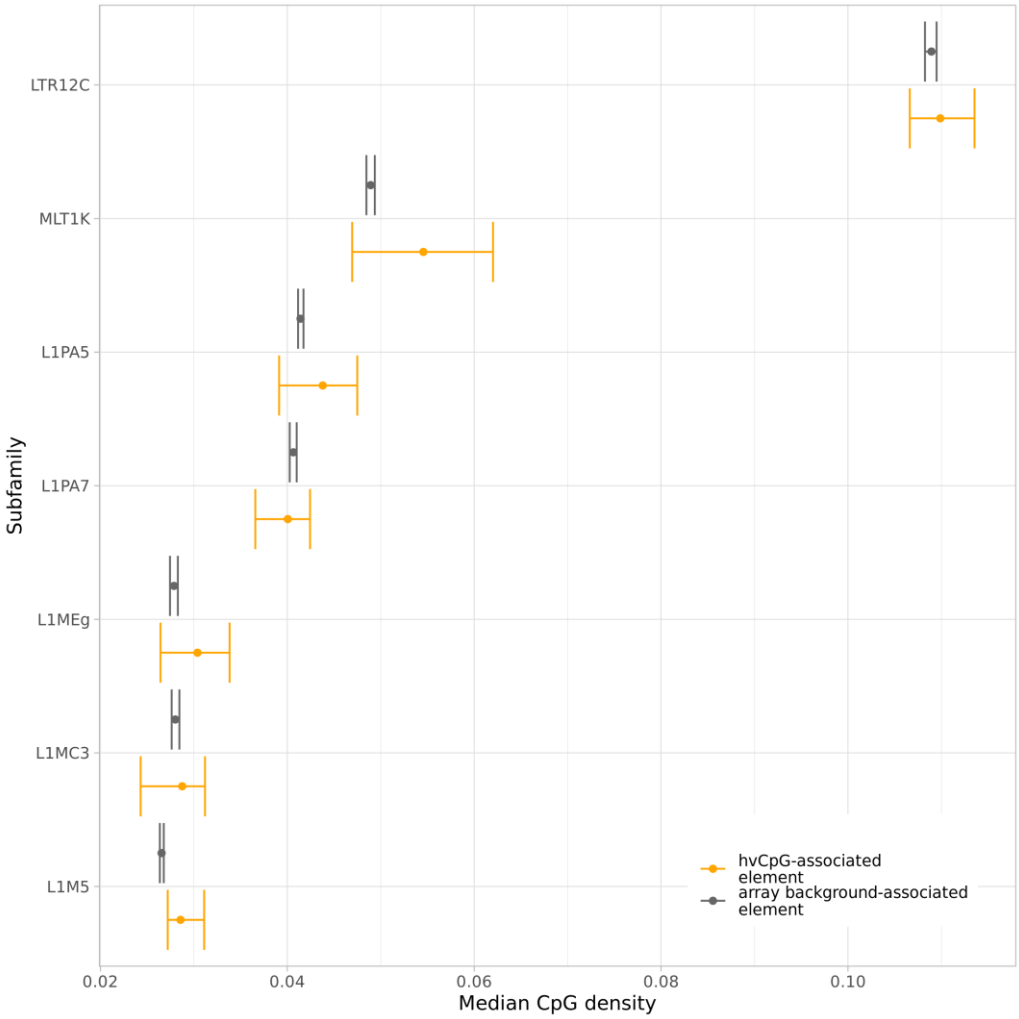


Figure B.4 Methylation variability is not linked to retrotransposon CpG density. Points indicate the median CpG density of subfamily elements within 3kb of an hvCpG (orange) and subfamily elements within 3kb of an array background CpG (grey). Error bars give bootstrapped confidence intervals with 1000 permutations. Note, we only considered the 7 subfamilies that hvCpGs were enriched for proximity to (Figure 4.2).

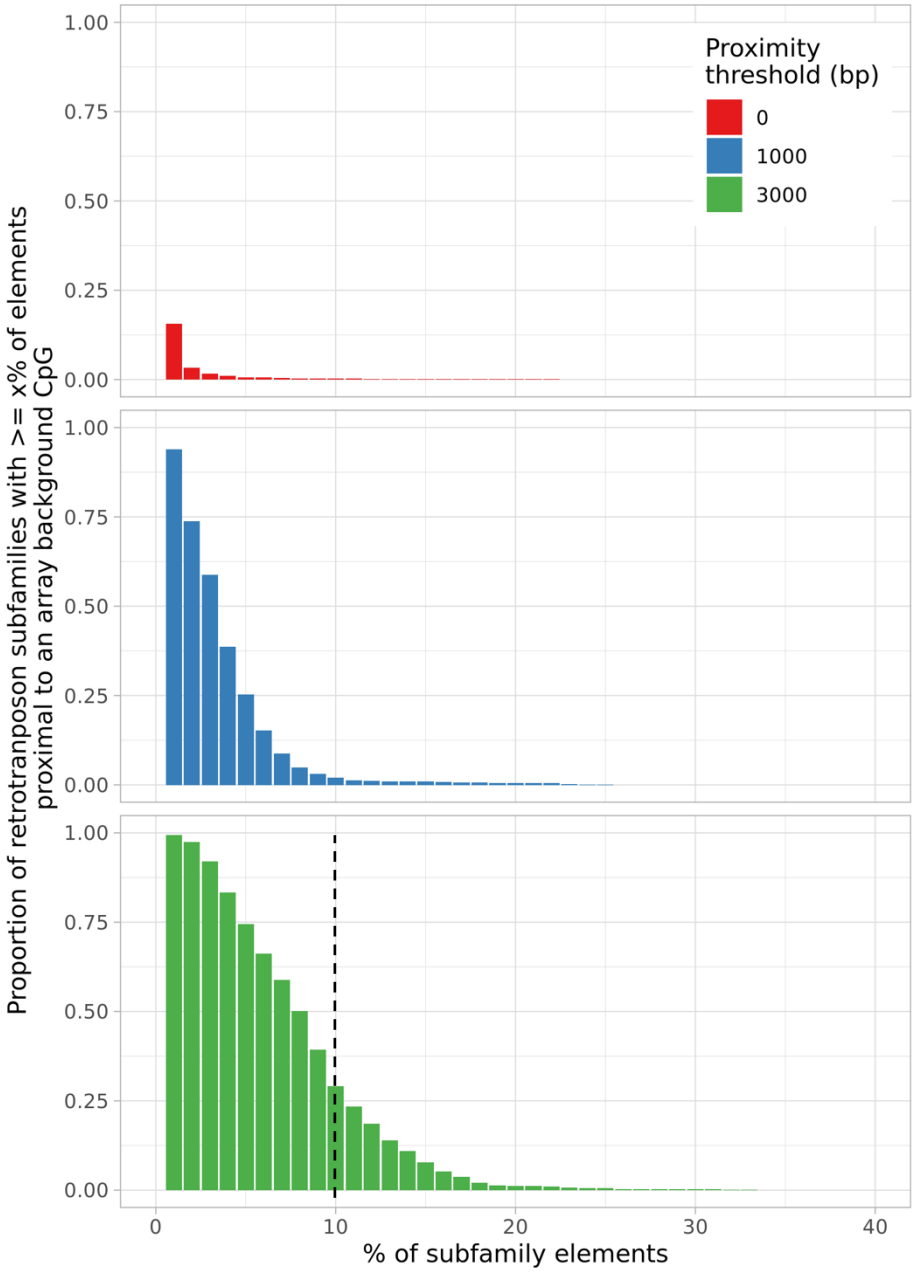


Figure B.5. Examination of coverage of retrotransposons on Illumina450K array. Y-axis shows the proportion of retrotransposon subfamilies that have at least x% of elements within 0kb (red), 1kb (blue) and 3kb (green) of an array background CpG. The dashed vertical line at x = 10% in bottom facet indicates the proportion of subfamilies that have at least 10% of elements within 3kb of an array background CpG, which we refer to as 'covered' subfamilies.

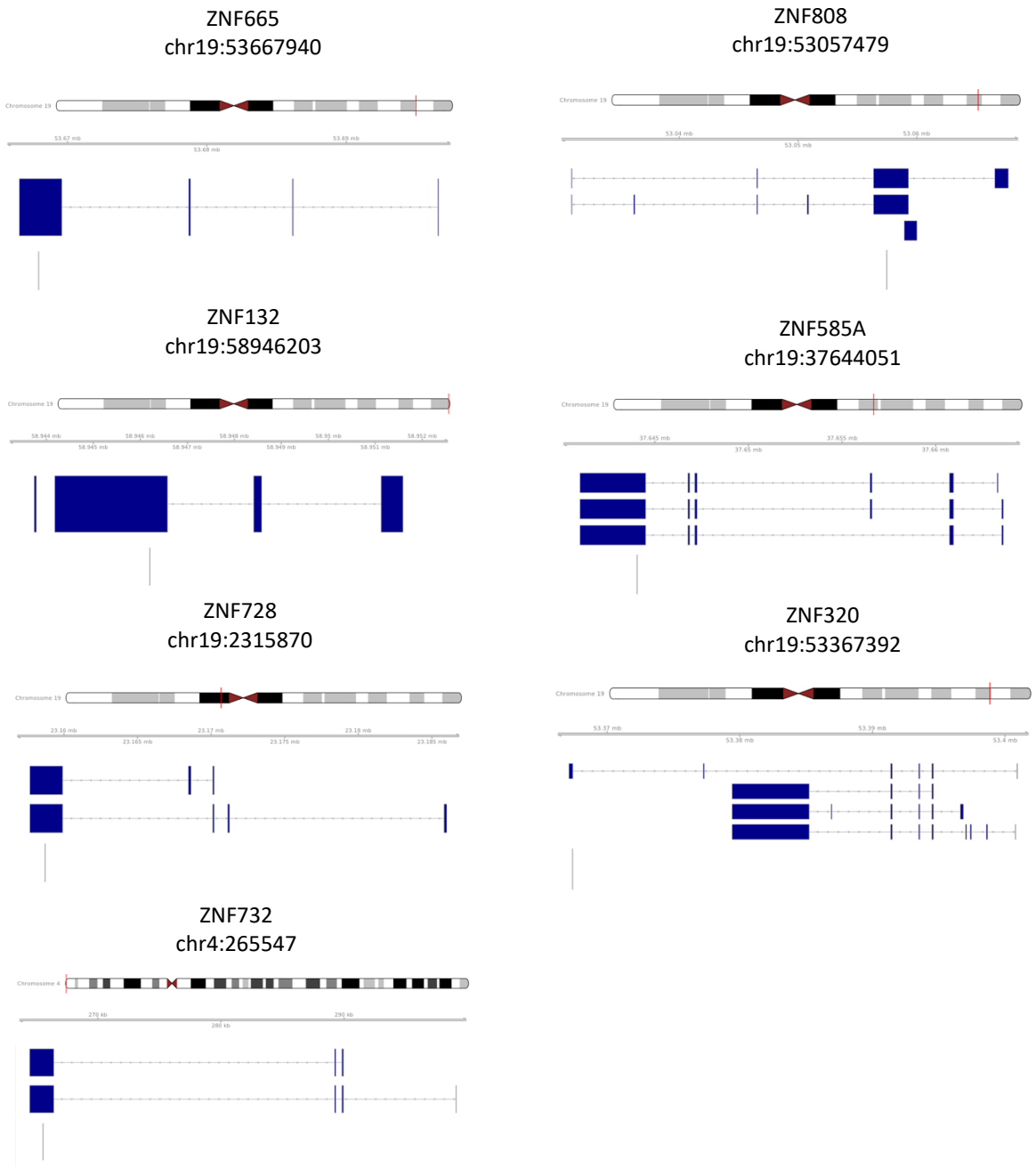


Figure B.6. Locations of the 7 hvCpG-associated KZFP mQTLs that are located in the protein coding sequence of a KZFP transcript. All mQTLs are located in the C-terminal zinc finger array DNA binding domain of KZFP genes with the exception of the ZNF320 mQTL which is located in the KRAB domain. Each track represents a gene transcript. Vertical line in bottom track represents the mQTL.

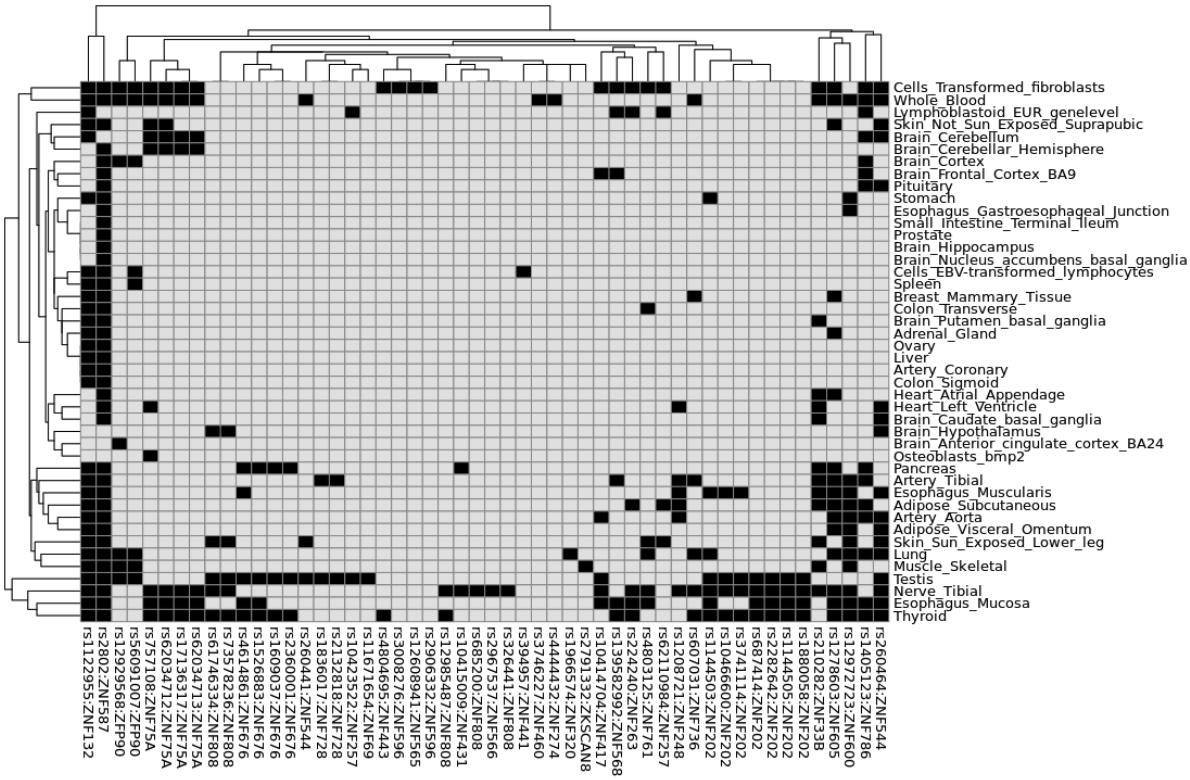


Figure B.7. The overlap between KZFP *trans* mQTLs and GTEx eQTLs across tissue types. Black squares indicate an overlap, whilst grey squares indicate the absence of an overlap. The x-axis gives the KZFP *trans* mQTL and the KZFP gene it is located in. The y-axis gives the GTEx tissue type.

