The Epigenomics of Human Ageing

University of Southampton

Academic Unit: Human Development and Health

A thesis for the degree of Doctor of Philosophy

ORCID: 0000-0002-2574-9611

Richard J. Acton

March 2021

Abstract

The global population is ageing and age is a primary risk factor for many life threatening diseases including cancer, cardiovascular and neurodegenerative conditions. This also encompasses conditions such as osteoporosis which increase the risk of death indirectly through elevated risk of the fracture of major bones and subsequent complications. The aim of this thesis was to extend the understanding of the epigenetic processes involved in ageing and ageing-related disease. Three specific areas were investigated.

Firstly, the relationship between early life epigenetic state and long term bone health was analysed. This investigation took the form of Epigenome-wide association studies (EWAS) conducted on umbilical cord blood DNA methylation. This identified candidate CpGs whose, DNA methylation state is associated with bone mineral content at 6 years of age ($p < 2.52 \times 10^{-8}$, n = 402) and periosteal circumference at 6 years of age ($p < 4.24 \times 10^{-8}$, n = 141) respectively.

Secondly, the changes in human tRNA gene DNA methylation with age were interogated. tRNAs permit the look-up of amino acids matching a given codon and as such are an essential core component of the translation of mRNA into protein by the ribosome. Human tRNA genes were found to be enriched for age-related DNA hypermethylation and three specific tRNA loci show genome-wide significant ($p < 4.34 \times 10^{-9}$) hypermethylation with age. Two of which, tRNA-iMet-CAT-1-4 and tRNA-Ser-AGA-2-6, were validated using the 450k Illumina DNA methylation array and replicated in an independent cohort, using targeted bisulfite sequencing.

Thirdly, a DNA methylation based age predictor based on the Alu family of SINE repeat elements was constructed. These repeats comprise a region of the genome not previously broadly accessible to DNA methylation assays used in the construction of age predictors. Age predictors using the DNA methylation state of Alu repeat elements were able to predict human age with an R of 0.65 and a median absolute error of 8.1 years. This predicition was possible using MeDIP-seq training data from 774 individuals and validated on an unrelated set of 664 individuals. An attempt was made to use ageing-related changes in Alu DNA methylation as a potential measure of 'biological' ageing. A Genome-wide Association study (GWAS) was performed for age accleration, the difference between predicted and chronological age. However, this age acceleration calculation was observed to be still strongly driven by actual age.

This work brings together epigenomic changes related to ageing from the beginning of the life-course through to later life, uniquely examining areas of the DNA methylome not previously studied in depth.

Publications arising from this Thesis

BioRχiv manuscript: The Genomic Loci of Specific Human tRNA Genes Exhibit Ageing-Related DNA Hypermethylation

RichardJ.Acton[1,2,3], *WeiYuan*[4,5], *FeiGao*[6], *YudongXia*[6], *EmmaBourne*[7], *EvaWozniak*[7], *JordanaBell*[4], *KarenLillycrop*[3], *JunWang*[6], *ElaineDennison*[2], *NicholasHarvey*[2], *CharlesA.Mein*[7], *TimD.Spector*[4], *PirroG.Hysi*[4], *CyrusCooper*[2], *ChristopherG.Bell*[1]

1. William Harvey Research Institute, Barts & The London School of Medicine and Dentistry, Charterhouse Square, Queen Mary University of London, London, U.K.
2. MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, U.K.
3. Human Development and Health, Institute of Developmental Sciences, University of Southampton, Southampton, U.K.
4. Department of Twin Research & Genetic Epidemiology, St Thomas Hospital, King's College London, London, U.K.
5. Institute of Cancer Research, Sutton, U.K.
6. BGI-Shenzhen, Shenzhen, China
7. Barts & The London Genome Centre, Blizard Institute, Barts & The London School of Medicine and Dentistry, Queen Mary University of London, London, U.K. *bioRχiv* [1]

Under Consideration at Nature Communications

Research Thesis: Declaration of Authorship

Name: Richard J. Acton

Title of thesis: The Epigenomics of Human Ageing

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as: a *BioRχiv* pre-print [1]

Signature: _____

Date: 2021-03-28

Contents

Abstract . . . 2
Publications arising from this Thesis . . . 3
Research Thesis: Declaration of Authorship . . . 4

Acknowledgements **29**

1 Introduction **31**
1.1 The Ageing Population and The Burden of Disease . . . 31
1.2 The biology of Ageing . . . 35
1.3 Epigenomics - Overview . . . 38
1.4 Fundamentals of DNA Methylation . . . 41
1.4.1 Structure and Context . . . 41
1.4.2 Distribution and Global Trends . . . 42
1.4.3 Pathways of DNA methylation and demethylation . . . 45
1.4.4 Maintenance and Fidelity . . . 46
1.4.5 DNA Methylation Assays . . . 50
1.5 DNA Methylation and Ageing . . . 53
1.5.1 Age-Related Differential Methylation and Epigenetic Clocks . . . 55
1.5.2 Genetic Influences on DNA Methylation . . . 58
1.5.3 DNAm Age as a Biomarker of Ageing . . . 60
1.5.4 DNA methylation clock measurements as outcomes for anti-ageing interventions . . . 62
1.6 Aims . . . 64

2 Methods **65**
2.1 Illumina DNA Methylation arrays . . . 65
2.2 MeDIP-seq . . . 70
2.2.1 Participants . . . 74
2.3 Targeted Bisulfite Sequencing . . . 75

3 Epigenome Wide Association Studies for Bone Health Outcomes in Umbilical Cord Blood and Tissue **79**
3.1 Abstract . . . 79
3.2 Introduction . . . 80
3.3 Primary Questions . . . 84
3.4 Methods . . . 84
3.4.1 Outline of EWAS . . . 84
3.4.2 Functional Normalisation . . . 85
3.4.3 Genetically Confounded and Multi-mapping Probes . . . 86
3.4.4 EWAS Models . . . 87
3.4.4.1 Cell-Type Correction . . . 87
3.4.4.2 Structure of models fitted for each EWAS . . . 88
3.4.5 Concordance of EPIC and 450k EWAS Results . . . 89
3.5 Results . . . 89
3.5.1 Combined MAVIDOS Summary data . . . 89
3.5.2 MAVIDOS phase I . . . 89

3.5.2.1 Whole array QC . . . 89
3.5.2.1.1 EPIC arrays . . . 91
3.5.2.1.2 450k Arrays . . . 94
3.5.2.2 Probe QC . . . 98
3.5.2.2.1 Probe QC - EPIC Arrays . . . 98
3.5.2.2.2 Probe QC - 450k Arrays . . . 102
3.5.2.3 Functional Normalisation . . . 105
3.5.2.4 EWASs . . . 107
3.5.2.4.1 Neonatal Bone Mineral Content . . . 107
3.5.2.4.2 Intervention / Placebo . . . 108
3.5.2.4.3 Maternal Vitamin D (34wks) . . . 109
3.5.2.4.4 Change in Maternal Vitamin D . . . 111
3.5.2.5 Concordance of EPIC and 450k EWAS results . . . 113
3.5.3 MAVIDOS phase II . . . 115
3.5.3.1 Whole Array QC . . . 115
3.5.3.2 Probe QC . . . 119
3.5.3.3 Functional Normalisation . . . 122
3.5.3.4 EWAS . . . 123
3.5.3.4.1 Bone Mineral Content at 4 years . . . 123
3.5.3.5 RXRA & CDKN2A . . . 125
3.5.4 Southampton Women's Survey (SWS) . . . 130
3.5.4.1 Whole Array QC . . . 130
3.5.4.2 Probe QC . . . 135
3.5.4.3 Functional Normalisation . . . 138
3.5.4.4 EWAS . . . 139
3.5.4.4.1 Total Bone Mineral Conent at 8 years . . . 140
3.5.4.4.2 Total Bone Mineral Conent at 6 years . . . 142
3.5.4.4.3 Periosteal Circumference at 6 years . . . 144
3.5.4.4.4 Cortical Denisty at 6 years . . . 146
3.6 Discussion . . . 148

4 The Genomic Loci of Specific Human tRNA Genes Exhibit Ageing-Related DNA Hypermethylation 151
4.1 Abstract . . . 151
4.2 Introduction . . . 153
4.3 Primary questions . . . 159
4.4 Methods . . . 160
4.4.1 Participants . . . 160
4.4.1.1 MeDIP-seq DNA methylomes . . . 160
4.4.1.2 Targeted Bisulfite sequencing . . . 160
4.4.2 tRNA Gene coordinates and other annotation information . . . 161
4.4.2.1 tRNA Gene Clustering . . . 161
4.4.2.2 tRNA gene mappability assessment . . . 162
4.4.3 DNA methylome data . . . 162
4.4.3.1 TwinsUK MeDIP-seq methylomes . . . 162
4.4.3.2 Analysis of DNA methylome data for Significant Ageing-related changes . . . 163
4.4.3.3 Permutation Analysis for Enrichment with Age-related Changes 163
4.4.3.4 Neonate and Centenarian Whole Genome Bisulfite Sequencing . 164
4.4.3.5 Sample pooling and EPIC array . . . 165
4.4.3.6 Targeted Bisulfite Sequencing . . . 165
4.4.3.7 TwinsUK Illumina 450k array methylomes . . . 166
4.4.4 Chromatin Segmentation Data . . . 166
4.4.5 Isolated Blood Cell Type Specific Data . . . 166
4.4.6 Cancer and Tissue Specific Methylation Data . . . 166
4.4.7 Assaying tRNA expression in blood with MINTmap . . . 167
4.4.8 Mouse RRBS Analysis . . . 167

4.5 Results . . . 168
4.5.1 DNA Methylation of Specific tRNA Gene Loci Changes with Age . . . 168
4.5.2 tRNA Genes are Enriched for Age Related DNA Hypermethylation . . . 174
4.5.2.1 tRNA gene clustering . . . 176
4.5.2.2 Age-related tRNAome DNA Hypermethylation is even observed in one Newborn versus one Centenarian . . . 177
4.5.2.3 Age-related Changes Independently Replicated with Targeted Bisulfite Sequencing . . . 178
4.5.2.3.1 Select Duplicates & Isodecoders of Hypermethylating tRNA loci remain unchanged . . . 181
4.5.2.4 DNA methylation 450k Array Data Validates the MeDIP-seq Results . . . 183
4.5.2.5 Ageing-Related tRNA Loci show increased Enhancer-Related Chromatin Signatures . . . 184
4.5.2.6 Age Hypermethylating tRNAs are more methylated in Lymphoid than Myeloid cells . . . 185
4.5.3 tRNA Gene DNA Methylation in Other Tissues . . . 187
4.5.3.1 tRNA Genes also Hypermethylate with Age in Solid Tissue . . . 188
4.5.3.2 tRNA Gene Methylation in Cancer . . . 190
4.5.4 Expression of tRNAs in Blood with Age . . . 192
4.5.4.1 MINTmap reference Fragment distribution . . . 194
4.5.5 Mice also show age-related tRNA gene DNA hypermethylation . . . 194
4.6 Discussion . . . 196

5 DNA methylation clocks in Alu Elements 201
5.1 Abstract . . . 201
5.2 Introduction . . . 202
5.3 Methods . . . 207
5.3.1 DNA methylome data . . . 207
5.3.2 Alu element annotation and pre-filtering . . . 207
5.3.3 Elastic net regression . . . 207
5.3.4 Sample Selection . . . 208
5.3.5 Binarisation . . . 211
5.3.6 Comparisons to other clocks in TwinsUK Illumina 450k array data . . . 212
5.3.7 Genome Wide Association studies of Alu Age Acceleration . . . 212
5.4 Results . . . 213
5.4.1 Comparison to other DNA methylation clocks . . . 215
5.4.2 Alu Age Acceleration . . . 216
5.4.3 Top results Genome Wide Association studies of Alu Age Acceleration are shared by Alu Age predictors . . . 220
5.5 Discussion . . . 225

6 Discussion 227
6.1 Epigenomic analysis of the developmental origins of long-term bone health . . . 227
6.2 The implications of ageing-related changes in the epigenetic state of tRNA genes 230
6.3 Assessing Biological ageing by DNA methylation changes within Alu repeat elements 233
6.4 Conclusion . . . 234

References 237

Appendices 291
6.5 *BioRχiv* manuscript: The Genomic Loci of Specific Human tRNA Genes Exhibit Ageing-Related DNA Hypermethylation . . . 291
6.6 Specific tRNA annotations used in the tRNA analyses . . . 292
6.7 tRNA targeted bisulfite sequencing primers . . . 292
6.8 Additional tRNA Age modelling results . . . 293
6.9 tRNA mouse models . . . 293

6.10 tRNA proportional enrichment fisher's test count data 293
6.11 TCGA sample used in tRNA analysis 293
6.12 Twins UK 450k array tRNA age models 293
6.13 MAVIDOS RXRA and CDKN2A probe models 293
6.14 Age acceleration GWAS SNPs 293