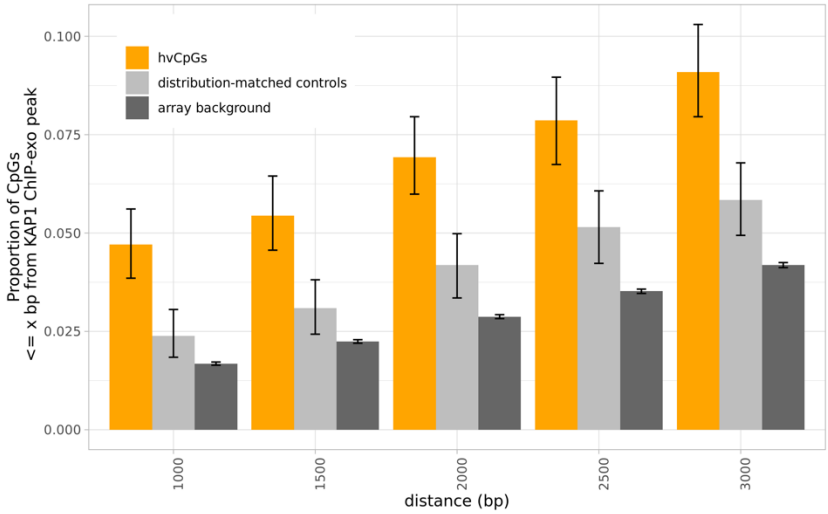


Figure B.8. De-clustered hvCpGs are enriched for proximal KAP1 binding sites relative to distribution-matched controls and array background CpGs. Bars show the proportion of 2388 de-clustered hvCpGs (orange), 2388 distribution-matched controls (light grey) and array background CpGs (dark grey) that are located ≤ x bp from the nearest KAP1 ChIP-exo peak in H1-ESCs. Error bars give bootstrapped 95% confidence intervals.

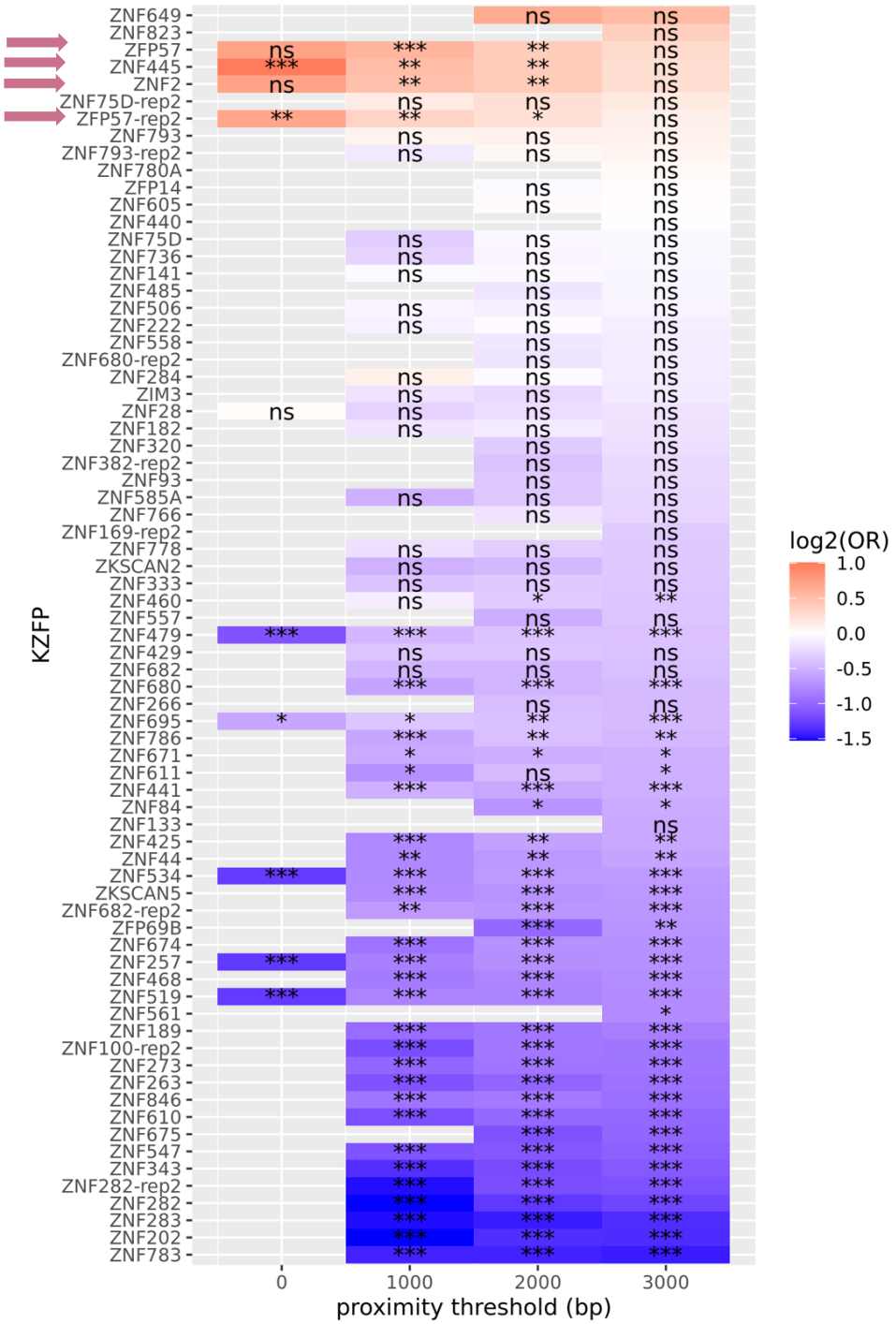
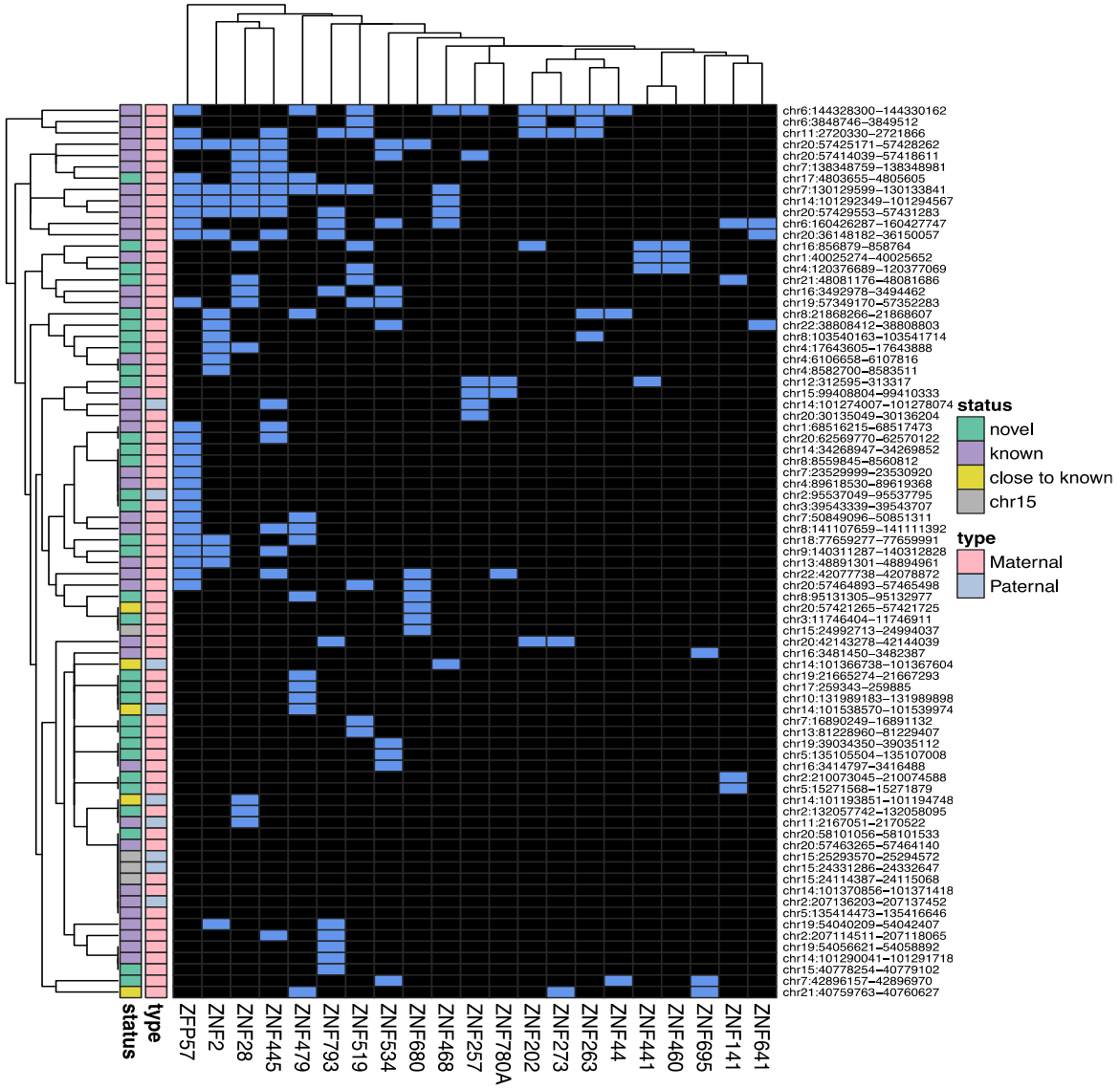
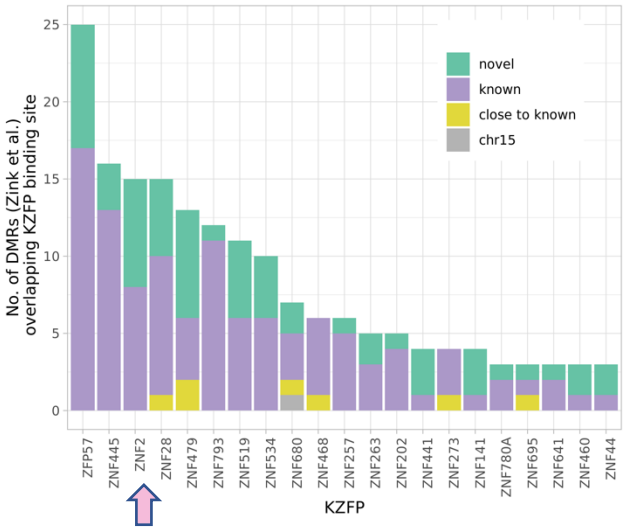


Figure B.9. hvCpGs are enriched for proximity to ChIP-exo peaks of certain KZFPs. FET were used to examine the enrichment of 2,640 de-clustered hvCpGs within x bp ('proximity threshold') from the nearest KZFP binding site (for each KZFP ChIP-exo experiment) relative to array background CpGs. Tile colour indicates enrichment odds ratios and asterisks give FET p-values adjusted for the false discovery rate (FDR): $q < 0.001$(***), $q < 0.01$ (**), $q < 0.05$ (*), 'ns' = not significant. For clarity, we have only shown odds ratios and q-values for KZFPs that have at least 1% of hvCpGs within the proximity threshold (x-axis) for a KZFP binding site. The blank tiles indicate where this threshold was not met. Pink arrows highlight ZFP57, ZNF445 and ZFN2. See Supplementary Table B.4 for further details.

A


status
novel
known
close to known
chr15
type
Maternal
Paternal
status
type
ZFP57
ZNF2
ZNF28
ZNF445
ZNF479
ZNF793
ZNF519
ZNF534
ZNF680
ZNF468
ZNF257
ZNF780A
ZNF202
ZNF273
ZNF263
ZNF44
ZNF441
ZNF460
ZNF695
ZNF141
ZNF641

B


No. of DMRs (Zink et al.) overlapping KZFP binding site
0
5
10
15
20
25
novel
known
close to known
chr15
ZFP57
ZNF445
ZNF2
ZNF28
ZNF479
ZNF793
ZNF519
ZNF534
ZNF680
ZNF468
ZNF257
ZNF263
ZNF202
ZNF441
ZNF273
ZNF141
ZNF780A
ZNF695
ZNF641
ZNF460
ZNF44
KZFP

C

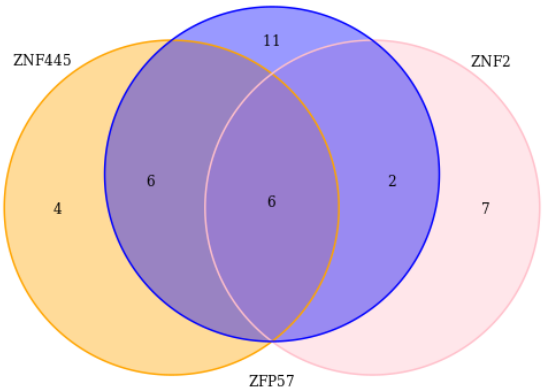

ZNF445
ZNF2
ZFP57
11
6
6
2
4
7

Figure B.10. The overlap between KZFP ChIP-exo peaks and regions of parent-of-origin specific methylation ('PofOm DMRs'). A) Heatmap showing the overlap between binding sites for 21 KZFPs and 79 PofOm DMRs identified by Zink et al. Y-axis gives the hg19 genomic coordinates for the 79 PofOm DMRs that overlap at least one KZFP ChIP-exo peak. Black squares indicate lack of binding, whilst blue squares indicate the overlap between a PofOm DMR and a KZFP ChIP-exo peak. Maternal DMRs are highlighted in pink and paternal DMRs are highlighted in blue. DMRs are also coloured by their known 'status' reported by Zink et al. DMRs defined as 'novel' had not been identified in previous screens. **B)** Bar plot showing the number of PofOm DMRs overlapping a KZFP ChIP-exo peak, highlighted by their 'status'. Pink arrow highlights ZNF2, which binds the third highest proportion of DMRs after ZFP57 and ZNF445. **C)** The overlap of PofOm DMRs bound by ZNF2, ZFP57 and ZNF445. ZNF2 has its own unique targets and also shares targets with ZFN445 and ZFP57.

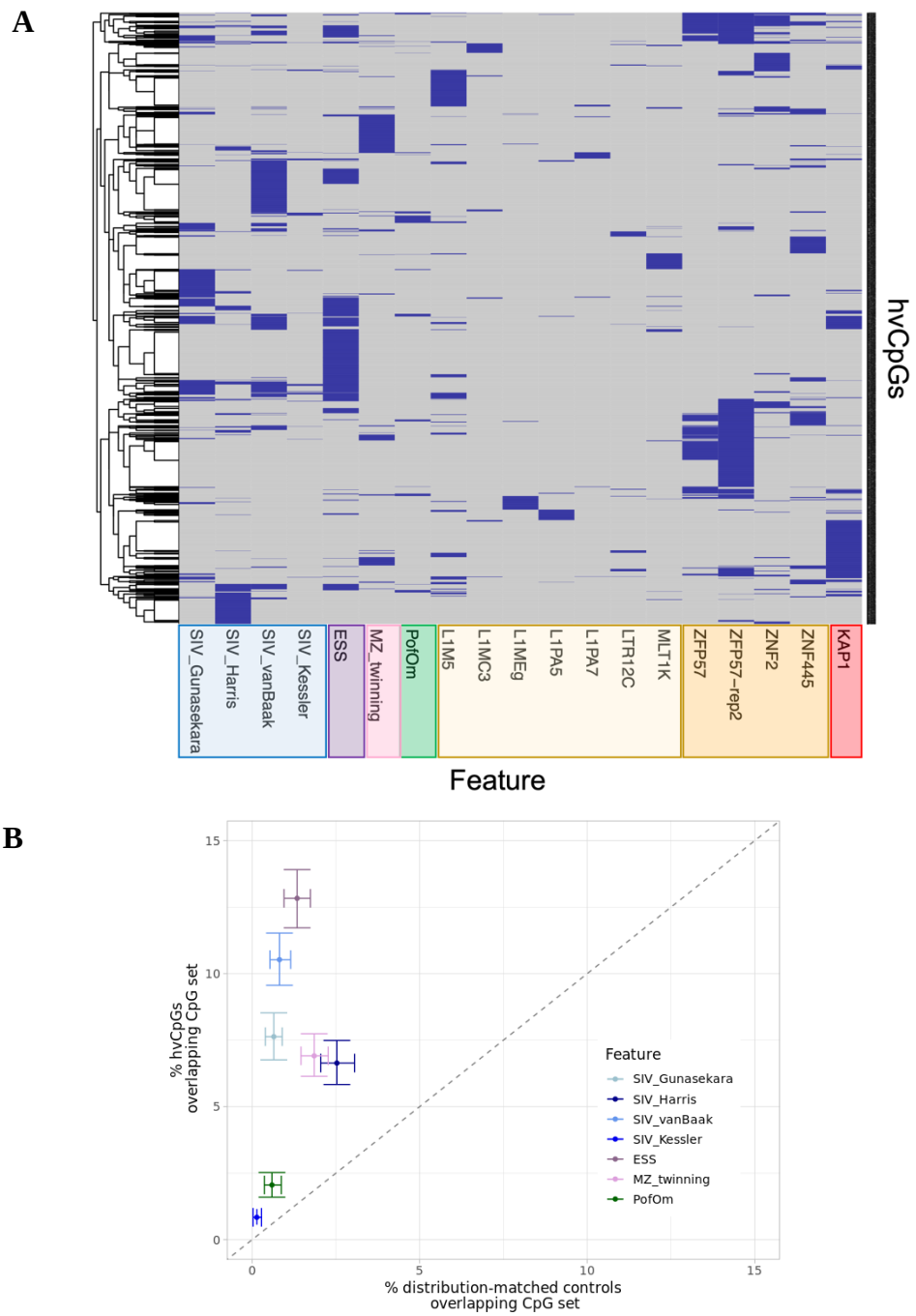


Figure B.11. Heatmap showing the overlaps between hvCpGs and notable features. A) Blue tiles indicate hvCpGs that overlap SIV loci identified in four independent studies ('SIV_Gunasekara', 'SIV_Harris', 'SIV_vanBaak' and 'SIV_Kessler'), ESS CpGs, MZ twinning CpGs and regions of PofOm (see Chapter 3, Supplementary Table 12). Blue tiles are also used to indicate hvCpGs that are proximal (within 3kb) to specific retrotransposon subfamilies (L1M5, L1MC3, L1MEg, L1PA5, L1PA7, LTR12C, MLT1K, see Table B.1), KZFP binding sites (ZFP57, ZNF2, ZNF445, see Table B.4) and KAP1 binding sites (see Figure 4.5A). To assess whether hvCpGs fall into subsets based on the features they show, a clustering algorithm was applied using default parameters in the 'pheatmap' R package. **B)** Extended version of Figure 3A (Chapter 3) showing hvCpG enrichments for notable CpG sets relative to distribution-matched controls. In this extended version, SIV CpGs have been stratified according to the study in which they were identified.

Supplementary Tables for Chapter 4

Supplementary Table B.1. Enrichment and depletion for proximal retrotransposon subfamilies at hvCpGs relative to the Illumina450K array background. Only the subfamilies with at least 1% of hvCpGs within 3kb of a subfamily element are shown. The subfamilies highlighted in green have FDR-adjusted p-value < 0.05 and correspond to those shown in Figure 4.2

subfamily	family	class	Prop. hvCpGs	Prop. array_back	Odds Ratio	FET p-value	FDR-adjusted p-value
L1MEg	L1	LINE	0.017	0.008	2.195	3.868E-06	1.352E-03
MIR	MIR	SINE	0.265	0.302	0.833	3.027E-05	6.090E-03
LTR12C	ERV1	LTR	0.013	0.006	2.243	3.723E-05	6.090E-03
MIR3	MIR	SINE	0.180	0.212	0.815	4.357E-05	6.090E-03
MIRb	MIR	SINE	0.335	0.371	0.855	1.437E-04	1.435E-02
L1PA5	L1	LINE	0.013	0.007	2.001	2.357E-04	1.697E-02
L1M5	L1	LINE	0.075	0.058	1.318	3.134E-04	1.826E-02
L2c	L2	LINE	0.177	0.204	0.836	4.045E-04	1.913E-02
MIRc	MIR	SINE	0.194	0.222	0.844	5.020E-04	2.064E-02
AluJb	Alu	SINE	0.214	0.243	0.850	5.339E-04	2.073E-02
L2b	L2	LINE	0.159	0.184	0.836	7.280E-04	2.544E-02
L1PA7	L1	LINE	0.011	0.005	1.981	1.154E-03	3.361E-02
L1MC3	L1	LINE	0.015	0.009	1.710	1.718E-03	4.188E-02
MLT1K	ERVL-MaLR	LTR	0.029	0.020	1.452	2.243E-03	4.899E-02
L1MA9	L1	LINE	0.013	0.008	1.740	2.370E-03	5.021E-02
L1MC1	L1	LINE	0.014	0.008	1.690	3.502E-03	6.616E-02
MLT1C	ERVL-MaLR	LTR	0.035	0.026	1.373	3.840E-03	6.882E-02
L1MC4	L1	LINE	0.034	0.026	1.348	6.744E-03	1.025E-01
THE1D	ERVL-MaLR	LTR	0.012	0.007	1.647	1.028E-02	1.272E-01
L1M4	L1	LINE	0.019	0.013	1.462	1.037E-02	1.272E-01
MSTB	ERVL-MaLR	LTR	0.010	0.006	1.668	1.215E-02	1.396E-01
L1MC4a	L1	LINE	0.036	0.028	1.316	1.218E-02	1.396E-01
L1MEc	L1	LINE	0.011	0.007	1.630	1.325E-02	1.447E-01
FLAM_A	Alu	SINE	0.026	0.034	0.742	1.356E-02	1.458E-01
MLT1D	ERVL-MaLR	LTR	0.028	0.022	1.304	2.767E-02	2.146E-01
L2a	L2	LINE	0.214	0.231	0.902	3.127E-02	2.254E-01
FRAM	Alu	SINE	0.016	0.021	0.717	3.637E-02	2.492E-01
AluSx3	Alu	SINE	0.056	0.065	0.847	5.289E-02	2.936E-01
L1MB3	L1	LINE	0.028	0.022	1.265	5.332E-02	2.936E-01
AluSx	Alu	SINE	0.239	0.255	0.916	5.696E-02	3.059E-01
L1MD	L1	LINE	0.011	0.007	1.447	6.609E-02	3.253E-01
L1MB5	L1	LINE	0.016	0.012	1.349	6.695E-02	3.272E-01
HAL1	L1	LINE	0.022	0.018	1.276	7.430E-02	3.463E-01
FLAM_C	Alu	SINE	0.055	0.063	0.863	9.041E-02	3.854E-01

L1MC	L1	LINE	0.018	0.014	1.282	9.554E-02	3.939E-01
MLT1H	ERVL-MaLR	LTR	0.011	0.008	1.375	9.920E-02	3.952E-01
L1ME2	L1	LINE	0.011	0.009	1.341	1.105E-01	4.050E-01
MLT1A	ERVL-MaLR	LTR	0.012	0.009	1.321	1.197E-01	4.143E-01
THE1B	ERVL-MaLR	LTR	0.015	0.012	1.286	1.238E-01	4.223E-01
MLT1A0	ERVL-MaLR	LTR	0.022	0.018	1.230	1.250E-01	4.241E-01
L1ME3	L1	LINE	0.012	0.009	1.311	1.266E-01	4.275E-01
AluSx1	Alu	SINE	0.219	0.231	0.931	1.381E-01	4.536E-01
L1ME2z	L1	LINE	0.011	0.009	1.291	1.733E-01	5.067E-01
MLT1B	ERVL-MaLR	LTR	0.024	0.020	1.178	2.125E-01	5.542E-01
AluSz6	Alu	SINE	0.078	0.085	0.917	2.476E-01	6.050E-01
AluYc	Alu	SINE	0.020	0.017	1.174	2.543E-01	6.066E-01
MLT1J	ERVL-MaLR	LTR	0.018	0.016	1.169	2.701E-01	6.219E-01
L1MB7	L1	LINE	0.027	0.023	1.144	2.705E-01	6.219E-01
L1ME1	L1	LINE	0.023	0.020	1.150	2.928E-01	6.519E-01
AluY	Alu	SINE	0.211	0.202	1.052	2.960E-01	6.548E-01
MLT1I	ERVL-MaLR	LTR	0.011	0.010	1.198	3.136E-01	6.680E-01
L3b	CR1	LINE	0.011	0.014	0.820	3.151E-01	6.680E-01
AluSq	Alu	SINE	0.048	0.053	0.908	3.163E-01	6.680E-01
AluSg	Alu	SINE	0.090	0.095	0.934	3.352E-01	6.912E-01
MSTA	ERVL-MaLR	LTR	0.013	0.011	1.169	3.496E-01	7.059E-01
AluSc	Alu	SINE	0.065	0.069	0.935	4.183E-01	7.736E-01
L1ME4a	L1	LINE	0.051	0.048	1.070	4.373E-01	7.984E-01
L4	RTE	LINE	0.022	0.020	1.104	4.387E-01	7.984E-01
AluSc5	Alu	SINE	0.015	0.014	1.101	5.035E-01	8.542E-01
AluSx4	Alu	SINE	0.011	0.013	0.868	5.396E-01	8.731E-01
L1MB8	L1	LINE	0.016	0.015	1.088	5.691E-01	9.025E-01
L2	L2	LINE	0.081	0.078	1.032	6.628E-01	9.857E-01
AluSg7	Alu	SINE	0.021	0.020	1.055	6.738E-01	9.978E-01
L1MC5	L1	LINE	0.022	0.023	0.946	7.455E-01	1.000E+00
AluSq2	Alu	SINE	0.121	0.119	1.017	7.634E-01	1.000E+00
L1ME3C	L1	LINE	0.011	0.010	1.053	7.726E-01	1.000E+00
L1ME3A	L1	LINE	0.013	0.014	0.940	8.009E-01	1.000E+00
AluSz	Alu	SINE	0.192	0.191	1.011	8.231E-01	1.000E+00
L3	CR1	LINE	0.059	0.060	0.981	8.695E-01	1.000E+00
AluSg4	Alu	SINE	0.019	0.018	1.011	8.845E-01	1.000E+00
AluJr	Alu	SINE	0.155	0.154	1.007	8.923E-01	1.000E+00
AluSp	Alu	SINE	0.115	0.115	1.006	9.267E-01	1.000E+00
AluJr4	Alu	SINE	0.033	0.033	0.987	9.566E-01	1.000E+00
AluJo	Alu	SINE	0.152	0.153	0.995	9.567E-01	1.000E+00
AluSc8	Alu	SINE	0.047	0.047	1.004	9.632E-01	1.000E+00