List of Figures

1.1 **The Population is Getting Older** Population pyramids for global population for A) 1950 and B) 2019 from populationpyramid.net [7] 32
1.2 **The Population is Likely to Continue Getting Older** Median projected increase in life expectancy from birth in the period 2010 to 2030 for 35 industrial nations reproduced from Kontis et al. 2017 figure 1 [2]. 32
1.3 **Mortality from Age-Related Conditions Increases According to a Logistic Function with Age** Mortality rate in deaths per 100,000 at different ages for Atherosclerosis, Essential Hypertension, Insulin-dependent diabetes, Non-insulin-dependent diabetes, Alzheimer's disease, and Parkinson's disease. Showing Model fits for exponential, logistic and the sum of 2 logistic functions. Underlying Cause of Death data (1999-2015) were form the Centers for Disease Control and Prevention, Wide-ranging OnLine Data for Epidemiologic Research (CDC WONDER) database. Reproduced from Belikov 2019 figure 1 [4]. 33
1.4 **Mortality from Infectious Diseases Increases With Age** Visits to General practitioners, hospital admissions and deaths from respiratory disease in persons aged: (a) 45-64, (b) 65-74 and (c) 75 years and over. Reproduced from Fleming et al. 2005 figure 3 [5]. 34
1.5 **Number and Frequency of Comorbidities increase with Age** Number of chronic disorders by age group. reproduced from Barnett et al. 2012 figure 1 [6]. 35
1.6 Epigenetic Changes as a hub for the hallmarks of ageing. (Reproduced from Booth & Brunet [30], figure 3) 38
1.7 The structure of 5-methylcytosine contrasted with Cytosine. (Figure created by the Author.) 42
1.8 Schematic representation of CpG Islands, shores, shelves and seas. CpG density declines from shores to seas and methylation increases (CpG density and methylation proportion not to scale). (Figure created by the Author.) 42
1.9 Stylised representation of the relationship between CpG Density, CpG methylation and functional DNA elements. (Figure reproduced from Baubec & Schubeler [87] figure 1) 44

1.10 The fraction of CpGs with a given methylation level. Data from Whole-Genome Bisulfite Sequencing (WGBS) of Mouse embryonic stem cells. Colours denote: unmethylated regions (blue), low-methylated regions (red), and fully methylated regions (green). These groups are three segments produced by a hidden markov model. low-methylated regions with 10-50% methylation are evolutionarily conserved distal regulatory elements with dynamic cell-type specific regulation unlike the consistently unmethylated regions at CGIs. (Figure reproduced from Stadler et al. [84] figure 1a) 44

1.11 Motif VI (ENV/PCQ) refers to the conserved catalytic motif in DNMTs. 'B:' Represents the enzyme supplied base (Figure reproduced from the review of DNMT biology by Lyko [93] (fig. 1b)) 45

1.12 BER = Bases Excision Repair; AM = Active Modification; AR = Active Removal, PD = Passive Dilution (Figure reproduced from the review of TET mediated active demethylation by Wu & Zhang [101] (fig. 1a)) 46

1.13 5mC levels over developmental time. 5mCH or 5-methylcytosine-(A, T or C) levels are represented on a separate (purple) Axis from the main 5mCG axis. PGC = Primordial Germ Cell. (Figure reproduced from [109] figure 3) 47

1.14 E_m = probability of methylation maintenance, E_m = probability of *de novo* methylation. M & U = the number of methylated and unmethylated molecules at specific CpG sites respectively. Modelling (**Grey points**) E_m, E_d values of 0.90,0.10; 0.95,0.05; 0.97,0.03 (outermost to innermost, two curves for each E_m, E_d set, one starting at $M = 1, U = 0$ and one a $M = 0, U = 1$. Experimental data (**Black points**) from the *HpaII* locus (site H3) in 17 clones of mouse cell line BML-2 which has a known methylation level of 50%. (Reproduced from Riggs & Xiong [117]) 48

1.15 Bimodal Distribution of DNA methylation level and fidelity. Methylation fidelity exhibits a bimodal distribution with most and least methylated regions exhibiting the highest fidelity. Additionally, fidelity for methylation levels of 10-50% is considerably better than fidelity for levels of 50-90%. These data are from the mouse cell line ES-E14TG2a which is capable of self-renewal when cultured with Leukaemia Inhibitory Factor (LIF) and spontaneous differentiation upon removal of LIF (denoted as E14-d0 at day 0 and E14-d6 at day 6 after the withdrawal of LIF, respectively). Fidelity represents the percentage of symmetrically methylated or unmethylated CpG dyads for a given position as determined by Hairpin-Bisulfite PCR. (Reproduced from Zhao et al. 2014 [122] (figure 3 a and b).) 50

1.16 Sodium Bisulfite conversion of C to U, 5mC and 5hmC are unaffected. (Image Credit New England Biolabs). 52

1.17 Mapping Chromosomal Regions of Differential DNA Methylation. Examples from Chromosomes 1, 3, 12 & 17 are shown for a 3 and 50-year-old twin pair. Methylation status is shown by competitive hybridization of AIMS (amplification of inter-methylated sites) products to metaphase chromosomes. Green and Red signals indicate hypermethylation and hypomethylation events between twins, Yellow indicates little difference. Red and Green blocks adjacent to ideograms indicate areas of significant DNA methylation change. (Reproduced from Fraga et al. [119] (figure 3).) 55

1.18 Global Hypomethylation with Age. NB = Newborn, Y26 = 26 year old, Y103 = 103 year old. (Reproduced from Heyn et al. [161] figure 1 b.) 57

1.19 Somatic mutations increase with Age. (Reproduced from Jaiswal et al. [182], figure 1) 59

1.20 C to T transitions are the most common substitutions that occur with age. (Reproduced from Jaiswal et al. [182], figure2 c) 60

1.21 Survival probability by quartiles of Δ_{Age} in LBC 1921 adjusted for sex, and chronological age. LBC = Lothian Birth Cohort. Using the Hannum and Horvath predictors' values for Δ_{Age}. (Reproduced from [190] figure 3.) 62

2.1 **Illumina Methylation Bead Chip Array Probe Designs** The probes on the array are 50bp in length. In type I probes, unlike type II probes the base that is incorporated (or not) in the elongation step is the base that **follows** the CpG site. The probes on the array are bound by the amplification products derived from the bisulfite converted DNA, not direct conversion products. Thus, Cs in the sample DNA represent methylated positions and Ts unmethylated positions, as Ts take the place of Us created by bisulfite conversion during the amplification process. **Type I**: Panel A shows the type I probe design. Each locus is represented on two beads, a methylated and an unmethylated bead. *Methylated Sample*: When sample DNA binds a methylated probe the G at the second position in the CpG on the methylated probe will be complementary to the C marking the location of the 5mC. Extension will then be able to occur incorporating a fluorescently labelled base complementary to the base just 5' of the 5mC position in the probe. On the unmethylated probe, the C marking the position of the 5mC will not be complementary to the A base in the probe and thus extension cannot occur. No fluorescent base will be incorporated at the unmethylated probe. *Unmethylated Sample*: When sample DNA binds the methylated probe the T marking the position of the unmethylated C is not complementary to the G at the second position in the CpG site on the probe. Thus, no fluorescent base is incorporated. Whereas, on the unmethylated probe the T marking the position of the unmethylated C will be complementary to the A at the second position in the CpG site on the probe. Thus, a fluorescently labelled base can be incorporated. **Type II**: Panel B shows the Type II probe design. For the Type II probes only one probe represents each locus. In the Type II probes the incorporated base is at the second position in the CpG site. A fluorescently labelled A is always incorporated opposite a T in the sample DNA marking the position of an unmethylated C, and a fluorescently labelled G is always incorporated opposite a C marking the position of an unconverted 5mC in the sample DNA. For the Type II probes, in contrast to the Type I probes, methylation is always signalled in the green channel and unmethylation in the red channel. (figure reproduced from Bibikova et al. [162].) 67

2.2 **Type I probes have a wider range and less variability than type II**. Type I probes shown in blue, Type II in red. Plot shows kernel density of beta values with gausian smoothing. (Reproduced from Dedeurwaerder et al. [215]) 69

2.3 MeDIP-seq provides superior coverage of repetitive regions of the genome than the Illumina array platform. This presumes saturation coverage for the MeDIP-seq assay. MD-s = MeDIP-seq, 450k = Illumina 450k DNA methylation bead chip array (Reproduced from Clark et al. [213]) The CpG islands, shores, and shelves major RefSeq gene features are covered almost as well in the 450k array as they are by MeDIP-seq but repetitive regions and other regulatory elements are where MeDIP-seq provides substantially greater coverage. 72

2.4 **Graphical Summary of the MeDIP-seq process** Purified DNA is fragmented by sonication, denatured and incubated with anti-5mC antibodies. It is then immunoprecipitated resulting in fragments containing methylated CpGs which are subsequently sequenced. 73

2.5 **Bisulfite sequencing alignment process A)** Unmethylated cytosines are converted to uracil in the bisulfite conversion process these are rendered as thymidine as T is used in the PCR amplification following bisulfite conversion. Following the conversion of a C to A T its corresponding G is swapped for an A in the complementary strand. The Genome is converted in silico to permit the alignment of these altered bases. Thus there is the C to T converted and a G to A converted genome each to which has two strands totalling 4 possible alignment targets for the converted reads. **B)** The unique best alignment for a read is determined from this set and mismatches of C in this alignment are used to call the methylation state. (Reproduced from Krueger et al. [236]) 76

2.6 **Microfluidic system for multiple parallel pairwise PCR reactions.** The Fluidigm 48.48 access array permits the combination of up-to 48 samples with up-to 48 different set of reagents, without multiplexing. In this case the reagents are locus specific primers for targeted sequencing. Samples are added to one side of the array and reagents to the other the flow cell permits all combinations of the two to be mixed. Fluidigm's controller flow cell and thermocycler allow all the PCR reactions to be carried out simultaneously and their products be extracted for sequencing. 77

2.7 **Four primer step-out PCR for targeted bisulfite sequencing on the Fluidigm access array.** The sample DNA is bisulfite converted. The region of interest amplified using site specific primers with the Fluidigm CS1 and CS2 primers in the first PCR step. The CS1/2 primers have regions complementary to the Illumina sequencing primers permitting them to be added in a second PCR step to generate the final PCR product ready for sequencing. 78

3.1 **Overall Fracture Incidence Increases With Age** Data from 5 million adults in the General Practice Research Database 1988-1998. **A** Incidence of all fractures at any site by sex and age. **B** Fractures by sex and age at select sites, pelvis, Femur/Hip & Vertebra confer the greatest increased mortality rates and Radius/Ulna has the greatest sexual dimorphism. (Adapted from van Staa et al. [238].) 81

3.2 **Bone Mass Over The Lifecourse** Bone mass increases from intrauterine development through to a peak in early adulthood. Intervention in early life to modulate growth trajectory and increase peak mass may provide a higher starting point and delay bone loss in later life. (Reproduced from [245].) 82

3.3 Diagrammatic representation of DNAm change arising from extrinsic or intrinsic changes in DNA methylation. Extrinsic changes are due to shifts in cell-type composition. Intrinsic changes in DNA methylation occur without changes in the proportions of cell-types. These two modes of change are of course not mutually exclusive and both can be occurring. 88

3.4 Predicted sex of each sample based on the sex chromosome copy numbers inferred from probe intensities for the EPIC array data. Mismatches between the predicted sex and that asserted in the sample annotation metadata are shown in red. Two predicted sex values differ from their annotations. Plot generated by `meffil` QC report. 91

3.5 Median methylated signal vs unmethylated signal per sample for the EPIC array data, solid red line indicates linear regression of median methylated signal vs median unmethylated signal with dotted red lines representing 3σ from the expected mean. Samples outside the expected range are indicated in the legend. Plot generated by `meffil` QC report. 92

3.6 Proportion of probes with detection p-values >0.01 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 93

3.7 Proportion of probes with a bead count of < 3 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 94

3.8 Predicted sex of each sample based on the sex chromosome copy numbers inferred from probe intensities for the 450k array data. No predicted sex values differ from their annotations. Plot generated by `meffil` QC report. 95

3.9 Median methylated signal vs unmethylated signal per sample for the 450k array data, solid red line indicates linear regression of median methylated signal vs median unmethylated signal with dotted red lines representing 3σ from the expected mean. Samples outside the expected range would be indicated in the legend. Plot generated by `meffil` QC report. 96

3.10 Proportion of probes with detection p-values >0.01 by sample for the 450k array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 97

3.11 Proportion of probes with a bead count of < 3 by sample for the 450k array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 98

3.12 Control probe signal by sample for each summary group for the EPIC data. Outliers would be circled in black. Plot generated by `meffil` QC report. 99

3.13 Undetectable probes across samples for EPIC data. Manhattan plot showing proportion of samples (y) in which a given probe (x) is not distinguishable from background noise, i.e. a detection p-value of > 0.01. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 100

3.14 Low bead count probes across samples for EPIC data. Manhattan plot showing the proportion of samples (y) in which a given probe (x) has a bead count of <3. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 101

3.15 An example of the DNAm distribution for a result from the `gaphunter()` function. This is an example chosen to best exemplify the sort of result which is strongly suggestive of a genetic variant with an impact on methylation status acting on this site. It is unrepresentative of typical results from `gaphunter()` in that the groups have a relatively even membership, many results have a small number of individuals in one or more groups making it hard to distinguish methylation outliers caused by rarer genetic variants from those with other causes. 102

3.16 Control probe signal by sample for each summary group for the 450k data. Outliers would be circled in black. Plot generated by `meffil` QC report. 103

3.17 Undetectable probes across samples for 450k data. Manhattan plot showing proportion of samples (y) in which a given probe (x) is not distinguishable from background noise, i.e. a detection p-value of > 0.01. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 104

3.18 Low bead count probes across samples for 450k data. Manhattan plot showing the proportion of samples (y) in which a given probe (x) has a bead count of <3. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 105