Supplementary Table B.2. KZFP mQTLs that are associated with hvCpGs in the GoDMC meta-analysis. hvCpGs with the same 'cluster id' are within 4kb of one another and refer to the hvCpG clusters introduced in Chapter 3.

mQTL	KZFP gene	hvCpG	CpG chr	CpG pos	cluster.id
chr1:227893320:SNP	ZNF678	cg03689146	chr7	52341810	1136
chr4:265547:SNP	ZNF732	cg20276780	chr1	41303356	235
chr4:374694:SNP	ZNF141	cg14316629	chr6	168436353	895
chr6:28108788:SNP	ZKSCAN8	cg02174092	chr10	43846539	1361
chr6:28216427:INDEL	ZKSCAN4	cg11383134	chr6	29648590	962
chr6:28216427:INDEL	ZKSCAN4	cg12644888	chr6	29648360	962
chr6:28216427:INDEL	ZKSCAN4	cg16885113	chr6	29648507	962
chr6:29644040:SNP	ZFP57	cg01817393	chr20	57427642	2552
chr7:63706590:SNP	ZNF679	cg18041277	chr8	57030824	1272
chr7:63771849:SNP	ZNF736	cg08415582	chr8	57030523	1272
chr7:63774243:SNP	ZNF736	cg05332308	chr16	62442964	2046
chr7:148779400:SNP	ZNF786	cg20288565	chr1	235805403	163
chr7:148894646:SNP	ZNF282	cg07769957	chr19	46443657	2346
chr8:189364:SNP	ZNF596	cg22034735	chr8	1321727	1276
chr8:194943:SNP	ZNF596	cg03128011	chr8	1321333	1276
chr8:194943:SNP	ZNF596	cg20088245	chr8	1321375	1276
chr9:97028679:SNP	ZNF169	cg03269218	chr10	96990700	1412
chr9:97028679:SNP	ZNF169	cg15233961	chr10	96990543	1412
chr9:97051328:SNP	ZNF169	cg03269218	chr10	96990700	1412
chr9:97051328:SNP	ZNF169	cg15233961	chr10	96990543	1412
chr9:114287818:SNP	ZNF483	cg03961283	chr1	223566761	140
chr9:114287818:SNP	ZNF483	cg04926766	chr1	223567002	140
chr9:114287818:SNP	ZNF483	cg13963793	chr1	223567173	140
chr9:114287818:SNP	ZNF483	cg22372182	chr1	223566669	140
chr9:114287818:SNP	ZNF483	cg26603179	chr1	223566646	140
chr9:114287818:SNP	ZNF483	cg27202708	chr1	223566709	140
chr9:114309210:SNP	ZNF483	cg05333568	chr1	223566794	140
chr10:38114707:SNP	ZNF248	cg16482344	chr10	42970987	1516
chr10:43087649:SNP	ZNF33B	cg02750792	chr3	39543967	435
chr10:43087649:SNP	ZNF33B	cg03054684	chr3	39543515	435
chr11:123598620:SNP	ZNF202	cg03689146	chr7	52341810	1136
chr11:123598620:SNP	ZNF202	cg26617364	chr7	52342124	1136
chr11:123598620:SNP	ZNF202	cg20581874	chr14	21191754	1924
chr11:123598620:SNP	ZNF202	cg26393629	chr14	21191859	1924
chr11:123601916:SNP	ZNF202	cg03248761	chr13	81228965	1848
chr11:123602199:SNP	ZNF202	cg21540359	chr7	52341648	1136
chr11:123602657:SNP	ZNF202	cg15078822	chr7	52341469	1136
chr11:123602945:SNP	ZNF202	cg21266330	chr4	72897698	584
chr11:123605723:INDEL	ZNF202	cg03600259	chr6	164052180	880
chr11:123609168:SNP	ZNF202	cg23513591	chr13	81229232	1848
chr11:123609364:SNP	ZNF202	cg16919449	chr10	36054185	1510
chr12:133498660:SNP	ZNF605	cg06840743	chr7	155150876	1053
chr12:133730500:SNP	ZNF268	cg23226914	chr1	247537224	179
chr12:133736546:SNP	ZNF268	cg11207515	chr7	146904205	1040
chr16:3316340:SNP	ZNF263	cg27228433	chr16	71475716	2051
chr16:3356441:SNP	ZNF75A	cg03454541	chr7	5111488	1094
chr16:3359124:SNP	ZNF75A	cg27571374	chr19	2889256	2437
chr16:3362017:SNP	ZNF75A	cg14981253	chr7	5112642	1094
chr16:3362017:SNP	ZNF75A	cg08684893	chr19	44616615	2341

chr16:3366513:SNP	ZNF75A	cg25786696	chr19	44406241	2340
chr16:3366513:SNP	ZNF75A	cg07504200	chr19	2889208	2437
chr16:31923839:SNP	ZNF267	cg00876757	chr16	33483591	2153
chr16:31923839:SNP	ZNF267	cg08704196	chr16	34787392	2162
chr16:68568537:SNP	ZFP90	cg04522432	chr6	29618347	960
chr16:68596689:SNP	ZFP90	cg18014983	chr6	29618382	960
chr19:9642388:SNP	ZNF426	cg17248924	chr5	110062384	776
chr19:9647670:SNP	ZNF426	cg06429214	chr5	110062417	776
chr19:9647670:SNP	ZNF426	cg23279021	chr5	110062343	776
chr19:11881451:SNP	ZNF441	cg13344095	chr6	26233352	923
chr19:12014355:SNP	ZNF69	cg13713218	chr9	100000033	1324
chr19:12014355:SNP	ZNF69	cg14625636	chr9	100000026	1324
chr19:12036566:SNP	ZNF763	cg19597498	chr3	96495651	461
chr19:12054516:SNP	ZNF763	cg07420274	chr2	171676306	247
chr19:12058536:SNP	ZNF763	cg04358463	chr19	56215104	2409
chr19:12058536:SNP	ZNF763	cg10606127	chr19	56215426	2409
chr19:12058536:SNP	ZNF763	cg19196221	chr19	56215575	2409
chr19:12140625:SNP	ZNF433	cg10555800	chr1	168356606	105
chr19:12179712:SNP	ZNF844	cg23475489	chr8	16142199	1238
chr19:12353395:SNP	ZNF44	cg01894875	chr1	2121521	25
chr19:12353395:SNP	ZNF44	cg05337761	chr1	2121349	25
chr19:12353395:SNP	ZNF44	cg08200543	chr1	2121449	25
chr19:12496934:SNP	ZNF799	cg02633068	chr15	27083996	1970
chr19:12544238:SNP	ZNF443	cg09929763	chr11	358332	1567
chr19:12588119:SNP	ZNF709	cg17125837	chr11	358595	1567
chr19:12588119:SNP	ZNF709	cg18815398	chr20	61506981	2559
chr19:19910658:INDEL	ZNF506	cg21442626	chr8	104032819	1301
chr19:20121260:SNP	ZNF682	cg22150661	chr8	104032860	1301
chr19:20129977:SNP	ZNF682	cg14877637	chr8	104032876	1301
chr19:21351758:SNP	ZNF431	cg20909645	chr8	99985049	1299
chr19:21998816:SNP	ZNF43	cg08767686	chr5	51426	633
chr19:22127841:SNP	ZNF208	cg23245905	chr10	99735133	1417
chr19:22192425:SNP	ZNF208	cg02026141	chr10	23492470	1492
chr19:22243448:SNP	ZNF257	cg26579653	chr14	74319300	1921
chr19:22252332:SNP	ZNF257	cg03379552	chr10	99734662	1417
chr19:22252430:INDEL	ZNF257	cg09834706	chr2	39470743	339
chr19:22366973:SNP	ZNF676	cg09945482	chr18	12777974	2328
chr19:22373841:SNP	ZNF676	cg15146462	chr11	86142587	1650
chr19:22373841:SNP	ZNF676	cg24983120	chr18	74499380	2308
chr19:22379587:SNP	ZNF676	cg12302982	chr2	39471028	339
chr19:22380261:SNP	ZNF676	cg18640030	chr8	67090752	1283
chr19:23158720:SNP	ZNF728	cg18625538	chr6	87832609	834
chr19:23158720:SNP	ZNF728	cg08355456	chr11	67383691	1632
chr19:23182380:SNP	ZNF728	cg13164087	chr6	87832666	834
chr19:23182380:SNP	ZNF728	cg24690094	chr11	67383802	1632
chr19:23928881:SNP	ZNF681	cg09744766	chr16	53407421	2171
chr19:36692190:SNP	ZNF565	cg08679971	chr19	38281047	2511
chr19:36947180:SNP	ZNF566	cg08679971	chr19	38281047	2511
chr19:36979894:SNP	ZNF566	cg05740763	chr12	13080422	1711
chr19:37251893:SNP	ZNF850	cg14218481	chr19	38281219	2511
chr19:37389471:SNP	ZNF829	cg10208301	chr11	6592745	1564
chr19:37389471:SNP	ZNF829	cg24637308	chr11	6592297	1564
chr19:37415240:SNP	ZNF568	cg01833234	chr11	6592585	1564
chr19:37644051:SNP	ZNF585A	cg24335895	chr19	36643771	2506

chr19:37863059:SNP	ZNF527	cg15215348	chr19	18344520	2481
chr19:52779505:SNP	ZNF766	cg11057205	chr8	134588243	1308
chr19:52946971:SNP	ZNF534	cg09169779	chr13	20751710	1865
chr19:52949319:SNP	ZNF534	cg01932197	chr7	83087065	1001
chr19:52952016:SNP	ZNF534	cg03954144	chr13	20751679	1865
chr19:53041107:SNP	ZNF808	cg24148447	chr8	34764927	1249
chr19:53050134:SNP	ZNF808	cg07770222	chr8	144120106	1192
chr19:53057479:SNP	ZNF808	cg22829325	chr8	144120484	1192
chr19:53059545:SNP	ZNF808	cg04612566	chr8	144120706	1192
chr19:53065661:SNP	ZNF808	cg03058664	chr8	144120635	1192
chr19:53287295:SNP	ZNF600	cg27423959	chr3	126945870	481
chr19:53350823:SNP	ZNF468	cg00060374	chr1	1355235	226
chr19:53367392:SNP	ZNF320	cg27423959	chr3	126945870	481
chr19:53667940:SNP	ZNF665	cg03868770	chr22	30783737	2598
chr19:53945215:SNP	ZNF761	cg06385380	chr7	48887585	1130
chr19:57804524:SNP	ZNF460	cg15321211	chr10	43905515	1362
chr19:57804524:SNP	ZNF460	cg15612221	chr18	37379468	2293
chr19:57901422:SNP	ZNF17	cg01016191	chr12	130707332	1672
chr19:57901422:SNP	ZNF17	cg20736492	chr12	130707407	1672
chr19:58038624:SNP	ZNF549	cg04391048	chr2	18766022	325
chr19:58207899:SNP	ZNF776	cg17481235	chr6	160241360	875
chr19:58223828:SNP	ZNF776	cg19579217	chr6	10720630	857
chr19:58315273:SNP	ZNF587	cg11885357	chr3	138152902	492
chr19:58348570:SNP	ZNF587	cg11586570	chr7	63498133	1158
chr19:58348965:SNP	ZNF587	cg21136104	chr3	142666320	498
chr19:58419841:SNP	ZNF417	cg19137348	chr6	1555311	921
chr19:58706186:SNP	ZNF274	cg13947469	chr7	63505871	1159
chr19:58715538:SNP	ZNF274	cg08747963	chr5	150326862	673
chr19:58715538:SNP	ZNF274	cg18313661	chr7	63560590	1160
chr19:58715682:SNP	ZNF274	cg13763098	chr7	56516117	1146
chr19:58741170:SNP	ZNF544	cg01285926	chr7	56242554	1143
chr19:58741170:SNP	ZNF544	cg11460282	chr7	56297661	1144
chr19:58741170:SNP	ZNF544	cg05655885	chr7	56515601	1146
chr19:58741170:SNP	ZNF544	cg23316599	chr7	57271379	1150
chr19:58741170:SNP	ZNF544	cg20164601	chr7	63642083	1161
chr19:58774762:SNP	ZNF544	cg06703343	chr22	36563048	2602
chr19:58946203:SNP	ZNF132	cg04143909	chr1	234367322	162
chr19:58946203:SNP	ZNF132	cg23445321	chr1	234367493	162
chr19:58946203:SNP	ZNF132	cg04044120	chr11	73668757	1641

Supplementary Table B.3. KZFP mQTLs that are associated with more than one cluster of hvCpGs.

mQTL	hvCpG	CpG chr	CpG pos.	KZFP	cluster.id
chr11:123598620:SNP	cg03689146	chr7	52341810	ZNF202	1136
chr11:123598620:SNP	cg26617364	chr7	52342124	ZNF202	1136
chr11:123598620:SNP	cg20581874	chr14	21191754	ZNF202	1924
chr11:123598620:SNP	cg26393629	chr14	21191859	ZNF202	1924
chr16:3362017:SNP	cg14981253	chr7	5112642	ZNF75A	1094
chr16:3362017:SNP	cg08684893	chr19	44616615	ZNF75A	2341
chr16:3366513:SNP	cg25786696	chr19	44406241	ZNF75A	2340
chr16:3366513:SNP	cg07504200	chr19	2889208	ZNF75A	2437
chr16:31923839:SNP	cg00876757	chr16	33483591	ZNF267	2153
chr16:31923839:SNP	cg08704196	chr16	34787392	ZNF267	2162
chr19:12588119:SNP	cg17125837	chr11	358595	ZNF709	1567
chr19:12588119:SNP	cg18815398	chr20	61506981	ZNF709	2559
chr19:22373841:SNP	cg15146462	chr11	86142587	ZNF676	1650
chr19:22373841:SNP	cg24983120	chr18	74499380	ZNF676	2308
chr19:23158720:SNP	cg18625538	chr6	87832609	ZNF728	834
chr19:23158720:SNP	cg08355456	chr11	67383691	ZNF728	1632
chr19:23182380:SNP	cg13164087	chr6	87832666	ZNF728	834
chr19:23182380:SNP	cg24690094	chr11	67383802	ZNF728	1632
chr19:57804524:SNP	cg15321211	chr10	43905515	ZNF460	1362
chr19:57804524:SNP	cg15612221	chr18	37379468	ZNF460	2293
chr19:58715538:SNP	cg08747963	chr5	150326862	ZNF274	673
chr19:58715538:SNP	cg18313661	chr7	63560590	ZNF274	1160
chr19:58741170:SNP	cg01285926	chr7	56242554	ZNF544	1143
chr19:58741170:SNP	cg11460282	chr7	56297661	ZNF544	1144
chr19:58741170:SNP	cg05655885	chr7	56515601	ZNF544	1146
chr19:58741170:SNP	cg23316599	chr7	57271379	ZNF544	1150
chr19:58741170:SNP	cg20164601	chr7	63642083	ZNF544	1161
chr19:58946203:SNP	cg04143909	chr1	234367322	ZNF132	162
chr19:58946203:SNP	cg23445321	chr1	234367493	ZNF132	162
chr19:58946203:SNP	cg04044120	chr11	73668757	ZNF132	1641

Supplementary Table B.4. Enrichment and depletion for proximal KZFP binding sites relative to the Illumina450K array background. Only the KZFPs with at least 1% of hvCpGs proximal to a binding site and with FDR-adjusted p-value < 0.05 are shown. We considered several proximity thresholds (0kb, 1kb, 2kb and 3kb). Results are ordered by Odds Ratio and those highlighted in green have OR > 1.