3.19 Residual variation remaining after functional normalisation of the top 20,000 most variable probes with m PCs from the control probe summary matrices for the EPIC array samples (n=137), for M = methylated and U = unmethylated probes. 106
3.20 Residual variation remaining after functional normalisation of the top 20,000 most variable probes with m PCs from the control probe summary matrices for the 450k array samples (n=60), for M = methylated and U = unmethylated probes. . . . 106
3.21 Distribution of Neonatal Bone Mineral Content (g) for individuals in the EWAS. 107
3.22 Results of EWAS for neonatal bone mineral content with SVA model. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. Size and transparency of points increases with $-log_{10}(p-value)$ such that the most significant CpGs are represented by the largest and least translucent points. x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 6.18×10^{-8} $(0.05 \div 808,585)$. 108
3.23 Results of EWAS for intervention/placebo group status with SVA model. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. Size and transparency of points increases with $-log_{10}(p-value)$ such that the most significant CpGs are represented by the largest and least translucent points. x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 6.18×10^{-8} $(0.05 \div 808,585)$. 109
3.24 Maternal circulating 25(OH)-vitamin D levels (nmol/l) at 11 and 34wks gestation, supplementation with 1000 IU/day cholecalciferol began at week 14. Each participant is shown at both time points linked by a line to indicate the direction of change. The violin plots indicate the density of the distribution of vitamin D values at each time point with the 25^{th}, 50^{th}, 75^{th} quantiles indicated with horizontal black lines. The colour indicates Intervention (Red) / Placebo (Blue) group 110
3.25 Results of EWAS for maternal circulating 25(OH)-vitamin D levels (nmol/l) levels at 34wks gestation with SVA model. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. Size and transparency of points increases with $-log_{10}(p-value)$ such that the most significant CpGss are represented by the largest and least transleucent points. x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 6.18×10^{-8} $(0.05 \div 808,585)$. . . . 111
3.26 Distribution of the changes in maternal circulating 25(OH)-vitamin D levels (nmol/l) levels from 11 to 34 weeks gestation. 112

3.27 Results of EWAS for change in maternal circulating 25(OH)-vitamin D levels (nmol/l) levels from 11 to 34wks gestation with SVA model. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. Size and transparency of points increases with $-log_{10}(p-value)$ such that the most significant CpGs are represented by the largest and least translucent points. x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 6.18×10^{-8} $(0.05 \div 808,585)$. 113

3.28 Concordance between the top 100,000 probes in common between the EWASs run on the EPIC (n=137) and 450k (n=60) data sets. bmc = bone mineral content, cc = Intervention / Placebo, difVD = Change in Vitamin D from 11 to 34wks, ExHighConc = Example of High Concordance generated using SVA vs iSVA results for the 450k intervention/placebo EWAS. Dotted line denotes concordance expected by chance (intersects 50% at 387,511, the number of shared probes). 114

3.29 Concordance between the top 10,000 probes in common between the EWASs run on the EPIC (n=137) and 405k (n=60) data sets. bmc = bone mineral content, cc = Intervention / Placebo, difVD = Chance in Vitamin D from 11 to 34wks, ExHighConc = Example of High Concordance generated using SVA vs iSVA results for the 450k intervention/placebo EWAS. Dotted line denotes concordance expected by chance (intersects 50% at 387,511, the number of shared probes). . . 115

3.30 Predicted sex of each sample based on the sex chromosome copy numbers inferred from probe intensities for the EPIC array data. Mismatches between the predicted sex and that asserted in the sample annotation metadata are shown in red. Two predicted sex values differ from their annotations. Plot generated by `meffil` QC report. 116

3.31 Median methylated signal vs unmethylated signal per sample for the EPIC array data, solid red line indicates linear regression of median methylated signal vs median unmethylated signal with dotted red lines representing 3σ from the expected mean. Samples outside the expected range are indicated in the legend. Plot generated by `meffil` QC report. 117

3.32 Proportion of probes with detection p-values >0.01 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 118

3.33 Proportion of probes with a bead count of < 3 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 119

3.34 Control probe signal by sample for each summary group for the EPIC data. Outliers would be circled in black. Plot generated by `meffil` QC report. 120
3.35 Undetectable probes across samples for EPIC data. Manhattan plot showing proportion of samples (y) in which a given probe (x) is not distinguishable from background noise, i.e. a detection p-value of > 0.01. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 121
3.36 Low bead count probes across samples for EPIC data. Manhattan plot showing the proportion of samples (y) in which a given probe (x) has a bead count of <3. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 122
3.37 Residual variation remaining after functional normalisation of the top 20,000 most variable probes with m PCs from the control probe summary matrices for the EPIC array samples (n=237), for M = methylated and U = unmethylated probes. 123
3.38 Distribution of whole body (minus head) bone mineral content in grams at 4 years of age. 124
3.39 Results of EWAS for whole body minus head bone mineral content (g) with SVA model (n = 237). Bidirectional Manhattan plot on which $-log_{10}(p - value)$ is plotted on the y axis and the sign of this value represents the direction of change. Size and transparency of points increases with $-log_{10}(p - value)$ such that the most significant CpGs are represented by the largest and least translucent points. x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 6.18×10^{-8} $(0.05 \div 808,585)$. 125
3.40 *RXRA* gene associated probes showing nominally significant (p<0.05) changes in DNA methylation with maternal vitamin D at 11 weeks gestation 128
3.41 *RXRA* gene associated probes showing nominally significant (p<0.05) changes in DNA methylation with maternal vitamin D at 34 weeks gestation 128
3.42 *CDKN2A* gene associated probes showing nominally significant (p<0.05) changes in DNA methylation with bone mineral density at 4 years 129
3.43 *CDKN2A* gene associated probes showing nominally significant (p<0.05) changes in DNA methylation with bone mineral density at 4 years 129
3.44 Predicted sex of each sample based on the sex chromosome copy numbers inferred from probe intensities for the EPIC array data. Mismatches between the predicted sex and that asserted in the sample annotation metadata are shown in red. Plot generated by `meffil` QC report. 132

3.45 Median methylated signal vs unmethylated signal per sample for the EPIC array data, solid red line indicates linear regression of median methylated signal vs median unmethylated signal with dotted red lines representing 3σ from the expected mean. Samples outside the expected range are indicated in the legend. Plot generated by `meffil` QC report. 133

3.46 Proportion of probes with detection p-values >0.01 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 134

3.47 Proportion of probes with a bead count of < 3 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 135

3.48 Control probe signal by sample for each summary group for the EPIC data. Outliers would be circled in black. Plot generated by `meffil` QC report. 136

3.49 Undetectable probes across samples for EPIC data. Manhattan plot showing proportion of samples (y) in which a given probe (x) is not distinguishable from background noise, i.e. a detection p-value of > 0.01. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 137

3.50 Low bead count probes across samples for EPIC data. Manhattan plot showing the proportion of samples (y) in which a given probe (x) has a bead count of <3. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report. 138

3.51 Residual variation remaining after functional normalisation of the top 20,000 most variable probes with m PCs from the control probe summary matrices for the EPIC array samples (n=464), for M = methylated and U = unmethylated probes. 139

3.52 Bone Mineral content at 6 and 8 years of are correlated with an $R^2 = 0.88$. . . . 140

3.53 Distribution of whole body (minus head) bone mineral content in kg (ttotbmcwhasa) at 8 years of age (n = 408), adjusted for sex and age, as measured by DXA. 141

3.54 Results of EWAS for whole body (minus head) bone mineral content in kg at 8 years of age (n = 408), adjusted for sex and age. The 'all' model results are shown here. Bidirectional Manhattan plot on which $-log_{10}(p - value)$ is plotted on the y axis and the sign of this value represents the direction of change. The x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 5.92×10^{-8}. 142

3.55 Distribution of whole body (minus head) bone mineral content in kg (stotbmcwhasa) at 6 years of age (n = 402), adjusted for sex and age, as measured by DXA. 143

3.56 Results of EWAS for whole body (minus head) bone mineral content in kg at 6 years of age (n = 402), adjusted for sex and age. The 'all' model results are shown here. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. The x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 5.92×10^{-8}. 144

3.57 Distribution of periosteal circumference at 38% from the distal end of the tibia (mm) at 6 years of age (s3peric), adjusted for sex and age, as measured by PQCT (n = 141). 145

3.58 Results of EWAS for periosteal circumference at 38% from the distal end of the tibia (mm) at 6 years of age (n = 141), adjusted for sex and age. The 'all' model results are shown here. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. The x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 5.92×10^{-8} 146

3.59 Distribution of cortical density at 38% from the distal end of the tibia ($mg\ cm^{-3}$) at 6 years of age (s3crtden), as measured by PQCT (n = 141). 147

3.60 Results of EWAS for cortical density at 38% from the distal end of the tibia ($mg\ cm^{-3}$) at 6 years of age (n = 141). The 'all' model results are shown here. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. The x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 5.92×10^{-8} 148

4.1 **Structure of a mature tRNA** Two and three dimensional representations of tRNA structure with matching colour coding. Adapted from the wikimedia foundation structure based on PDBID: 1ehz. tRNAs are 'charged' when an amino acid is attached at the CCA site at the 3' end. 154

4.2 **tRNA Transcription** Cartoon representation of the RNA polIII transcription initiation complex and structure of the type II RNA polymerase III promoter. Colour coding here corresponds to that in figure 4.1 illustrating that promoter is internal as the A box corresponds approximately to the D-loop and the B box to the T-loop in the tRNA structure. 155

4.3 **tRNA gene copy number is imperfectly correlated with amino acid and codon usage frequency A)** Amino acid usage frequency in the human exome vs tRNA gene copy number **B)** Codon usage frequency in the human exome vs tRNA gene copy number Usage Frequency Data [324], tRNA gene count data from GtRNAdb [304]. 156

4.4 **The Types and Functions of tRNA derived small RNAs** Reproduced from Cristodero et al. [341]. 157

4.5 **tRNA gene cluster numbers at different bin sizes** total: total number of tRNA clusters. singletons: number of tRNAs in clusters alone. moreThan1: number of tRNAs in clusters with more than one tRNA. maxSize: the number of tRNAs in the largest cluster. 162

4.6 CpG Density is higher in windows directly overlapping tRNA genes compared to that of non-tRNA overlapping windows in arbitrary flanking sequences (+/-5kb). This difference in CpG density between tRNA loci and other regions of the genome is a potential source of bias if age related DNA methylation changes vary with CpG density which they may as baseline DNA methylation levels also vary with CpG density. 164

4.7 **Study Structure** tRNAs differentially methylated with age initially identified in MeDIP-seq, validated (where covered in 450k array) and replicated in targeted bisulfite sequencing of pooled samples. Tissue specificity of these effects was explored in TCGA and foetal tissue data. 168

4.8 **The genetic code as instantiated in the human tRNAome.** The triplet genetic code leads to the incorporation of specific amino acids into an elongating protein via corresponding tRNAs. n is the number of tRNA genes which encode a given amino acid, the number in parentheses is how many of those may be pseudogenes based on their tRNAscan score [376], and the number in square brackets is the number in blacklisted regions [357]. There are a total of 610 tRNAs and closely related sequences in GtRNAdb [304], 416 of which are high confidence tRNAs, 116 of which are potential pseudogenes, and 107 are in blacklisted regions [357]. Notably 7 of the 61 non-STOP codons are missing from the human tRNAome therefore these codons are handled by wobble base matching (*e.g.* GCG Arg, ACC Gly). Also of note are the suppressor and selenocysteine tRNAs. The 20 methionine tRNAs are split equally between initiator methionine and internally incorporating methionine tRNAs, which are structurally distinct [299]. There are also 23 nuclear encoded mitochondrial tRNAs. 169

4.9 Example of mappability data from the encode mappability tracks [358] for the initiator methionine tRNA genes. 170

4.10 Mappability score density of the tRNAome increases with read length and is greater when flanking regions ($\pm 500bp$) are included. Mappability score density is computed as the area under the encode mappability tracks [358] over the length of the region. 171

4.11 Human tRNAome overview. From the outside in: Chromosome ideograms scaled by the number of tRNA genes (total = 598), as excludes chromosome X (10), Y (0) and contig chr1_gl000192_random (2; see Methods 4.4.2). tRNA genes within 20kbp of one another are grouped with breaks inserted between these clusters. Radial grey lines represent the location of tRNA genes in the genome. $-log_{10}(p-value)$ for the blood cell-type and batch corrected age model are shown for each window overlapping a tRNA gene in green. Mean methylation across all samples (n=3001) in RPM (reads per million base pairs) is shown in blue. Genome-wide significant cell-type & batch corrected ($p < 4.34 \times 10^{-9}$) tRNAs show in red. The 158 Loci covered by 213 probes on the 450k array which directly overlap a tRNA gene are shown with green triangles. The 84 loci targeted for bisulfite sequencing in this study are indicated in magenta. Mappability score density is computed as the area under the encode mappability tracks [358] over the length of the region. 173

4.12 **A)** tRNA genes are enriched for age-related hypermethylation compared to the genomic background, (Fisher's Exact Test p < 1.05×10^{-27}, n = 3001). **B)** tRNA genes show more significant hypermethylations than CpG Density matched permutations. Each permutation represented a random set of windows matching the CpG density of the functional unit (bivalent domains, polycomb group target promoters & the tRNA genes). These are subsequently assessed for signficant age-related DNAm changes (see Methods 4.4.3.3). The red line is the observed number of significant loci. 175

4.13 **A)** tRNA-iMet-CAT-1-4 as well as **B)** tRNA-Ser-AGA-2-6 and tRNA-Ile-AAT-4-1 exhibit age related DNA hypermethylation in MeDIP-seq data. The sharp peaks suggest that this effect is localised to individual tRNA genes. One of the windows overlapping tRNA-Ile-AAT-4-1 also partially overlaps tRNA-Ser-AGA-2-6. Results shown are from the blood cell-type and batch corrected model. Median Methylation in reads per million is calculated across all samples in the model. The window structure track (Wins) illustrates the tiled nature of the MeDIP-seq windowing. Some annotations of the 3' UTR of CTC1 extend further than illustrated here covering tRNA-Thr-AGT-1-2. CpG density and CpG position are included to illustrate that similarly CpG dense tRNA genes are exhibiting differing age related DNAm change patterns. 176

4.14 **A)** Almost all tRNA genes (42/44, counts shown in red) study-wide significant in the MeDIP batch corrected age model, and all of the genome-wide significant and blood-cell type corrected sets of tRNA genes reside in one of 12 major tRNA gene clusters. (Defined by joining all tRNAs within 5Mb of one another and requiring at least 5 tRNA genes per cluster with a density of at least 5 tRNA genes per Mb). **B)** Available study-wide significant (SWS) tRNA genes (n = 14) are more methylated in a centenarian than in a neonate or a 26 year old. Whole Genome Bisulfite Sequencing Data in a newborn, as adult and a centenarian. Each point represents the methylation level at an individual CpG within a tRNA gene. . . . 177

4.15 Mean chronological age is tightly correlated with DNAm Horvath clock [168] predicted age for the 8 pooled samples. (See Table 4.1 for pool details). 179

4.16 Individual CpG methylation increases (nominally significant $p < 0.05$) in tRNA-iMet-CAT-1-4. 180

4.17 Combined CpGs within tRNA loci results (experiment-wide Bonferroni $p = 7.14 \times 10^{-4}$); **(A-D)** Comparison of select tRNA-iMet-CAT loci: Hypermethylation is specific to iMet-CAT-1-4 (c) not other isodecoders (A, B, & D); **(F-I)** Comparison of select tRNA-Ser-AGA loci: Hypermethylation is specific to Ser-AGA-2-6 A (viii) and to a lesser extent Ser-AGA-2-1 (F), whilst not other isodecoders (G, H); **(K-P)** Comparison of select tRNA-Ile loci: Hypermethylation is specific to Ile-AAT-10-1 (K), Ile-AAT-4-1 (M) displays hypomethylation contrary to previous MeDIP findings, Ile-TAT-2-2 & 2-3 lack hypermethylation (previously non-significant in blood-corrected MeDIP, although significant in uncorrected), whilst no change in Ile-AAT-11-1 (L) and Ile-TAT-2-1 (N); **(R-U)** Comparison of select tRNA-Leu loci: Hypermethylation in Leu-AAG-3-1 (R) consistent with 450k and Leu-TAG-2-1 (U) consistent with MeDIP, whilst no change in Leu-TAA-1-1 (S) & Leu-TAA-3-1 (T). 182

4.18 Volcano-like plot. tRNAs are labelled if they are significant here or were in the MeDIP-seq data (Red). Model slope: the model coefficient for the methylation values. Unfilled circles indicate those probes in the general mask generated by Zhou et al. [223]. Significance threshold: $0.05/103 \approx 4.58 \times 10^{-4}$ (the number of tRNA genes examined). 184

4.19 Chromatin segmentation data from the Epilogos [368] 'Blood & T-cell' 15 State model (tRNA genes +/- 200bp). Frequency with which a model state was the predominant state at a given tRNA. Proportions of predominant tRNA state for the 14 study-wide significant age-hypermethylating tRNAs covered compared to other 371 available tRNAs. 185

4.20 Heatmap [379] of mean methylation of probes covering each tRNA in 7 cell-type fractions from 6 Male individuals. Data from GSE35069 [369]. Of the 16 study-wide significant hypermethylating tRNAs, 8 are covered by this dataset. 186
4.21 Heatmap Mean Methylation of probes covering each tRNA in 7 cell-type fractions from 6 Male individuals. Showing all 150 tRNAs covered by 213 probes on the Illumina 450k array. Data from GSE35069 [369] downloaded using `GEOquery` [370]. Generated with the `ComplexHeatmap` `R` package [379]. 187
4.22 Mean Methylation of 43 tRNAs in 19 tissues. Possible pseudogene (tRNA-Asn-ATT-1-1) is shown in a separate cluster beneath the main heatmap [379]. 189
4.23 Mean Methylation of 115 tRNAs in 11 tissues. Possible pseudogenes are shown in a separate cluster beneath the main heatmap [379]. 190
4.24 Global properties of tRNA methylation data for 45 tRNA genes across 19 tissues with matched normal and tumour samples from 733 cases in TCGA [372,373]. . . 192
4.25 tRNA fragments derived from the Gln-CTG family of tRNAs, selected as tRNA-Gln-CTG-7-1 is one of the 16 study-wide significant age-hypermethylating tRNA genes. Pane titles contain the MINTbase Plates, unique identifiers of the tRNA fragments [329]. 193
4.26 Comparison of the fragment size distributions between our custom reference **A)** and the original the MINTmap reference **B)**. 194
4.27 DNA methylation of CpGs in 3 tRNA which significantly hypermethylate with age in mice, in data from Petkovich et al. [375]. 6 CpGs reach Bonferroni significance and 7 show nominally significant increases. 195

5.1 **Proportion of the Human genome given over to different repeat classes** 'Rest' indicates the non-repeat features. Reproduced form Jorda et al. [410] Figure 1 b. 202
5.2 **Structure of an Alu element** A complete Alu element contains the A and B box sequence elements from an RNA polIII promoter, a central A rich region, a short insertion in the right arm, and a poly A region involved in retrotransposition and insertion. There is not an explicit polIII terminator in the Alu sequence so transcription continues through arbitrary flanking sequence until a run of at least 4 Ts, is encountered [417,418]. FLAM sequences are the origin of the Left arm and FRAMs the origin of the Right [416]. 203
5.3 **Alu element Age and history of expansion in primates** Mya (milion years ago). Reproduced from Batzer & Deininger 2002 [418]. 204

5.4 **Alu elements are CpG dense** FAM, FLAM and FRAM sequences are the precursors which gave rise to the modern Alu families, they are less CpG dense than the AluS and AluY families and closer to the global average of ~1.8% Chapter 1.4.2. 205

5.5 Comparison of the overlap of different repeat classes [437] with four DNA methylation clocks (Horvath clock [168], Hannum et al. clock [167], PhenoAge clock [193], epiTOC clock [372]) as well as Illumina array CpGs (27k [154], 450k [162], EPIC [210]) and all CpGs in the genome. Reproduced from Bell et al. [434] figure 2H. 206

5.6 **Construction of training and replication sets** Twin pairs were split one twin in each group. The first group of twins was split into two groups with a random number generator. The twins of the individuals in one random half of the group were assigned to be the replication set for the other half, producing two sets of unrelated training and replication groups. Models were also used to predict the age of related individuals to gauge the effect of genetic factors on the predictive power of the models. 208

5.7 **Age distribution of training and replication sets** Training 1 (n = 774, 64M/710F); Training 2 (n = 774, 65M/709F); Replication 1 (n = 664, 55M/609F), unrelated to Training 1; Replication 2 (n = 644, 55M/589F), unrelated to Training 2. M = Male, F = Female 209

5.8 **Zygosity Make-up of sets** Replication 1 (n = 664, 225 DZ, 439 MZ); Replication 2 (n = 644, 204 DZ, 440 MZ), Training 1 (n = 774, 256 DZ, 518 MZ); Training 2 (n = 774, 269 DZ, 505 MZ). 210

5.9 **Correlation of Predicted with Chronological Age Across Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus. 214

5.10 **Correlation of Alu Age Acceleration with Chronological Age Across Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. Alu Clock here refers to the unfiltered model trained on the Training 1 set. . . . 215

5.11 **Correlation of Alu Age Acceleration with Chronological Age Across Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. Alu Clock here refers to the unfiltered model trained on the Training 1 set. . . . 216

5.12 **Correlation of Alu Age Acceleration with Chronological Age Across Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus. 217

5.13 **Correlation of Age Acceleration with Chronological Age Across Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. Alu Clock here refers to the unfiltered model trained on the Training 1 set. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus. 218

5.14 **Correlation of Alu Age Accelerations with Age Acceleration from Other Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. Alu Clock here refers to the unfiltered model trained on the Training 1 set. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus. 219

5.15 Manhattan plot showing the results of a GWAS for Alu Age acceleration computed with the predictor trained using Training set 1 on unfiltered windows. 220

5.16 Manhattan plot showing the results of a GWAS for Alu Age acceleration computed with the predictor trained using Training set 1 on windows filtered to have an element length >180bp and 99.9% non-zero RPM values at each locus. 221

5.17 Top four shared SNPs are all located just upstream of the the *CCDC91* gene. Adapted from the output of LocusZoom [450]. 223

5.18 *Phenotypes Common to the top 4 shared SNPs.* N = number of the top 4 SNPs in which a term appears in the top 10 Phenotypes Top four that SNP in PheWeb [453]. 224

6.1 https://doi.org/10.1101/870352 292

List of Tables

3.1 Summary statistics for variables used in the SWS EWAS. Blood cell count information is estimated based on the methylation data as described in the methods. 89

3.2 Summary statistics for variables used in the MAVIDOS EWAS divided by the sex of the child. Blood cell count information is estimated based on the methylation data as described in the methods. n=375, Female n=170, Male n=205. 90

3.3 Results for modelling Maternal Vitamin D levels with umbilical cord tissue DNA methylation levels and covariates at probes associated with the *RXRA* and *CDKN2A* genes. vit D: maternal circulating vitamin D nmol/l 126

3.4 Results for modelling bone mineral content and density at 4 years of age with umbilical cord tissue DNA methylation levels and covariates at probes associated with the *RXRA* and *CDKN2A* genes. Corrected: BMC/D ~ beta + wageb + Sex + awbmi + smokpreg + egest + Bcell + CD4T + CD8T + Gran + Mono + NK + Endothelial + Epithelial + Stromal, Uncorrected: BMC/D ~ beta, beta: methylation value, BMC/D Bone Mineral Content (g)/Density(gcm-2) at 4 years (DXA)‘ 127

3.5 Summary statistics for variables used in the SWS EWAS. Blood cell count information is estimated based on the methylation data as described in the methods. (Ob.) obstetric exam, (init.) initial survey. 130

3.6 Summary statistics for variables used in the SWS EWAS divided by the sex of the child. Blood cell count information is estimated based on the methylation data as described in the methods. n = 464, Female n=226, Male n=238. (Ob.) obstetric exam, (init.) initial survey. 131

4.1 Summary information on participants in each pool. 165

4.2 Study-wide significantly hypermethylating tRNAs in blood cell-type and batch corrected model MeDIP-seq. With Corresponding results in Twins UK 450k array (blood cell-type corrected) and targeted bisulfite sequencing results. Age models for array and targeted BiS-seq were calculated using all probes / CpGs overlapping the indicated tRNA gene. 'Slope' corresponds to the beta value for methylation in the linear model, orange colouring indicates hypermethylation and blue hypomethylation. p-values coloured such that low values are dark blue and high values are yellow. blank grey cells indicate missing data. 172

4.3 Pairwise Differences in Methylation between Age groups by tRNA. p-values are for pairwise methylation differences (see Methods 4.4.3.6)[366]. 181

5.1 **Alu Clock Model Summaries** *A)* Models fitted with Training 1 set predicting Replication 1 ages. *B)* Models fitted with Training 2 set predicting Replication 2 ages. R is Pearson's correlation coefficient. RMSE = root mean squared Error, MAE = median absolute error. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus. 213

5.2 **Summaries of Alu Clock Models Predicting Twin Groups** for the purpose of comparing predictive of age models in related and unrelated samples. *A)* Models fitted with Training 1 set predicting Replication 2 ages. *B)* Models fitted with Training 2 set predicting Replication 1 ages. R is Pearson's correlation coefficient. RMSE = root mean squared Error, MAE = median absolute error. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus. 213

5.3 Top 10 SNPs from model 1. 222

5.4 Top 10 SNPs from model 2. 223

Acknowledgements

I would like to acknowledge Dr Nevena Krstic for her work extracting the DNA for the MAVIDOS EPIC array analysis and Dr Millie Parsons for her assistance with with the MAVIDOS sample metadata, as well as Dr. Beth Curtis and Professor Nick Harvey of the MRC-LEU for their assistance with framing the research questions for the vitamin D and bone development outcomes work with the MAVIDOS samples. The MRC-LEU is supported by the Medical Research Council (MRC). I gratefully acknowledge the individuals from TwinsUK, Mavidos and the Hertfordshrie cohort. I would like to thank Dr. Pirro Hysi for his work performing the age acceleration GWAS. I would like to acknowledge Nikki Graham for her assistance identifying suitable samples for targeted bisulfite sequencing work and her work extracting DNA from those samples. I acknowledge the use of the IRIDIS High-Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. I would like to thank the MRC Doctoral fund for supporting this work. I thank my family for their invaluable support, encouragement, and occasional proofreading. Thanks also to Dr Michael Glinka for his advice, encouragement and friendship. I thank my supervisors Professors Karen Lillycrop and Cyrus Cooper for their input and advice on my project. Finally, I would like to acknowledge the extensive help and support of my primary supervisor Dr Chris Bell.