KZFP ChIP-exo experiment	prop.hvCpGs	prop. array_back	Odds Ratio	FET p-value	FDR-adjusted p-value	proximity threshold
ZNF445	0.016	0.008	2.011	4.783E-05	5.915E-04	0
ZFP57-rep2	0.025	0.016	1.625	2.682E-04	2.741E-03	0
ZFP57	0.045	0.031	1.499	4.509E-05	4.159E-04	1000
ZNF445	0.041	0.029	1.443	3.756E-04	2.847E-03	1000
ZNF2	0.044	0.031	1.407	6.238E-04	4.311E-03	1000
ZNF445	0.056	0.043	1.331	1.066E-03	7.160E-03	2000
ZNF2	0.056	0.043	1.329	1.258E-03	7.993E-03	2000
ZFP57	0.059	0.045	1.319	1.373E-03	8.109E-03	2000
ZFP57-rep2	0.095	0.077	1.265	5.732E-04	4.082E-03	1000
ZFP57-rep2	0.122	0.105	1.178	6.833E-03	3.149E-02	2000
ZNF460	0.054	0.067	0.787	4.970E-03	2.433E-02	2000
ZNF695	0.056	0.072	0.775	2.133E-03	1.319E-02	1000
ZNF460	0.063	0.080	0.773	1.031E-03	6.374E-03	3000
ZNF479	0.131	0.165	0.761	1.278E-06	1.365E-05	2000
ZNF479	0.155	0.195	0.760	1.882E-07	2.032E-06	3000
ZNF695	0.069	0.089	0.751	1.400E-04	1.135E-03	2000
ZNF786	0.047	0.062	0.750	1.320E-03	8.109E-03	2000
ZNF680	0.087	0.115	0.733	3.218E-06	3.151E-05	3000
ZNF695	0.076	0.101	0.731	1.038E-05	9.378E-05	3000
ZNF786	0.052	0.069	0.726	1.908E-04	1.546E-03	3000
ZNF479	0.095	0.128	0.719	2.536E-07	3.137E-06	1000
ZNF680	0.070	0.095	0.714	5.026E-06	4.921E-05	2000
ZNF671	0.027	0.038	0.700	1.933E-03	1.033E-02	3000
ZNF611	0.021	0.030	0.697	6.008E-03	2.824E-02	3000
ZNF441	0.110	0.151	0.696	1.524E-09	2.238E-08	3000
ZNF441	0.067	0.094	0.693	8.244E-07	9.686E-06	1000
ZNF671	0.023	0.033	0.689	3.488E-03	1.821E-02	2000
ZNF84	0.017	0.025	0.685	8.778E-03	4.045E-02	3000
ZNF671	0.017	0.025	0.681	8.275E-03	4.137E-02	1000
ZNF425	0.034	0.050	0.675	1.284E-04	1.078E-03	3000
ZNF441	0.089	0.127	0.672	1.144E-09	1.793E-08	2000
ZNF786	0.034	0.050	0.667	7.976E-05	6.942E-04	1000
ZNF695	0.016	0.024	0.666	6.213E-03	4.867E-02	0
ZNF680	0.047	0.069	0.660	2.383E-06	2.434E-05	1000

ZNF44	0.026	0.039	0.658	3.377E-04	2.491E-03	3000
ZNF425	0.027	0.041	0.652	1.807E-04	1.370E-03	2000
ZNF682-rep2	0.022	0.035	0.634	2.765E-04	2.166E-03	1000
ZNF534	0.164	0.238	0.630	3.375E-20	8.812E-19	2000
ZNF534	0.190	0.271	0.630	2.521E-22	6.187E-21	3000
ZKSCAN5	0.046	0.071	0.624	8.079E-08	9.493E-07	3000
ZNF44	0.022	0.034	0.622	1.650E-04	1.292E-03	2000
ZNF682-rep2	0.027	0.043	0.620	2.245E-05	2.029E-04	2000
ZNF682-rep2	0.031	0.049	0.619	5.893E-06	5.539E-05	3000
ZFP69B	0.018	0.029	0.611	3.981E-04	2.835E-03	3000
ZNF84	0.012	0.020	0.611	3.808E-03	1.946E-02	2000
ZKSCAN5	0.039	0.062	0.609	1.553E-07	1.921E-06	2000
ZNF674	0.050	0.079	0.605	2.193E-09	3.031E-08	3000
ZNF257	0.148	0.224	0.604	2.633E-22	6.187E-21	3000
ZNF611	0.012	0.019	0.603	3.484E-03	1.997E-02	1000
ZNF674	0.043	0.070	0.602	1.523E-08	2.237E-07	2000
ZNF257	0.126	0.193	0.601	3.882E-20	9.124E-19	2000
ZNF468	0.080	0.127	0.594	9.589E-15	1.878E-13	3000
ZKSCAN5	0.028	0.047	0.589	1.236E-06	1.383E-05	1000
ZNF519	0.147	0.226	0.589	2.523E-24	9.884E-23	3000
ZNF561	0.011	0.018	0.589	2.854E-03	1.369E-02	3000
ZNF534	0.122	0.190	0.587	1.884E-21	6.325E-20	1000
ZNF44	0.016	0.027	0.581	2.245E-04	1.819E-03	1000
ZNF425	0.017	0.030	0.580	8.557E-05	7.182E-04	1000
ZNF468	0.067	0.111	0.572	1.097E-14	2.344E-13	2000
ZNF519	0.127	0.202	0.570	1.781E-24	8.369E-23	2000
ZNF519	0.102	0.166	0.565	2.439E-21	6.369E-20	1000
ZNF257	0.089	0.149	0.558	3.036E-20	7.134E-19	1000
ZNF189	0.043	0.075	0.555	2.693E-11	4.520E-10	3000
ZNF468	0.050	0.088	0.543	1.019E-13	1.996E-12	1000
ZNF846	0.028	0.050	0.542	2.202E-08	2.874E-07	2000
ZNF100-rep2	0.020	0.037	0.539	1.100E-06	1.231E-05	2000
ZNF189	0.037	0.066	0.537	5.727E-11	9.613E-10	2000
ZNF100-rep2	0.023	0.042	0.535	1.903E-07	2.032E-06	3000
ZNF273	0.081	0.141	0.533	1.207E-21	3.545E-20	2000
ZNF273	0.091	0.159	0.531	4.079E-24	1.369E-22	3000
ZNF846	0.023	0.043	0.530	6.805E-08	9.407E-07	1000
ZNF674	0.030	0.056	0.528	7.394E-10	1.158E-08	1000
ZNF263	0.125	0.214	0.524	2.851E-32	1.675E-30	3000

ZNF846	0.029	0.055	0.518	3.950E-10	6.189E-09	3000
ZNF189	0.028	0.054	0.511	3.003E-10	5.041E-09	1000
ZFP69B	0.013	0.024	0.503	1.601E-05	1.505E-04	2000
ZNF610	0.039	0.075	0.501	3.449E-14	6.754E-13	2000
ZNF610	0.044	0.083	0.501	1.231E-15	2.629E-14	3000
ZNF273	0.061	0.115	0.496	2.373E-21	6.369E-20	1000
ZNF675	0.014	0.028	0.495	2.516E-06	2.570E-05	3000
ZNF263	0.105	0.191	0.494	8.417E-34	9.890E-32	2000
ZNF547	0.019	0.039	0.481	8.210E-09	1.072E-07	3000
ZNF343	0.055	0.110	0.468	4.506E-23	1.324E-21	3000
ZNF547	0.017	0.035	0.466	1.982E-08	2.740E-07	2000
ZNF479	0.013	0.027	0.455	4.825E-07	7.558E-06	0
ZNF263	0.080	0.159	0.455	5.165E-34	6.069E-32	1000
ZNF547	0.013	0.029	0.455	2.141E-07	2.796E-06	1000
ZNF675	0.011	0.024	0.454	2.504E-06	2.559E-05	2000
ZNF282-rep2	0.056	0.116	0.454	9.022E-26	4.240E-24	3000
ZNF610	0.029	0.062	0.448	6.629E-15	1.416E-13	1000
ZNF100-rep2	0.014	0.030	0.445	6.172E-08	9.064E-07	1000
ZNF282-rep2	0.048	0.102	0.444	9.319E-24	3.650E-22	2000
ZNF343	0.047	0.099	0.441	1.254E-23	4.208E-22	2000
ZNF282	0.064	0.137	0.429	3.882E-33	3.041E-31	3000
ZNF534	0.016	0.039	0.411	3.213E-11	8.389E-10	0
ZNF519	0.012	0.028	0.410	1.563E-08	2.825E-07	0
ZNF257	0.012	0.029	0.407	7.100E-09	1.390E-07	0
ZNF282	0.053	0.121	0.406	2.209E-32	1.730E-30	2000
ZNF343	0.034	0.083	0.391	2.276E-24	8.914E-23	1000
ZNF202	0.128	0.274	0.389	3.928E-73	9.231E-71	2000
ZNF283	0.016	0.040	0.388	1.829E-12	3.307E-11	3000
ZNF202	0.141	0.298	0.386	1.563E-79	3.672E-77	3000
ZNF783	0.033	0.083	0.374	6.733E-26	3.165E-24	1000
ZNF783	0.039	0.098	0.374	4.617E-30	2.712E-28	2000
ZNF783	0.042	0.107	0.367	1.316E-33	1.546E-31	3000
ZNF283	0.013	0.035	0.367	5.291E-12	9.565E-11	2000
ZNF282-rep2	0.031	0.081	0.355	3.072E-27	1.805E-25	1000
ZNF283	0.010	0.028	0.355	2.798E-10	5.041E-09	1000
ZNF282	0.036	0.097	0.348	6.379E-33	4.997E-31	1000
ZNF202	0.102	0.244	0.348	1.793E-78	4.213E-76	1000

Appendix C

Supplementary Figures for Chapter 5

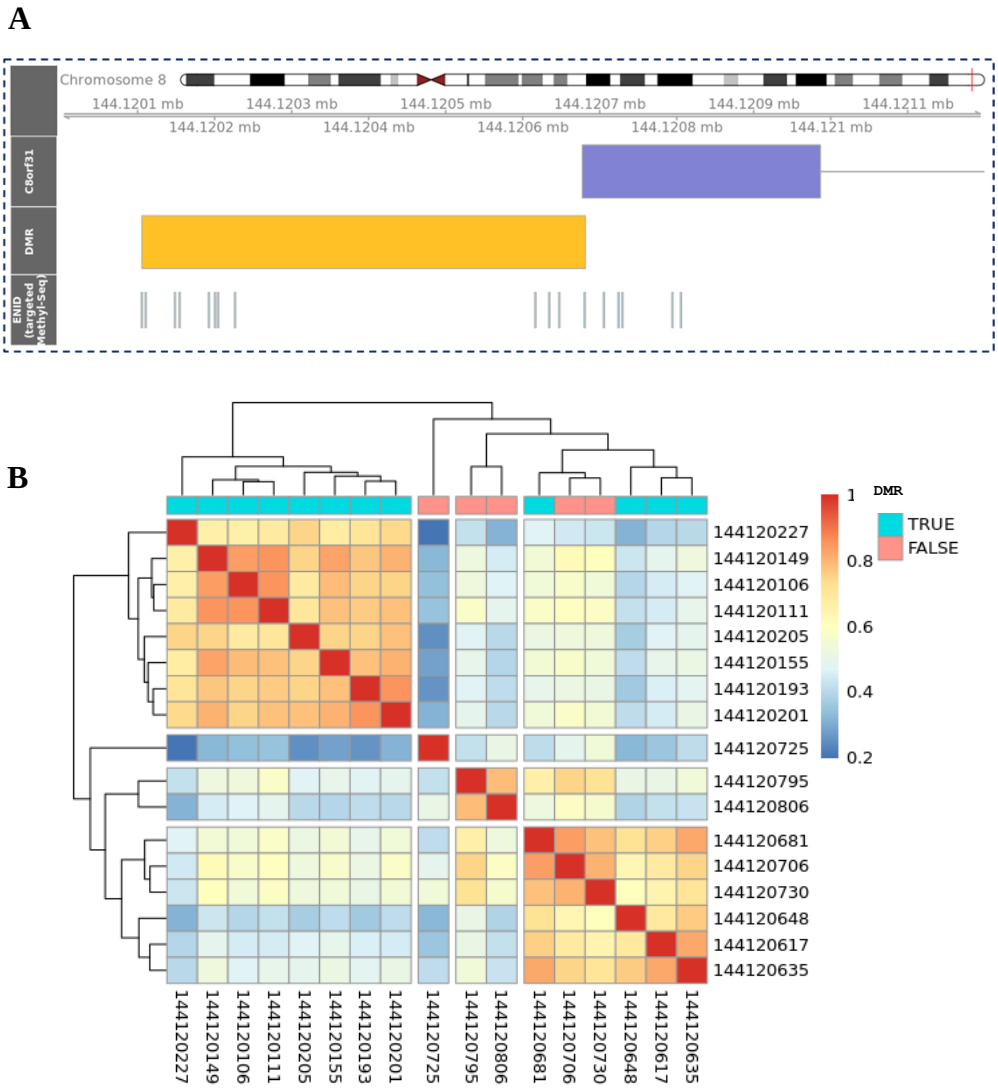


Figure C.1. CpGs covered in the targeted Methyl-Seq dataset (ENID cohort) fall into 4 correlated CpG blocks. A) Positions of the 17 CpGs covered in targeted Methyl-Seq data (bottom track) relative to the fasting insulin DMR (orange) and the first exon of *C8orf31* (purple). These CpGs had > 30x coverage in at least 300 ENID individuals (median = 479, IQR = [452, 501)]. **B)** Pairwise methylation correlation (Pearson r) at C8orf31 CpGs calculated across 250 individuals that had > 30x coverage at all 17 CpGs. CpGs fall into 4 correlated block each with a pairwise r > 0.7. Tile colour gives Pearson r correlation. CpGs located inside the DMR identified by Antoun et al. are highlighted in turquoise and those outside the DMR are highlighted in pink.