Chapter 1

Introduction

1.1 The Ageing Population and The Burden of Disease

The global population is ageing as is evident from the changing shape of the world population pyramid (figure 1.1). The increase in average lifespan underlying this demographic shift is projected to continue in industrialised nations with a probability of >65% for women and 85% for men [2] (figure 1.2). Ageing is the main risk factor for cancer, cardiovascular, neurodegenerative diseases, and many other conditions including osteoarthritis [3]. Furthermore, mortality from these conditions increases according to a logistic function with age [4] (figure 1.3). In addition to the non-infectious diseases, immunosenescence and frailty in old age contribute to increased mortality from infectious diseases, as exemplified by dramatically higher hospital admissions and deaths for respiratory disease during seasonal flu epidemics in the over 75s [5] (figure 1.4). Ageing is a common underlying risk factor for many conditions and as this would lead one to expect the number of concomitant disorders and the proportion of persons with multiple morbidities increases with age [6] (figure 1.5).

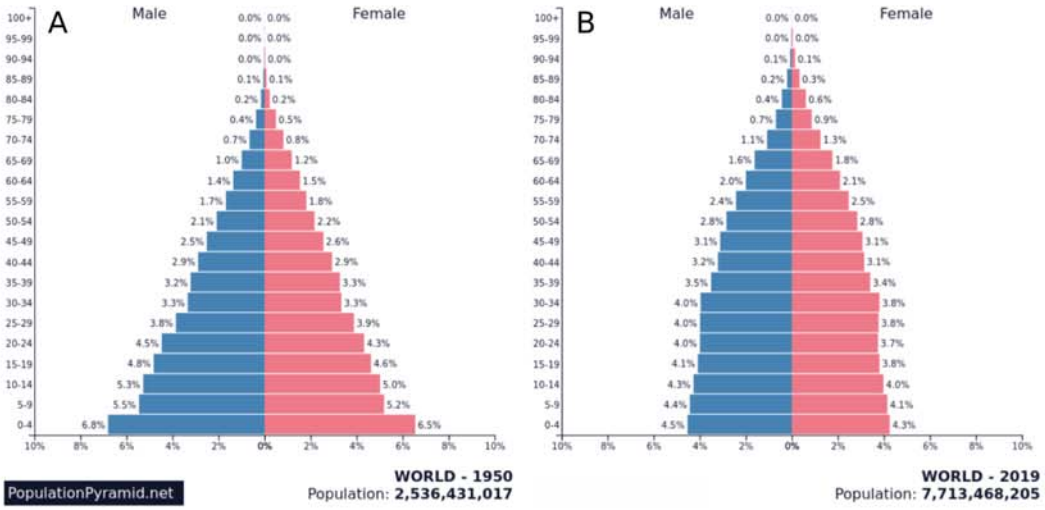

Figure 1.1: **The Population is Getting Older** Population pyramids for global population for A) 1950 and B) 2019 from populationpyramid.net [7]

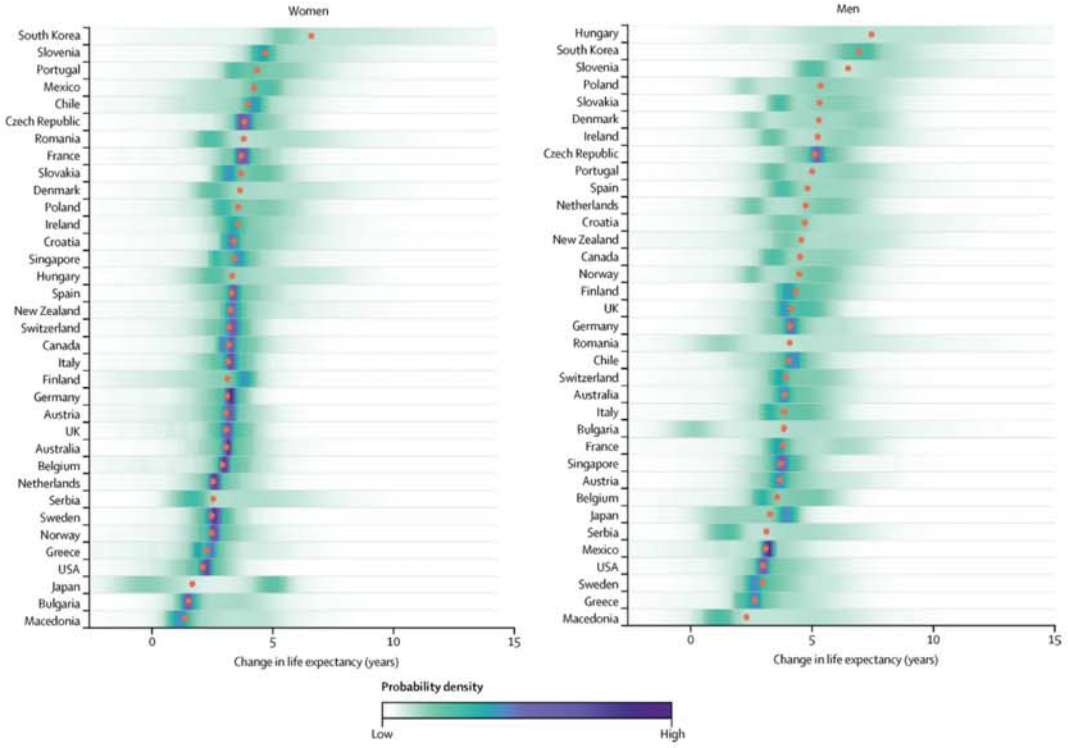

Figure 1.2: **The Population is Likely to Continue Getting Older** Median projected increase in life expectancy from birth in the period 2010 to 2030 for 35 industrial nations reproduced from Kontis et al. 2017 figure 1 [2].

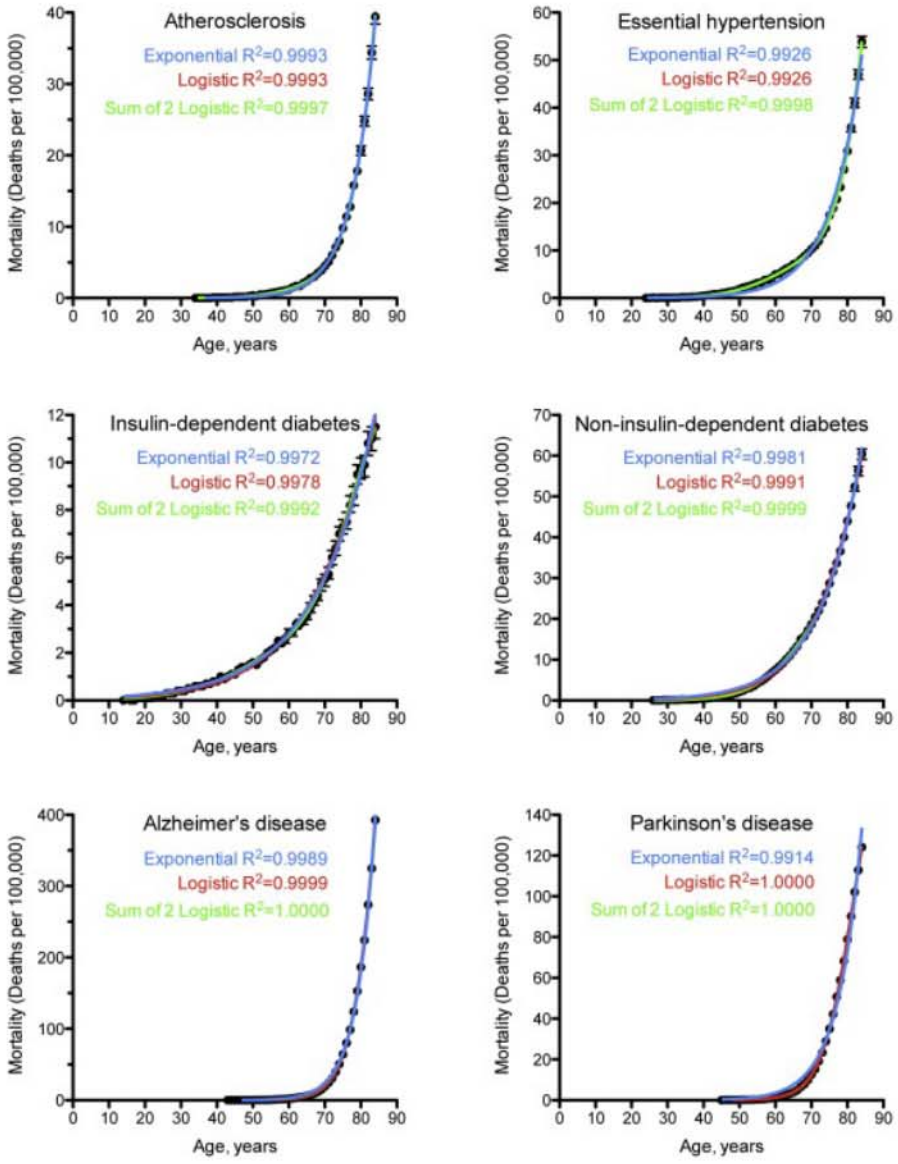

Figure 1.3: **Mortality from Age-Related Conditions Increases According to a Logistic Function with Age** Mortality rate in deaths per 100,000 at different ages for Atherosclerosis, Essential Hypertension, Insulin-dependent diabetes, Non-insulin-dependent diabetes, Alzheimer's disease, and Parkinson's disease. Showing Model fits for exponential, logistic and the sum of 2 logistic functions. Underlying Cause of Death data (1999-2015) were form the Centers for Disease Control and Prevention, Wide-ranging OnLine Data for Epidemiologic Research (CDC WONDER) database. Reproduced from Belikov 2019 figure 1 [4].

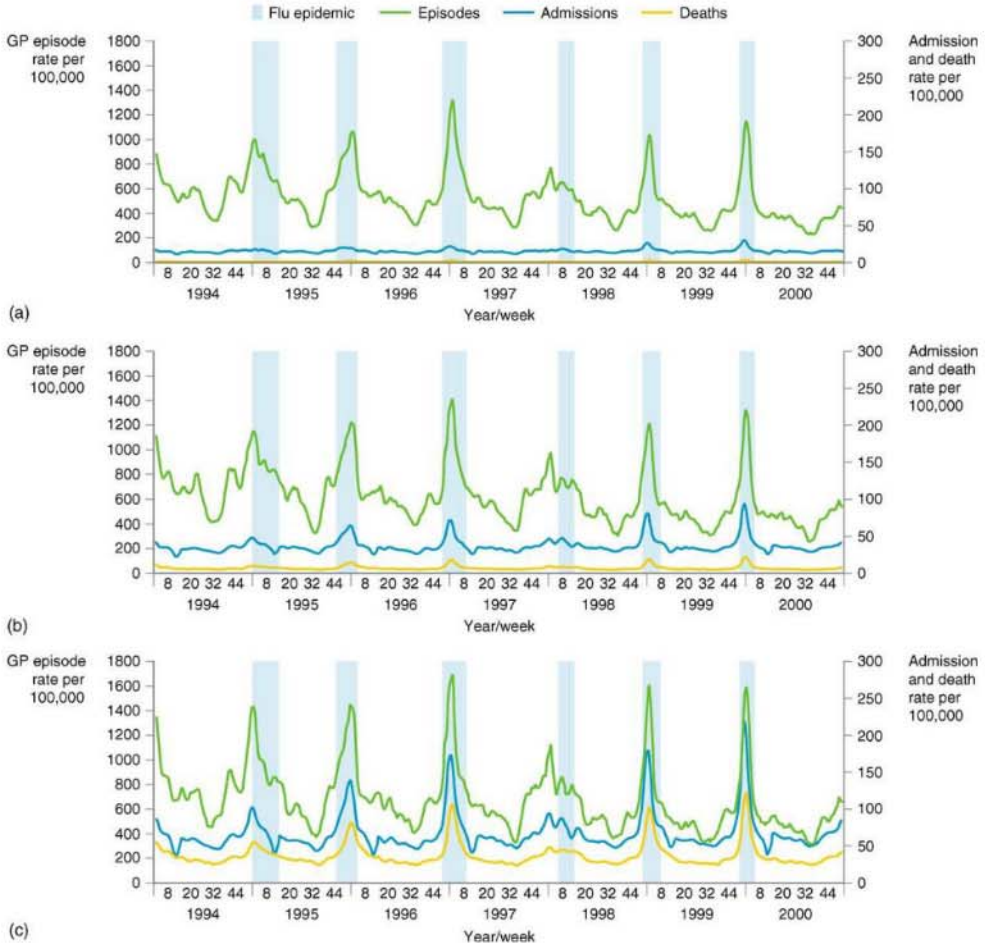

Figure 1.4: **Mortality from Infectious Diseases Increases With Age** Visits to General practitioners, hospital admissions and deaths from respiratory disease in persons aged: (a) 45-64, (b) 65-74 and (c) 75 years and over. Reproduced from Fleming et al. 2005 figure 3 [5].

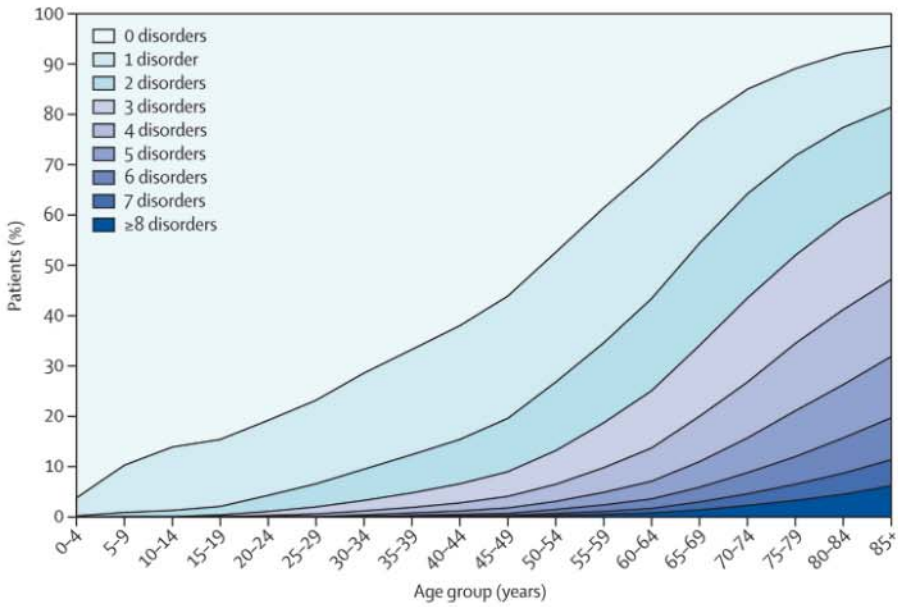

Figure 1.5: **Number and Frequency of Comorbidities increase with Age** Number of chronic disorders by age group. reproduced from Barnett et al. 2012 figure 1 [6].

Given the rapid increases in the rates of numerous diseases and the rate of death from those diseases as we age, there is a clear impetus to develop interventions to prevent and treat age-related diseases. The common factor among all these conditions is their relationship to age, but age is not merely a shared risk factor but frequently the largest; outside of rare genetic conditions which predispose to early-onset [3]. Thus understanding the underlying biological mechanisms of the ageing process and intervening in those mechanisms, has the potential to be vastly more effective and economical [8] than attempting to tackle diseases individually. The discovery in 1993 by Cynthia Kenyon and others [9] that a single mutation in the *daf-2* gene of *C. elegans* could double the lifespan of this organism is often cited as an inflexion point in the confidence of the ageing research field that interventions in the process of ageing itself may be a realistic therapeutic target. What then do we know about the underlying biology of ageing?

1.2 The biology of Ageing

Ageing is characterised by a progressive deterioration of physiological integrity with time, resulting in an increased risk of mortality and morbidity. Theories of how and why this occurs are numerous with a history in the scientific literature stretching back to the late 1800s [10], setting aside, for now, theoretical frameworks for the process of ageing we will first consider the nature of these changes. Changes at the organismal level have their origins in changes at the molecular and cellular scales. Lopez-Otin et al. 2013 [11] identified nine hallmarks of ageing in their 2013 review. These hallmarks meet the criteria of: Manifesting during normal ageing; Accelerating

ageing when aggravated; Retarding ageing when ameliorated. The hallmarks are further subdivided and hierarchically arranged into primary, antagonistic and integrative categories. Primary hallmarks are of unequivocally negative effect. Antagonistic hallmarks arise as a response to the primary and are initially protective against these changes but become problems in their own right as they occur at higher levels over time. Lastly, the integrative hallmarks directly impact on tissue function and homoeostasis. The hallmarks are as follows:

Primary Hallmarks - Causes of Damage

1. *Genomic instability*: Damage to the genomes of our cells occurs over time and accumulates as we age, this damage ranges from point mutations to large structural mutations, as well as mutations in the mitochondrial genome [12].

2. *Telomere attrition*: Telomere shortening occurs during the ageing process, somatic cells do not generally express telomerase so with each division telomeres grow shorter. Telomere exhaustion limits the replicative capacity of cells inducing replicative senescence, limiting regenerative potential but also protecting from the uncontrolled growth of cancer [13].

3. **Epigenetic alterations**: Patterns of DNA methylation and of a number of histone modifications exhibit age-related changes. These contribute to the dysregulation of gene expression and derepression of heterochromatin regions which also increases the risk of genomic instability [14].

4. *Loss of proteostasis*: Production of protein folding chaperones in response to stress decreases with age, as does the activity of the two major proteolytic systems, the ubiquitin-proteasome system and the autophagy-lysosome system. This leads to increased prevalence of misfolded, aberrant and aggregated proteins with age [15].

Antagonistic Hallmarks - Responses to damage

5. *Deregulated nutrient sensing*: The IIS (insulin & insulin-like growth signalling) and TOR (Target Of Rapamycin) pathways signal nutrient abundance, promoting anabolism. The AMPK (5' adenosine monophosphate-activated protein kinase) and Sirtuins signal nutrient scarcity, promoting catabolism. Broadly speaking inhibiting the pro-anabolic pathways and activating the pro-catabolic pathways can extend lifespan [16,17].

6. *Mitochondrial dysfunction*: Mitochondrial ATP production drops off with and ROS (Reactive Oxygen Species) production increases. Elevated ROS do not appear to accelerate ageing nor do elevated antioxidants retard it. Indicating the primary function of ROS in ageing is not straightforwardly a result of oxidative damage, but may result from its role in stress signalling. Dysfunctional mitochondria exhibit an increased propensity to polarise in response to stress, impacting on apoptotic signalling and inflammatory responses. This contributes to the increases in senescent cells and their chronic inflammatory phenotype

[18].

7. *Cellular senescence*: Senescent cells are in stable cell-cycle arrest induced by shortened telomeres, other forms of DNA damage, INK4/ARF derepression, or a variety of mitogenic/oncogenic signals. Induction of a senescent state in pre-cancerous cells protects against cancer and senescent cells are initially cleared by the immune system. However, these cells accumulate in tissue as we age without contributing to function and with a pro-inflammatory secretory phenotype that promotes chronic inflammation [19].

Integrative Hallmarks - Proximal Causes of the Ageing phenotype

8. *Stem cell exhaustion*: Adult stem cells capable of producing new cells for the regeneration of tissues undergo asymmetric divisions to replace themselves and produce replicative progenitor cells to replenish tissue. Stem cells may accumulate damage and cease replicating or leave their quiescent state becoming overly replicative and potentially becoming senescent. This disrupts the supply of new cells for the renewal of tissue [20].

9. *Altered intercellular communication*: The aforementioned cell-autonomous alterations alter external signalling behaviour of cells across intercellular signalling paradigms. This can lead to feedback loops that exacerbate the problem. The high background of pro-inflammatory signals from senescent cells dilutes already declining immune response and fewer senescent cells are cleared [21].

Many genetic changes have been associated with changes in longevity since Cynthia Kenyon's characterisation of the *daf-2* mutation which approximately doubled the lifespan of *C. elegans* [22–24].

In Humans Genome Wide Association Studies have identified relatively few genes associated with longevity [25,26]. APOE is perhaps the most consistently identified, followed by FOXO3A, however FOXO3A appears not to be as predictive for the majority of the population and is only associated with longevity in the very long lived [27]. The portion of variance in longevity attributable to genetics may adhere to the emerging pattern of the genetic contibution to common diseases being due to a multitude of common variants with small effects. With rarer mutations having more extreme effects on longevity [28]. There is also an effect of ultra-rare/private protein-truncating variants, a high burden of which negatively affects longevity [29].

This work and therefore the following review of the literature focuses on the 3rd hallmark, epigenetic alterations, specifically the role of DNA methylation in ageing. Booth & Brunet [30] contend that epigenetic processes are a hub through which all of the other hallmarks of ageing are mediated and feed-back on one another (Figure 1.6). One of these interactions is between DNA damage and chromatin state including DNA methylation. Sinclair and Oberdoerffer [31] make the case that DNA damage and repair processes are disruptive to the epigenome and, with Hayano et al., link double-stranded DNA breaks to accelerated epigenetic ageing in a

recent pre-print [32]. Kane & Sinclair [33] make the case that reprogramming of the epigenome to a 'younger' state is a promising mode of intervention in the ageing process. Whilst loss of genetic information is effectively irreversible, barring gene therapy-like interventions, loss of epigenetic information is not necessarily so. Dysregulation from the loss of epigenetic information is arguably proximate to age-related changes and may run ahead of significant genetic information loss such that intervening in the former may have beneficial effects in its own right and stave off the latter.

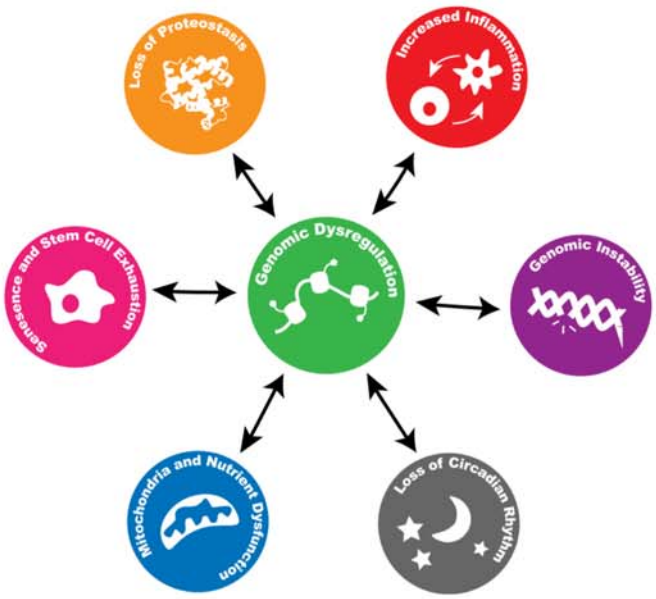

Figure 1.6: Epigenetic Changes as a hub for the hallmarks of ageing. (Reproduced from Booth & Brunet [30], figure 3)

Prior to examining the specific relationship of epigenetic changes to ageing we take a step back to define and outline epigenomics and examine the fundamentals of DNA methylation biology.

1.3 Epigenomics - Overview

Epigenetics generally refers to modifications to DNA and chromatin which do not affect the primary sequence of DNA bases [34], but which are to varying degrees stable and heritable. The term derives from epigenesis and genetics originating with Conrad Waddington [35]. Epigenesis refers to the idea that organisms develop through the progressive differentiation of cells from the egg into adult tissues. As modern genetics revealed that every cell contained a complete copy of the genome which was differentially utilised by the cells of adult organisms, the terms were fused to reflect the study of how this occurs. The usage has evolved a little further since molecular

biology began to elucidate the mechanisms involved in this process and the term now frequently refers to the study of these mechanisms and their effects in less explicitly developmental contexts. Arthur Riggs et al. considered that heritability should be a criterion for a mark to be considered epigenetic, but this excludes many phenomena now commonly referred to using the term [36]. Requiring heritability results in a further definitional dispute over degrees of heritability; mitotic, meiotic, intergenerational, transgenerational and to what degree of fidelity? Adrian Bird proposed the definition: "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states." as a useful compromise [34]. Epigenomics refers to the totality of the epigenetic modifications present in a particular cell, tissue type or genome. This definition appears to have caught on with at least some as: 'These days, "epigenetics" more generally refers to all non-genomic information storage in cells including gene networks, chromatin structure and post-translational modifications to histones.' according to Alice Kane and David Sinclair [33].

Whilst an organism can generally be thought of as having a single genome, with exceptions such as Somatic mutations and Chimerism, it will have at least as many epigenomes as it has cell types [37]. There are on the order of 10^{13} cells in the human body [38]. Estimates of the number of cell types vary with the resolution at which one deems cells functionally distinct [39]. At this point in time, however, there is no clear definition of what constitutes a distinct cell type, indeed they are beginning to be defined by the distinct patterns of gene expression and epigenetic modifications they exhibit [40]. Consequently, the number epigenomes that could be considered distinct will likely ultimately approximate to the number of epigenomes which can usefully distinguish between sub-populations of cells. A consortium has been established to produce a Human Cell Atlas [39] which aims to define all human cell types, and a cell type ontology [41] exists.

There are a number of epigenetic modifications which can be subdivided into four broad categories:

1. **DNA modifications**

   In mammals DNA methylation primarily takes place on Cytosine residues. Cytosine methylation occurs principally in an mCpG sequence context but can also occur in mCpH, mCHG and mCHH (H=A/C/T) contexts; particularly in cells of the nervous system [42] and embryonic stem cells where as many as 25% of all cytosines can be methylated in non-canonical contexts. In contrast to methylation in a differentiated non-neuronal tissue such as foetal lung fibroblasts which is 99.98% in CpG context [43,44]. 5-methylcytosine (5mc) can be oxidised to produce another modified DNA base 5-hydroxymethylcytosine (5hmC) [45,46]. 5hmC may act as an intermediate to demethylation and potentially has regulatory functions in its own right [47]. 5hmC can be oxidised further to formylcytosine (5fC) and carboxylcytosine (5caC) [48]. The extent to which these modifications are sta-

ble and functional is still being explored. Other DNA bases can be methylated such as N6-methyladenine, but this occurs at a substantially lower frequency that 5mC and much less is known about their potential functions [49]. N6-methyladenine may play a role in condensing chromatin [50].