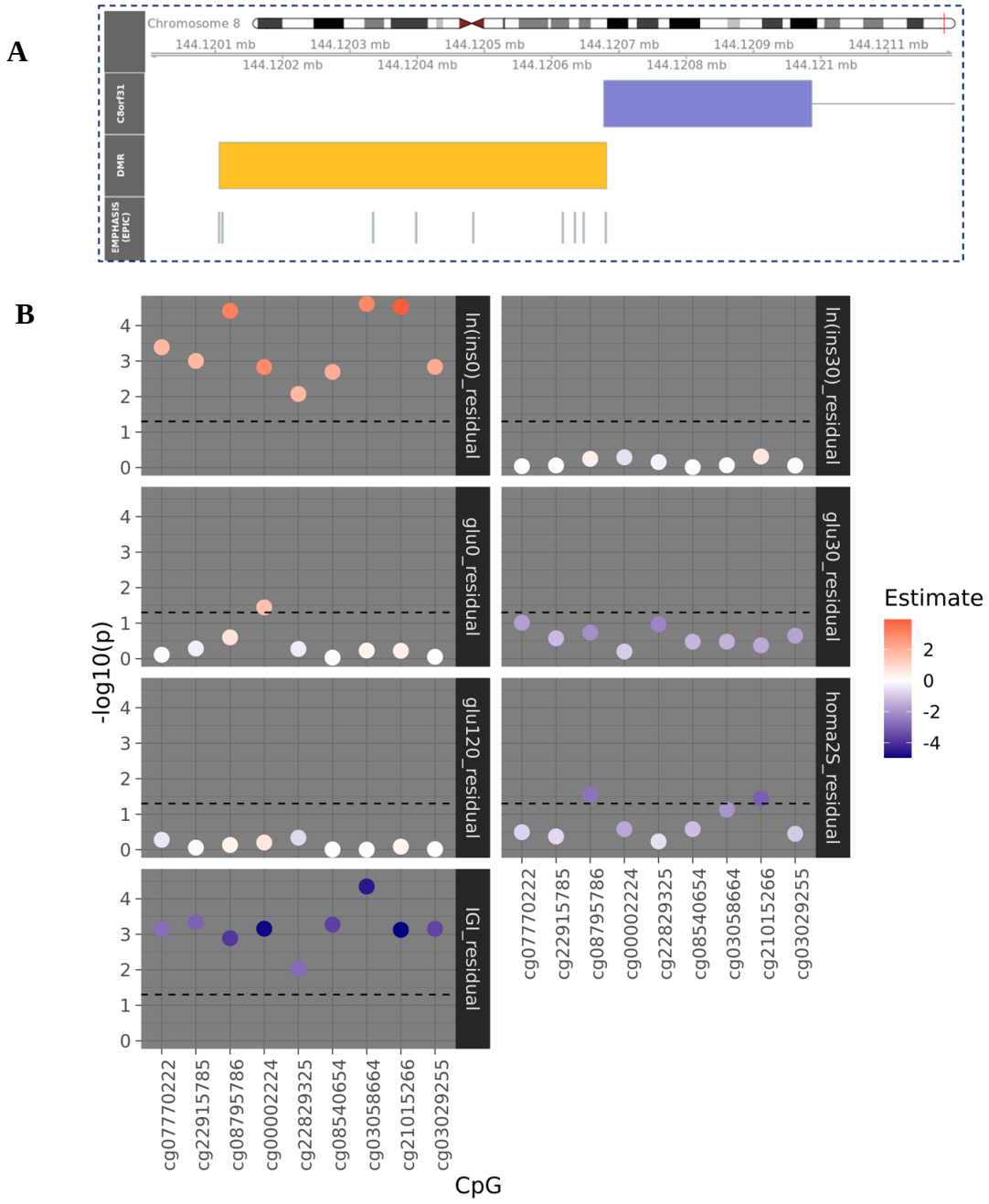


Figure C.2. Association between *C8orf31* DMR methylation and phenotypes linked to glucose metabolism in the EMPHASIS cohort of Gambian children. **A)** Positions of the 9 IlluminaEPIC CpGs relative to the fasting insulin DMR (orange) and the first exon of *C8orf31* (purple). **B)** Y-axis gives p-values from a linear model in which phenotypic measurements (facet labels) were used as outcome variables and methylation Beta values at CpGs located in the *C8orf31* DMR in the IlluminaEPIC manifest (x-axis) were used as predictor variables. Note, outcome variables are the residuals obtained after adjusting linear phenotypic measurements for BMI, sex and age (see Methods). Point colour gives the effect estimate of the association and the dashed horizontal line indicates a nominal p-value threshold < 0.05. Ins0 = fasting insulin, ins30 = 30-min insulin, glu0 = fasting glucose, glu30 = 30-min glucose, glu120 = 2hr-glucose IGI = insulinogenic index ((ins30-ins0)/(glu30-glu0)), HOMA2-S = homeostasis model assessment of insulin sensitivity.

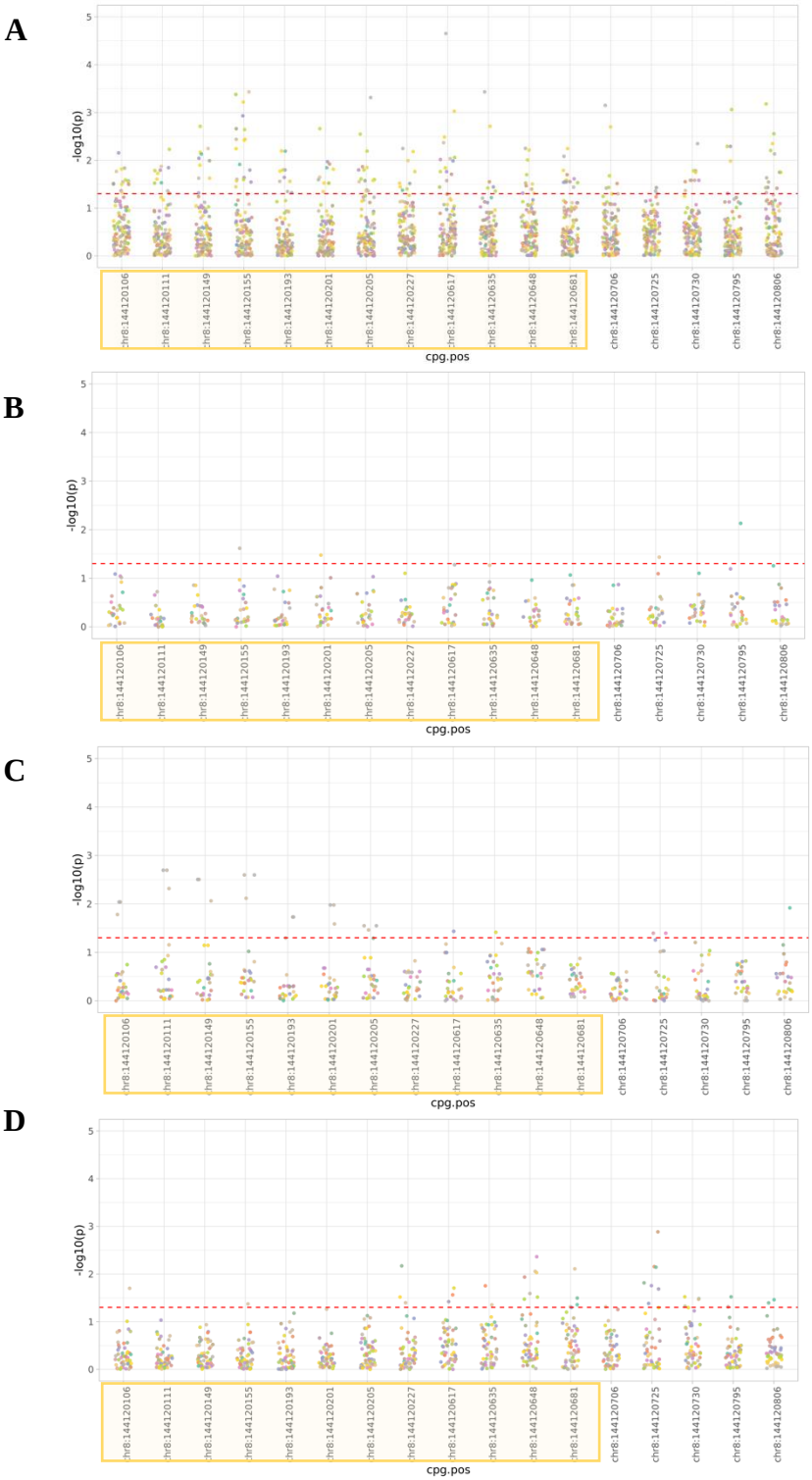


Figure C.3. Manhattan plots showing p-values from linear models testing association between methylation at *C8orf31* DMR CpGs and genotype at KZFP genes known to target MER11 retrotransposons. A) *ZNF578* (nSNPs = 115). **B)** *ZNF525* (nSNPs = 24). **C)** *ZNF440* (nSNPs = 27). **D)** *ZNF468* (nSNPs = 52). Point colour indicates the tested SNP. Red dashed line indicates nominal p-value threshold < 0.05. CpGs located in the DMR are highlighted in the yellow box.

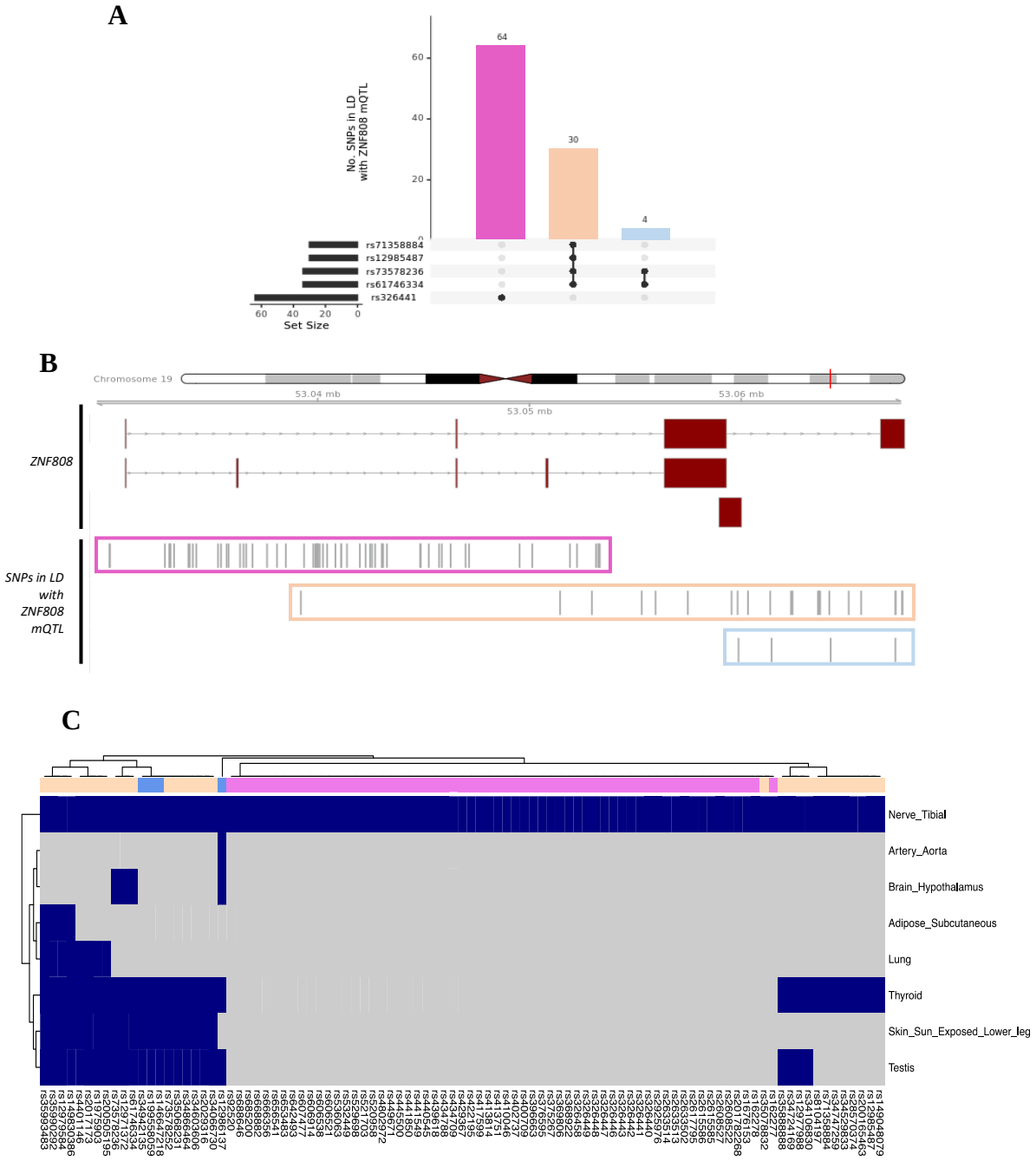


Figure C.4. Overlap between ZNF808 mQTLs and eQTLs. A) UpSet graph showing the number of SNPs that are in LD with each GoDMC ZNF808 mQTL associated with a CpG in the *C8orf31* DMR. In total, 98 SNPs were LD-linked to at least one mQTL. LD-linked SNPs were identified in HaploReg which uses the EUR phase1 1000 Genomes reference panel (r2 > 0.8). **B)** The genomic positions of the 98 SNPs in relation to *ZNF808* (hg19 coordinates). Exons are in dark red. **C)** The overlap between LD-linked SNPs (x-axis) and GTEx eQTLs across tissue types (y-axis). Blue squares indicate overlaps. SNPs are coloured according to the set of *ZNF808* mQTLs they are in LD with (panel A).