2. **Histone tail modifications**

   Histone proteins form disk-shaped octamers around which ~150bp of DNA can be wrapped to form a nucleosome. The 'tails' are generally the N termini of the histone proteins, outside of the core globular domains, which protrude from the nucleosome structure [51]. Histone tails can be subject to post-translational modification which commonly takes place at lysine residues, other residues are subject to modifications but lysines are among the best characterised. Examples of modifications include methylation with between one and three methyl groups per lysine, Acetylations, Ubiquitylation, SUMOylations, and phosphorylations [52]. Many of these marks can be generally classified as permissive or repressive but are frequently found in combinations of marks of opposing effect, rendering the interpretation of the 'Histone Code' extremely challenging [53–55]. One class of approaches to interpreting chromatin state is segmentation, this covers a variety of methods for combining mostly ChIP data for an array of histone modifications generated by large consortia projects such as ENCODE [56]. Segmentation makes use of pattern discovery techniques such as Hidden Markov Models to divide the genome into discrete sections assigning these sections to a set of categories which can then be compared to existing annotations to provide functional labels [57,58]. When successful the pattern discovery algorithm independently recapitulates our existing ontology of functional elements and hopefully provides new insights by for example highlighting regions not previously considered to be in a given functional category. This is a "top-down" approach making use of the "sum" of available chromatin state data rather than trying to interpret the significance of a single type of mark.

3. **Histone Variants**

   Alternatives to the canonical histones can be substituted in the nucleosome, examples of such alternative histones include H2A.Z and H3.3 [59]. Alternative histones can alter chromatin structure and dynamics by altering nucleosome stability, binding different factors, and presenting a different substrate for histone modifications, thus altering their pattern [60].

4. **Some non-coding RNAs (ncRNA)**

   A subset of long ncRNAs, particularly those which persist in close association with chromatin can fall within the definition of epigenetic features [61]. An example of such a long non-coding RNA (lncRNA) is Xist which is a key regulator of X inactivation [62]. The catagorisation of these as epigenetic is somewhat controversial, as they are not direct

chemical modifications to chromatin. However Xist, for example, facilitates the mitotically heritable repression of the inactivate X chromosome satisfying other criteria for being considered an epigenetic modification.

1.4 Fundamentals of DNA Methylation

1.4.1 Structure and Context

DNA methylation is the most well-studied epigenetic modification [63]. Specifically CpG methylation, the addition of a Methyl group to the 5 carbon of a cytosine base (figure 1.7), in a CpG dinucleotide context. (hereafter 'DNA Methylation' or 'DNAm' refers to 5mC CpG methylation unless stated otherwise.) DNA methylation is stable and relatively easy to characterise from biological samples by comparison with other epigenetic marks. The methylation status of ancient DNA has even been characterised, for example, native American remains ranging in age from 230 to 4500 years were measured using bisulfite sequencing [64]. C undergoes spontaneous hydrolytic deamination over time to U, and 5mC to T. once sufficient time has passed for the deamination process to convert the majority of the 5mC sites, this can be used to infer the methylation state of post-mortem DNA using DNA polymerases that cannot bypass uracils and thus only causes C $\rightarrow$ T misincorporation events at former 5mC loci. This approach was employed to estimate the age at death of a a 4000 year of old Paleo-Eskimo belonging to the Saqqaq culture [65].

5mC is quite chemically stable such that for most tissue sample preservation techniques for which DNA can be extracted from the sample DNAm can also be characterised [66]. DNA extracted with normal laboratory methods can be assayed for DNAm. Histone modification status, on the other hand, is less robust against environmental stresses on samples and often requires considerably more laborious sample preparation to characterise [67]. DNA methylation also offers the appearance of relative simplicity in comparison to the complex picture of the 'histone code,' the simple binary nature of DNA methylation state makes it easier to model and possibly to interpret. However, interaction between DNAm and histone modifications is well documented [68–70], this interplay between DNAm and histone modifications renders the interpretation of DNAm more complex, as its effects may be conditional on the chromatin environment.

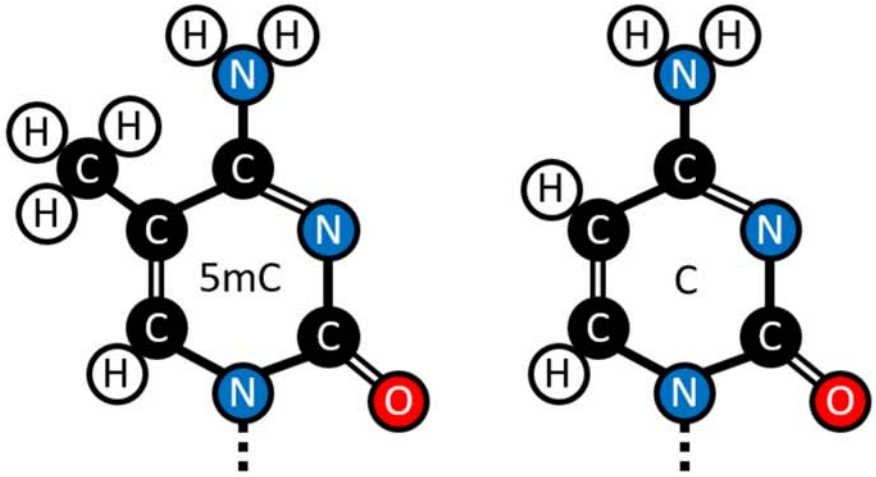

Figure 1.7: The structure of 5-methylcytosine contrasted with Cytosine. (Figure created by the Author.)

1.4.2 Distribution and Global Trends

There are 28,299,634 CpGs [71] in the hg19 assembly of the human genome [72]. Given that the GC content of the human genome is 42% the prior probability of getting a CpG dinucleotide is: $0.21 \times 0.21 = 0.0441$, 4.41%. CpGs represent $\sim 1.8\%$ of the dinucleotides in the human genome ($28,299,634 \div (3.23 \times 10^9 \div 2) \approx 0.0175$). Making them ~ 2.5 fold less frequent than would be expected *a priori*. Methylated cytosines are prone to deamination to thymines, resulting in mismatch lesions where a T is opposite a G. This can result in C to T mutations when DNA repair processes resolve this mismatch by changing the G to an A instead of correcting the T back to a C [73]. This increased mutagenic potential generally means that CpGs are selected against, accounting for at least some part of their under-representation in the genome. Methylated cytosines had a mutation rate in sperm of 18.5% compared to 5.44% for unmethylated CpGs [74], and C $\rightarrow$ T transitions account for ~33.1% of all SNPs [75]. The three major classes of repeat elements SINEs, LINEs and LTRs contain some 46% of all CpG sites, with a further 5% in other repetitive elements [71].

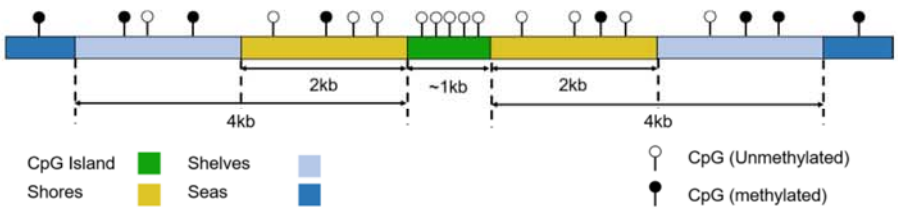

Figure 1.8: Schematic representation of CpG Islands, shores, shelves and seas. CpG density declines from shores to seas and methylation increases (CpG density and methylation proportion not to scale). (Figure created by the Author.)

CpGs are not uniformly distributed in the genome, they occur at higher frequency in some regions. "CpG islands" or CGIs are regions of high CpG density. They are defined somewhat arbitrarily, the definition used by UCSC is a sliding 500bp window [76]. There have existed other definitions [77]. The total number of CpGs in the UCSC repeat masked CGI annotation list of 28,691 CGIs is 1,990,729. Therefore, these CpGs comprise $\sim 7.0\%$ of the total number of CpGs the genome. The mean percentage of the sequence of these CGIs that is comprised of CpG dinucleotides is $\sim 18.5\%$ and their mean length is 761bp. CGIs overlap the promoters of $\sim 70\%$ of genes [78], this, however, leaves roughly half of all CGIs as "orphans" not associated with a known transcription site. These orphan CGIs frequently constitute cell-type specific enhancers [79] and alternate promoters [80,81]. They are found in both intergenic and intragenic regions. CpGs are distributed quite sparsely through the genome occurring at low density in "seas" and at increasing density in CGIs and their flanking regions, see figure 1.8. It is a characteristic of CGIs that they are generally unmethylated.

Early work showed 70-80% of CpGs are constitutively methylated [82,83] Stadler et al. produced a more detailed picture of the distribution of CpG methylation in mouse embryonic stem cells (ESCs) [84] see figures 1.9 & 1.10. Stadler et al. categorised regions of the genome into fully, low and un methylated regions (FMRs, LMRs, UMRs) in mouse embryonic stem cells using a hidden markov model. LMRs were evolutionarily conserved and highly enriched for the chromatin signatures of distal regulatory regions such as enhancers (High H3K4me1 relative to H3K4me3, p300, H3K27ac over H3K27me3) [84]. Irizarry et al. found that the 2kb regions flanking CpG islands which they termed "CpG island shores" exhibited greater tissue-specific differential methylation than the islands themselves [85]. This nomenclature has subsequently been expanded further with "CpG island shelves" which are 2kb - 4kb from the CGIs, and "seas" referring to the rest of the genome, see figure 1.8. Ziller et al. identified ~5.6 million CpGs, ~21.8% of autosomal CpGs, which are dynamically regulated across diverse cell-types, these clustered into ~716,000 differentially methylated regions. More that three quarters of these regions where under 1kb in size and located away from transcription start sites and >70% of which had methylation levels >75% [86]. Data from three different approaches taken by Irizarry et al. [85], Stadler et al. [84], and Ziller et al. [86] indicate that the most dynamic DNA methylation changes are occurring at the level of small distal regulatory features. A small fraction of CpGs, on the order of 2-6% exhibiting constitutively unmethylated states and on the order of 70-90% constitutively methylated leaving 4-28% of sites subject to dynamic regulation.

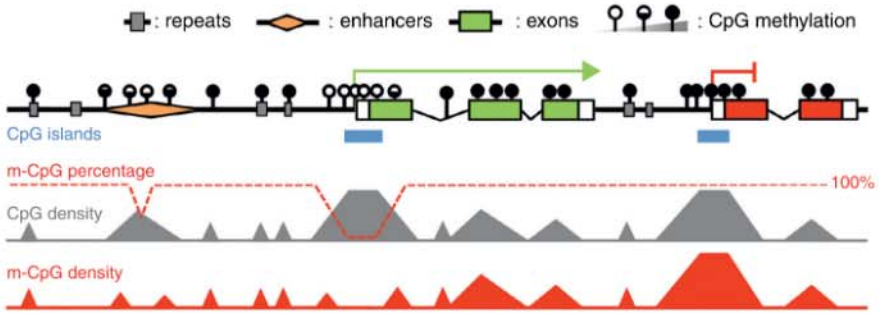

Figure 1.9: Stylised representation of the relationship between CpG Density, CpG methylation and functional DNA elements. (Figure reproduced from Baubec & Schubeler [87] figure 1)

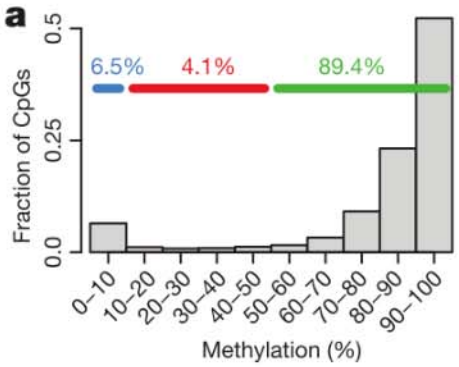

Figure 1.10: The fraction of CpGs with a given methylation level. Data from Whole-Genome Bisulfite Sequencing (WGBS) of Mouse embryonic stem cells. Colours denote: unmethylated regions (blue), low-methylated regions (red), and fully methylated regions (green). These groups are three segments produced by a hidden markov model. low-methylated regions with 10-50% methylation are evolutionarily conserved distal regulatory elements with dynamic cell-type specific regulation unlike the consistently unmethylated regions at CGIs. (Figure reproduced from Stadler et al. [84] figure 1a)

The regulatory role played by DNA methylation varies with context and the scale at which it is examined. For example methylation at a single CpG in a transcription factor binding site can determine whether or not the factor can bind at that site. There are methylation-sensitive transcription factors which can have their affinity for DNA either increased or decreased by methylation [88,89]. In addition, larger-scale changes in DNAm can through interactions with chromatin modifiers alter DNA compaction and more general accessibility of DNA regions for binding, and/or impacting on their topological organisation. DNAm also varies with nucleosome occupancy with lower methylation in linker sequences than on nucleosome-associated DNA [90,91]. The oxidation products of 5mC can also affect chromatin organisation, 5-Formylcytosine can impact nucleosome positioning through covalently binding to histones [92]. This raises the

question if some DNA methylation occurs specifically to act as an intermediate step for regulatory mechanisms which utilise its oxidation products.