Figure C.5. Overlap between ZNF808 mQTLs and cell-type specific promoters. The overlap between LD-linked SNPs (x-axis) and ChromHMM-predicted promoter regions across cell types from Roadmap Epigenome (y-axis). SNPs are coloured according to the set of *ZNF808* mQTLs they are in LD with (see Figure C.4A). Blue squares indicate overlaps.

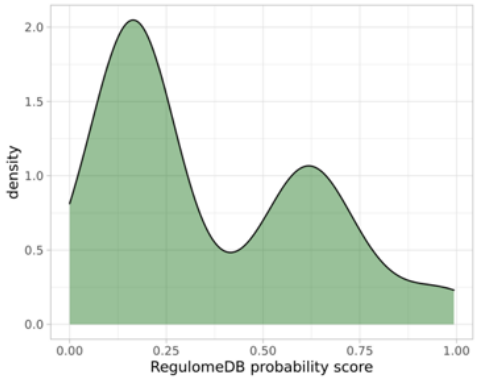


Figure C.6. The distribution of RegulomeDB scores for ZNF808 mQTLs and LD-linked SNPs.

A

Regulatory motifs altered

Position Weight Matrix ID (Library from Kheradpour and Kellis, 2013)	Strand	Ref	Alt	Match on: chr8:144,122,698 –144,122,709
				Ref: CGTCCCCAAAGCAGACATTTCAGAGGCCTCCCCTTAGGGATACATTCTCTTTCTCAGGG Alt: CGTCCCCAAAGCAGACATTTCAGAGGCCTTCCCTTAGGGATACATTCTCTTTCTCAGGG
AP-2_known4	-	11.7	11.3	NTYSCCYSRGGSDNW
EBF_known1	+	10	12.4	TCCCYDGRGK
EBF_known2	+	2	13.8	BTCCCYWRRGD

B

Figure C.7. The mQTL (rs6558371) located in the ZNF808 ChIP-exo peak located in MER11C is predicted to alter binding affinity for Early B Cell factor (EBF). A) Screen shot from HaploReg v 4.1 showing the regulatory motifs predicted to be altered by this mQTL, and their effect on binding affinity. Green square shows that **r**s6558371 has a particular effect on the binding affinity of EBF. **B)** Sequence logo for EBF_known2 motif. Note, this logo is the reverse complement of the motif in the purple box in the table. The C → T variant occurs at the conserved T nucleotide ('A' in the sequence logo), explaining why it increases EBF binding affinity from 2 to 13.8 (green box in panel A).

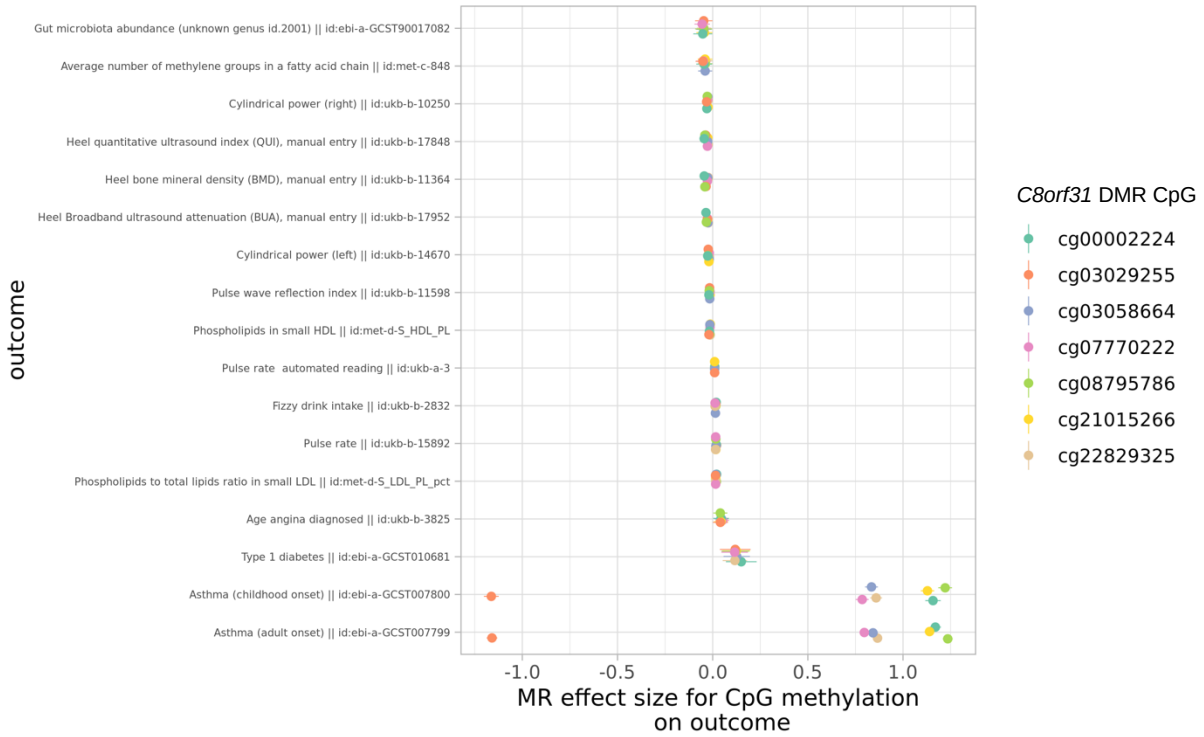


Figure C.8. Phenome-wide Two-sample Mendelian Randomisation. Forest plot showing the mean casual effect of each CpG on a given outcome (y-axis). Horizontal lines are 95% confidence interval and point colours indicate the DMR CpG. Only outcomes that are nominally associated ($p < 0.05$) with all 7 Illumina450K DMR CpGs and which have a minimum absolute MR effect size of 0.05 are shown for clarity. See 'Methods' for further details.

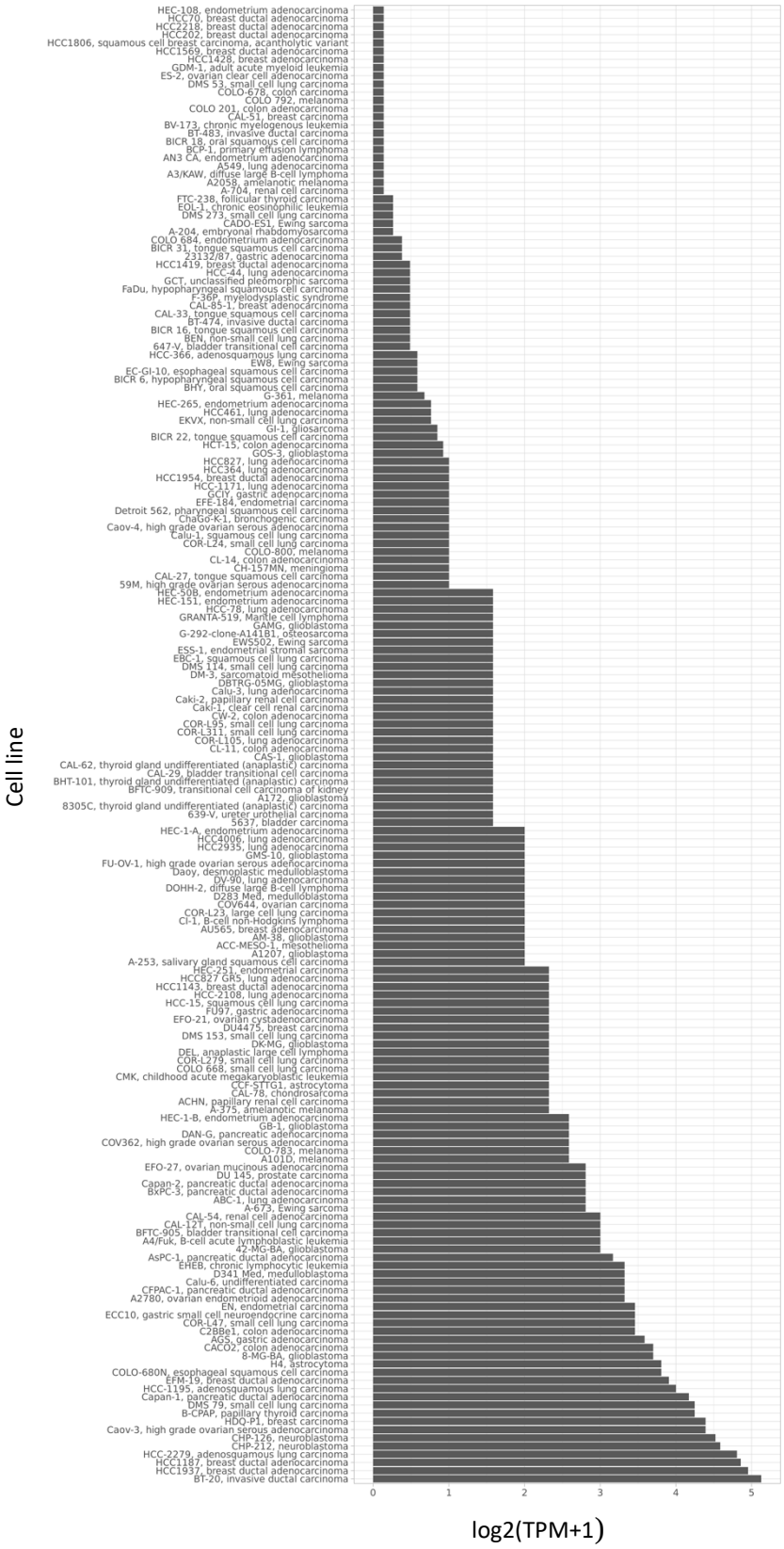


Figure C.9. Gene expression of *C8orf31* in cancer cell lines part of the Cancer Cell Line Encyclopedia (CCLE). TPM = transcripts per million

Figure C.10. BLAST analysis of nucleotide sequence of *C8orf31* transcript 1. Screenshot of BLAST's 'Multiple Sequence Alignment Viewer' showing hits with > 80% sequence identity and with > 20% query sequence coverage. Red lines are mismatches, blue brackets are insertions relative to the query sequence, and green brackets are unaligned regions. Green box highlights the alignments of *LY6S* transcripts.

Supplementary Tables for Chapter 5

Supplementary Table C.1. Descriptions of GWAS traits used for Two-sample Mendelian Randomisation analysis. All GWAS were carried out in European cohorts and were accessed using MR-Base.

Trait	Trait description	MR-Base id	PMID	Sample Size	N. SNPs
Type 1 diabetes	Pancreas makes little or no insulin.	ebi-a-GCST010681	32005708	24840	12783129
Type 2 diabetes	Caused by insulin resistance (ability of cells in peripheral tissues to uptake glucose in response to insulin) and/or reduced insulin production/secretion by pancreatic beta cells in response to glucose.	ebi-a-GCST006867	30054458	655666	5030727
Fasting insulin (blood)	High fasting insulin levels indicate a degree of insulin resistance.	ebi-a-GCST90002238	34059833	151013	29664438
Fasting glucose (serum)	High fasting glucose levels are associated with diabetes.	ieu-b-113	22581228	58074	2625495
HbA1C	Glycated haemoglobin - commonly used as a biomarker of glucose control (glycation rate is associated with blood glucose levels).	ieu-b-103	20858683	46368	2529804
HOMA-IR	Homeostasis model assessment of insulin resistance - derived from fasting glucose and fasting insulin measurements.	ieu-b-118	20081858	37037	2455342
HOMA-B	Homeostasis model assessment of beta-cell function - derived from fasting glucose and fasting insulin.	ieu-b-117	20081858	36466	2454220
Proinsulin levels	Proinsulin is biologically inactive and is cleaved into active insulin and C peptide. Higher C peptide levels are associated with increased beta cell function.	ebi-a-GCST001212	21873549	10701	2479861
Acute Insulin Response (AIR)	Incremental area under the insulin curve during the first 10 min after intravenous glucose stimulation. This measure was adjusted for insulin sensitivity (calculated using the 'MinMod' measure) and BMI.	ebi-a-GCST006677	28490609	2087	9629468
		ebi-a-GCST004575 (insulin sensitivity adjusted)	28490609	2087	9663724
		ebi-a-GCST006678 (insulin sensitivity and BMI adjusted)	28490609	2087	9629245
Insulin Secretion Rate (ISR)	Uses measured serum C-peptide concentrations to provide an estimate of the rate of insulin secretion before insulin clearance. This measure was also adjusted for BMI.	ebi-a-GCST004488	28490609	527	6906448
		ebi-a-GCST006679 (BMI adjusted)	28490609	527	6906448

Supplementary Table C.2. The association between insulinogenic index (IGI)/ fasting insulin levels (ins0) with methylation at CpGs in the *C8orf31* DMR using data from the EMPHASIS cohort.

model	CpG	Hg19 pos (chr 8)	Pr(>|t|)	Estimate
IGI_residual ~ cpg + sex + age + plate + row	cg03058664	144120635	4.519E-05	-0.046
	cg22915785	144120111	4.551E-04	-0.030
	cg08540654	144120617	5.345E-04	-0.037
	cg00002224	144120399	6.965E-04	-0.049
	cg03029255	144120681	7.025E-04	-0.036
	cg07770222	144120106	7.236E-04	-0.027
	cg21015266	144120648	7.506E-04	-0.050
	cg08795786	144120335	1.289E-03	-0.039
	cg22829325	144120484	9.054E-03	-0.028
ln(ins0)_residual ~ cpg + sex + age + plate + row	cg03058664	144120635	2.566E-04	0.189
	cg21015266	144120648	2.566E-04	0.500
	cg08795786	144120335	2.566E-04	0.229
	cg07770222	144120106	2.033E-03	0.051
	cg22915785	144120111	3.961E-03	0.051
	cg03029255	144120681	4.211E-03	0.078
	cg00002224	144120399	4.211E-03	0.183
	cg08540654	144120617	5.051E-03	0.071
	cg22829325	144120484	1.873E-02	0.050

Supplementary Table C.3. Notable characteristics of Illumina450K CpGs located in the *C8orf31* DMR, including mQTL associations and hypervariability in datasets used in Chapter 3.

CpG)	CpG position (hg19)	Gene position (UCSC)	hvCpG	prop datasets	n. tissues	n. ethnicities	GoDMC mQTL (blue = *cis* mQTL, pink = *ZNF808 trans* mQTL, orange = *ZNF578 trans* mQTL)
cg07770222	chr8:144120106	TSS1500	TRUE	0.72	12	7	chr19:53050134:SNP, chr8:144116165:SNP, chr8:144309595:SNP
cg08795786	chr8:144120335	TSS1500	FALSE	0.59	8	7	chr19:53021563:SNP, chr8:144118449:SNP
cg00002224	chr8:144120399	TSS1500	FALSE	0.5	7	6	chr19:53065075:SNP, chr8:144118449:SNP
cg22829325	chr8:144120484	TSS200	TRUE	0.7	4	7	chr19:53057479:SNP, chr8:144309595:SNP, chr8:144116165:SNP
cg03058664	chr8:144120635	TSS200	TRUE	0.65	4	7	chr19:53065661:SNP, chr8:144116165:SNP, chr8:144309595:SNP
cg21015266	chr8:144120648	TSS200	FALSE	0.45	4	4	chr19:53065075:SNP, chr8:144116165:SNP
cg03029255	chr8:144120681	1stExon 5'UTR	FALSE	0.55	4	5	chr19:52997170:SNP, chr8:143735190:SNP, chr8:144193737:SNP

Supplementary Table C.4. Summary statistics for *ZNF808 trans* mQTLs identified in the GoDMC meta-analysis associated with 5 Illumina450K CpGs located in the *C8orf31* DMR.

rsID	SNP position (hg19)	CpG	Allele1	Allele2	Freq1 (MAF)	Effect
rs71358884	chr19:53065075:SNP	cg00002224	t	c	0.0665	0.50
rs12985487	chr19:53065661:SNP	cg03058664	g	a	0.0663	0.48
rs73578236	chr19:53059545:SNP	cg04612566	t	c	0.0674	0.454
rs326441	chr19:53050134:SNP	cg07770222	a	g	0.2373	0.21
rs71358884	chr19:53065075:SNP	cg21015266	t	c	0.0663	0.45
rs61746334	chr19:53057479:SNP	cg22829325	a	g	0.0678	0.40

Supplementary Table C5. *C8orf31* GWAS associations obtained from https://www.genecards.org/cgi-bin/carddisp.pl?gene=LINC02904

The term 'GeneHancer' refers to a SNP ocated in an enhancer of *C8orf31*.

Phenotype	Gene Relation	log10(p-value)	SNP IDs	Mapped gene	PMID
low density lipoprotein cholesterol measurement	GeneHancer	17.4	rs11781667	PLEC	26690388
body weight	GeneHancer	11.5	rs2976384	PSCA, JRK	34594039
HMG CoA reductase inhibitor use measurement	GeneHancer	10	rs11787365	PLEC	31015401
body mass index	GeneHancer	9.3	rs2976384	PSCA, JRK	34594039
total cholesterol measurement	GeneHancer	9	rs11781667	PLEC	34594039
optic disc size measurement	GeneHancer	7.3	rs13264644	PSCA, JRK	31809533
3-hydroxy-1-methylpropylmercapturic acid measurement	GeneHancer	7	rs72690918	PSCA, JRK	26053186
cranial vault morphology measurement	GWAS	6	rs187705391	C8orf31, LY6E	29698431
eye colour measurement	GWAS	5.7	rs3758430	DMBT1,HTRA1	30895295
acute myeloid leukaemia	GWAS	5.5	rs12547949	C8orf31	27903959
fasting blood glucose change measurement	GWAS	5.3	rs4977129	LY6H, LYGL	31263163

Supplementary Table C6. Summary statistics for GoDMC 2SMR genetic instruments.

SNP	beta.exposure	se.exposure	effect_allele.exposure	eaf.exposure	pval.exposure	exposure
rs71358884	0.5032667	0.016039	T	0.0665	1E-200	cg00002224
rs4077724	0.5317837	0.0091747	T	0.3472	1E-200	cg00002224
rs382838	0.4355818	0.0297372	A	0.0517	1E-200	cg03029255
rs4736342	0.1087127	0.0089503	A	0.6378	1E-200	cg03029255
rs4076116	0.706516	0.0131192	G	0.1947	1E-200	cg03029255
rs12985487	0.47772	0.014222	G	0.0663	1E-200	cg03058664
rs11782280	0.6577705	0.0153903	G	0.4064	1E-200	cg03058664
rs6990605	0.1156561	0.0095946	G	0.3262	1.842E-33	cg03058664
rs326441	0.2057797	0.00803	A	0.2373	1E-200	cg07770222
rs11782280	0.6679562	0.0153363	G	0.4062	1E-200	cg07770222
rs6990605	0.1189108	0.0095916	G	0.3262	2.702E-35	cg07770222
rs71358878	0.433237	0.0183249	A	0.0558	1E-200	cg08795786
rs4077724	0.6250454	0.0089367	T	0.3475	1E-200	cg08795786
rs71358884	0.4539442	0.0144694	T	0.0663	1E-200	cg21015266
rs11782280	0.6375974	0.0155255	G	0.4065	1E-200	cg21015266
rs61746334	0.3968692	0.0138283	A	0.0678	3.83E-181	cg22829325
rs6990605	0.1225502	0.0095915	G	0.3262	2.204E-37	cg22829325

rs11782280	0.7248674	0.0149702	G	0.4063	1E-200	cg22829325

Supplementary Table C7. 2SMR results and horizontal pleiotropy tests. Note, only exposure-outcome paris with MR p-val < 0.05 are shown. All tests were performed using the Inverse Variance Weighted (fixed effects method). At least 3 genetic variants need to be associated with a given exposure in order to perform the MR Egger tests, explaining the 'NAs' in the 'egger_intercept' and 'egger_pval' column.

outcome	exposure	nsnp	b	se	MR pval	Q	Q_df	Q pval	Egger intercept	Egger pval
Type 1 diabetes || id:ebi-a-GCST010681	cg22829325	3	0.117	0.033	0.000	9.065	2	0.0108	0.082	0.307
	cg07770222	3	0.116	0.036	0.001	5.799	2	0.0551	0.061	0.257
	cg08795786	2	0.120	0.037	0.001	2.807	1	0.0938	NA	NA
	cg21015266	2	0.124	0.037	0.001	3.761	1	0.0525	NA	NA
	cg00002224	2	0.150	0.041	0.000	2.438	1	0.1185	NA	NA
	cg03029255	3	0.118	0.041	0.004	3.098	2	0.2125	0.029	0.622
	cg03058664	3	0.126	0.035	0.000	7.123	2	0.0284	0.076	0.302
Insulin secretion rate || id:ebi-a-GCST004488	cg22829325	3	0.135	0.058	0.020	3.244	2	0.198	0.058	0.641
	cg07770222	3	0.139	0.063	0.027	2.065	2	0.356	0.060	0.429
	cg08795786	2	0.165	0.082	0.043	1.147	1	0.284	NA	NA
	cg21015266	2	0.148	0.065	0.023	1.839	1	0.175	NA	NA
	cg00002224	2	0.205	0.089	0.021	1.001	1	0.317	NA	NA
	cg03058664	3	0.148	0.062	0.017	1.996	2	0.369	0.042	0.684
Insulin secretion rate (BMI adjusted) || id:ebi-a-GCST006679	cg22829325	3	0.134	0.058	0.021	3.319	2	0.190	0.059	0.636
	cg07770222	3	0.138	0.063	0.028	2.153	2	0.341	0.062	0.422
	cg08795786	2	0.165	0.083	0.048	1.177	1	0.278	NA	NA
	cg21015266	2	0.147	0.065	0.024	1.910	1	0.167	NA	NA
	cg00002224	2	0.205	0.090	0.023	1.046	1	0.306	NA	NA
	cg03058664	3	0.147	0.062	0.018	2.084	2	0.353	0.044	0.680
Acute insulin response || id:ebi-a-GCST004575	cg07770222	3	0.120	0.050	0.016	1.050	2	0.5916	0.015	0.680
	cg08795786	2	0.112	0.041	0.006	0.390	1	0.5322	NA	NA
	cg00002224	2	0.127	0.045	0.005	0.464	1	0.4960	NA	NA
Acute insulin response (insulin sensitivity adjusted) || id:ebi-a-GCST006677	cg07770222	3	0.117	0.050	0.019	1.287	2	0.5255	0.026	0.515
	cg08795786	2	0.113	0.041	0.006	2.125	1	0.1449	NA	NA
	cg00002224	2	0.134	0.045	0.003	1.109	1	0.2923	NA	NA
Acute insulin response (insulin sensitivity and BMI adjusted) || id:ebi-a-GCST006678	cg07770222	3	0.121	0.050	0.016	0.834	2	0.6589	0.020	0.591
	cg08795786	2	0.129	0.041	0.002	2.203	1	0.1378	NA	NA
	cg00002224	2	0.150	0.045	0.001	1.342	1	0.2466	NA	NA

Supplementary Table C.8. 2SMR directionality test outputs from the 'Two-sample MR' package Note, only those traits with significant causal associations with CpGs in the *C8orf31* DMR are shown.

Outcome	exposure	snp_r2.exposure	snp_r2.outcome	Correct causal direction	Steiger_pval
Type 1 diabetes || id:ebi-a-GCST010681	cg22829325	0.1629	0.0009	TRUE	0
	cg07770222	0.1623	0.0007	TRUE	0
	cg08795786	0.0786	0.0005	TRUE	1.2663E-183
	cg21015266	0.1583	0.0006	TRUE	0
	cg00002224	0.0738	0.0006	TRUE	5.9119E-174
	cg03029255	0.1952	0.0005	TRUE	0
	cg03058664	0.1641	0.0008	TRUE	0
Insulin secretion rate || id:ebi-a-GCST004488	cg22829325	0.1629	0.0164	TRUE	1.43565E-11
	cg07770222	0.1623	0.0133	TRUE	1.9869E-12
	cg08795786	0.0786	0.0100	TRUE	2.10688E-05
	cg21015266	0.1583	0.0134	TRUE	6.36488E-12
	cg00002224	0.0738	0.0120	TRUE	0.000132958
	cg03029255	0.1952	0.0099	TRUE	2.88857E-17
	cg03058664	0.1641	0.0145	TRUE	3.18543E-12
Insulin secretion rate (BMI adjusted) || id:ebi-a-GCST006679	cg22829325	0.1629	0.0164	TRUE	1.41022E-11
	cg07770222	0.1623	0.0133	TRUE	2.00054E-12
	cg08795786	0.0786	0.0097	TRUE	1.79998E-05
	cg21015266	0.1583	0.0133	TRUE	6.24525E-12
	cg00002224	0.0738	0.0118	TRUE	0.000120955
	cg03029255	0.1952	0.0098	TRUE	2.66796E-17
	cg03058664	0.1641	0.0146	TRUE	3.22016E-12
Acute insulin response || id:ebi-a-GCST004575	cg22829325	0.1629	0.0018	TRUE	1.77536E-62
	cg07770222	0.1623	0.0033	TRUE	5.75245E-58
	cg08795786	0.0786	0.0039	TRUE	5.88005E-23
	cg21015266	0.1583	0.0016	TRUE	3.5686E-60
	cg00002224	0.0738	0.0041	TRUE	4.6217E-21
	cg03029255	0.1952	0.0002	TRUE	8.26789E-88
	cg03058664	0.1641	0.0016	TRUE	6.02956E-64
Acute insulin response (insulin secretion adjusted) || id:ebi-a-GCST006677	cg22829325	0.1629	0.0014	TRUE	4.16133E-64
	cg07770222	0.1623	0.0032	TRUE	4.36218E-58
	cg08795786	0.0786	0.0047	TRUE	8.3143E-22
	cg21015266	0.1583	0.0012	TRUE	4.42433E-62
	cg00002224	0.0738	0.0047	TRUE	3.64866E-20
	cg03029255	0.1952	0.0009	TRUE	1.25845E-81
	cg03058664	0.1641	0.0015	TRUE	1.88362E-64
Acute insulin response (insulin secretion and BMI adjusted) || id:ebi-a-GCST006678	cg22829325	0.1629	0.0016	TRUE	3.57814E-63
	cg07770222	0.1623	0.0032	TRUE	3.49771E-58
	cg08795786	0.0786	0.0058	TRUE	2.17951E-20
	cg21015266	0.1583	0.0015	TRUE	9.30592E-61
	cg00002224	0.0738	0.0059	TRUE	1.00769E-18
	cg03029255	0.1952	0.0010	TRUE	2.85413E-81
	cg03058664	0.1641	0.0018	TRUE	2.24228E-63