1.4.3 Pathways of DNA methylation and demethylation

CpG methylation is produced and maintained by DNA methyltransferase (DNMT) enzymes. All the DNMT enzymes use S-adenosylmethionine as the source of the methyl donor group. DNMTs form a covalent intermediate between a conserved cysteine residue and the target base, through a nucleophilic attack on the C6 position in the cytosine ring [93]. This is followed by the transfer of the S-adenosylmethionine methyl group to C5, and deprotonation of the C5 to reform the double bond between C5 and C6; which is mediated by a base provided by the enzyme (Figure 1.11).

DNMT1 is associated with the replication machinery [94] and reproduces the methylation state of the parent strand on the daughter strand during replication. The largest of the DNMT family DNMT1 contains a 'replication foci targeting sequence' (RFTS) domain required for its targeting to replication forks. DNMT1 specialises in recognising hemimethylated DNA and methylating the unmethylated C in a palindromic CpG dinucleotide site, the UHRF1 protein assists in the recognition of these sequences [95]. DNMT3a and DNMT3b are responsible for *de novo* DNA methylation along with DNMT3L a catalytically inactive, but DNA-binding subunit [96]. The location of *de novo* methylation by the DNMT3s is influenced by a number of factors including the Chromatin state and other DNA binding factors [93].

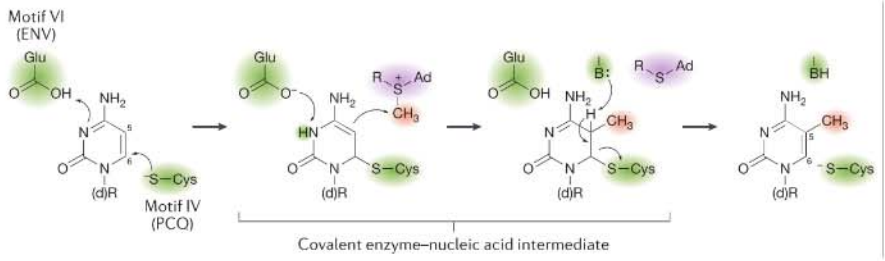

Figure 1.11: Motif VI (ENV/PCQ) refers to the conserved catalytic motif in DNMTs. 'B:' Represents the enzyme supplied base (Figure reproduced from the review of DNMT biology by Lyko [93] (fig. 1b))

DNMT2 is unlike the other members of the DNMT family in that it targets an RNA substrate. DNMT2 is a tRNA methyltransferase which methylates a specific subset of tRNA genes, mostly Asp isoacceptors, at a site adjacent to the anticodon which protects them from endonucleolytic cleavage under stress conditions. Unprotected tRNAs produce fragments which compete with small interfering RNAs interfering with their signalling pathways [97].

The mechanisms of the demethylation of DNA were more recently characterised than those

which govern its methylation [98,48]. They are more complex and less well understood. TET2 mutations are quite common in cancer, ranking 65th across 12 major cancer types [99] and more common in diverse myeloid malignancies, where loss of its catalytic activity favours tumourigenesis [100]. A distinction is drawn between active and passive demethylation, in passive demethylation, 5mC bases are diluted out in the process of DNA replication. In the leading model of active demethylation, they are oxidised one or more times by an enzyme from the TET (Ten-Eleven-translocase) family. They are then either passively removed by DNA replication or actively removed by a DNA glycosylase (thymine DNA glycosylase TGD) to create an apurinic site which is restored to a C by the Base Excision Repair (BER) pathway. This cycle of cytosine methylation and demethylation is illustrated in Figure 1.12 [101].

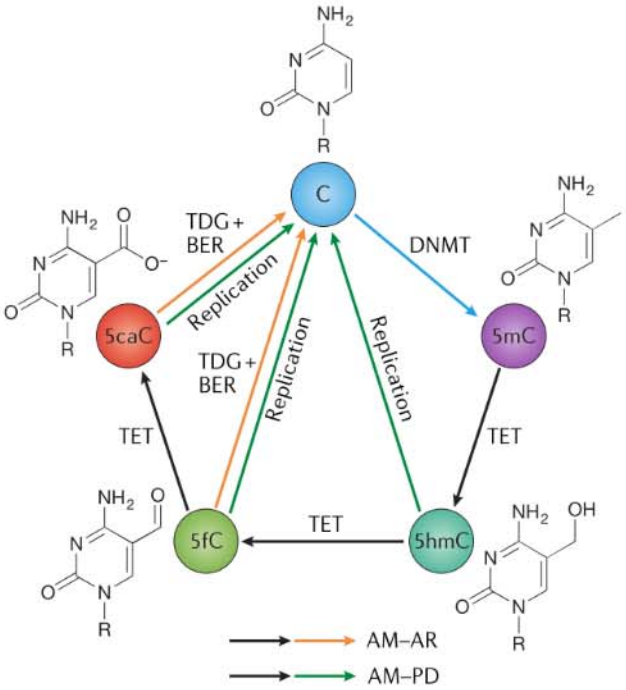

Figure 1.12: BER = Bases Excision Repair; AM = Active Modification; AR = Active Removal, PD = Passive Dilution (Figure reproduced from the review of TET mediated active demethylation by Wu & Zhang [101] (fig. 1a))

1.4.4 Maintenance and Fidelity

DNA methylation is highly dependent on the underlying DNA sequence. Sequence features of particular importance to determining DNA methylation status are; transcription factor and other DNA binding protein recognition motifs, and CpG density [102,86]. Alterations in underlying DNA sequence such as SNPs and copy number variants (CNVs) can have a significant impact on methylation level and the susceptibility of the methylation level to change. In addition, CNVs can result in dosage effects on measures of DNAm, causing regions to appear, respectively,

substantially more or less methylated when fewer or greater copies than expected are present [103–108]. Despite the strong influence of sequence on methylation, global CpG methylation exhibits change over developmental time as illustrated in figure 1.13 and exhibits tissue-specific changes in distribution and amount [86].

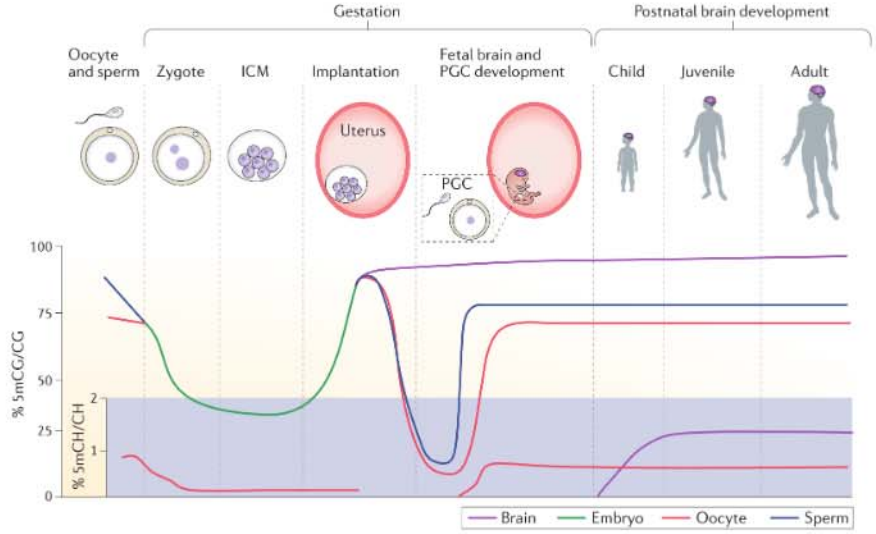

Figure 1.13: 5mC levels over developmental time. 5mCH or 5-methylcytosine-(A, T or C) levels are represented on a separate (purple) Axis from the main 5mCG axis. PGC = Primordial Germ Cell. (Figure reproduced from [109] figure 3)

In mitosis, DNA methylation is inherited by daughter cells with an error rate on the order of 1×10^{-3} per site per generation [110], several orders of magnitude less than that of DNA replication (error rate of 1×10^{-7} - 1×10^{-8} [111]). DNMT1 methylates hemimethylated DNA at CpG sites in S phase copying the DNA methylation state of the parent strand onto the daughter strand. DNMT1 provides the catalytic activity and is precisely targeted to hemimethylated DNA in conjunction with the UHRF1 (ubiquitin-like, containing PHD and RING finger domains 1) protein which specifically binds hemimethylated CpG sites via its SRA (SET and ring-associated) domain using a flip-out mechanism to recognise 5mC [112,95,113,114].

The fidelity of DNA methylation copying can be assayed by Hairpin-Bisulfite PCR (Polymerase Chain Reaction) [115]. Laird et al. looked at two alleles of a portion of the CpG island from the human *FMR1* gene in uncultured lymphocytes, one hypermethylated and one hypomethylated. In the hypermethylated allele, they found that 96% of sites methylated in the parent strand remained methylated in the daughter strand and 86% of unmethylated sites remained unmethylated. By contrast in the hypomethylated allele, there were no methylated sites to be retained and >99% of unmethylated sites remained so following replication. Laird et al.'s work suggests a higher degree of overall methylation fidelity for hypomethylated DNA and a propensity for unmethylated sites in hypermethylated DNA to become methylated.

Change in DNAm levels over time (divisions) can be modelled using the differential equations [116], which predict that a fully methylated site and a fully unmethylated site will converge on an equilibrium level. This level is determined by the probability of maintenance of the methylation state and of *de novo* methylation for a given locus. This stochastic model of DNAm is in agreement with experimental findings [115,117], Figure 1.14.

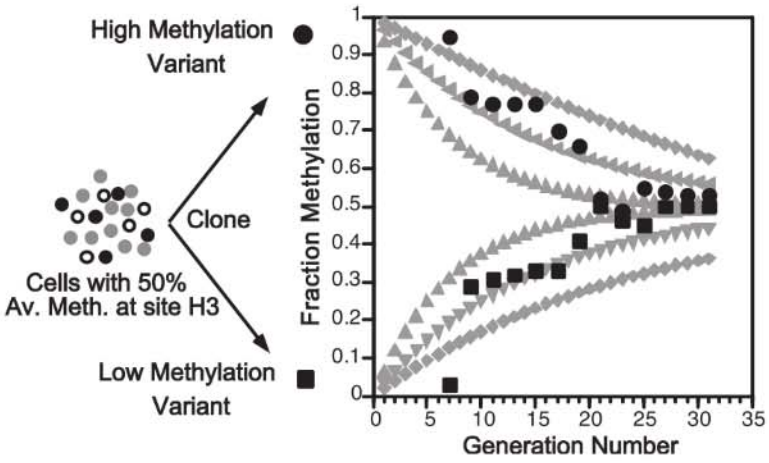

Figure 1.14: E_m = probability of methylation maintenance, E_m = probability of *de novo* methylation. M & U = the number of methylated and unmethylated molecules at specific CpG sites respectively. Modelling (**Grey points**) E_m, E_d values of 0.90,0.10; 0.95,0.05; 0.97,0.03 (outermost to innermost, two curves for each E_m, E_d set, one starting at $M = 1, U = 0$ and one a $M = 0, U = 1$. Experimental data (**Black points**) from the *HpaII* locus (site H3) in 17 clones of mouse cell line BML-2 which has a known methylation level of 50%. (Reproduced from Riggs & Xiong [117])

Jenkinson et al. [118] used an information-theoretic approach, modelling DNAm as a binary communications channel using a 1-dimensional Ising model from the field of statistical physics. This permitted them to examine properties of DNAm not accessible to conventional means of analysis which typically capture the mean methylation level and perhaps the variability for a given locus. When considering methylation fidelity using this lens the maintenance of a given methylation state can be seen as an information processing task which requires the consumption of free energy in order to reduce the probability of error in the transmission of that methylation state. Thus Relative Dissipated Energy (RDE) can serve as a measure of the work expended by a cell at a given locus in order to preserve the current methylation state of that locus. They found that the Transcription Start Sites (TSS) exhibited high levels of RDE and high information capacity as well as low levels of entropy. These values trended in the opposite direction as they moved outwards to CGIs, shores and seas. Jenkinson et al. [118] also found a global increase in entropy with age, but not with cell passage in culture suggesting an increase in entropy is associated with epigenetic age independent of mitotic age. This observation is in agreement with

the finding that epigenetic state becomes more stochastic and diverges with age [119,120], indeed genes whose methylation were most divergent with age were enriched for ageing associations. They also observed a loss of entropic sensitivity with age. Entropic sensitivity is an indicator of how plastic DNAm state is to extrinsic effects. Jenkinson et al. noted that there is a general loss of phenotypic plasticity with age, but cited no specific instance of loss of DNAm plasticity to environmental effects with age. Hahn et al. [121] provided a possible example of this effect when they reported that dietary restriction caused fewer differences in the methylation state of older than younger mice.

Zhao et al [122] applied Hairpin-Bisulfite PCR genome-wide in mouse embryonic stem cells (ESCs). They also found high degrees of methylation fidelity in hypomethylated regions such as CGIs and Promoters, as well as a high degree of fidelity in sites bound by transcription factors. This is in agreement with the findings of Jenkinson et al. [118] who noted that entropy (methylation stochasticity) was lower and more variable in CGIs and TSS (transcription start sites).

Methylation inheritance fidelity is reduced in cancer [110] and increases with differentiation [122]. The RDE (relative dissipated energy) at CGIs and TSSs is higher in differentiated tissues such as the brain, implying low entropy, and lower in embryonic stem cells, implying greater entropy [118] (see Figure 1.15). In addition, the correlation between CpG sites increases in cancer [118], suggesting reduced higher-level regulatory control and tendency to fall back on lower level feedback mechanisms leading to a return to the baseline equilibrium described above as the cell is expending less energy to maintain an out of equilibrium methylation state.

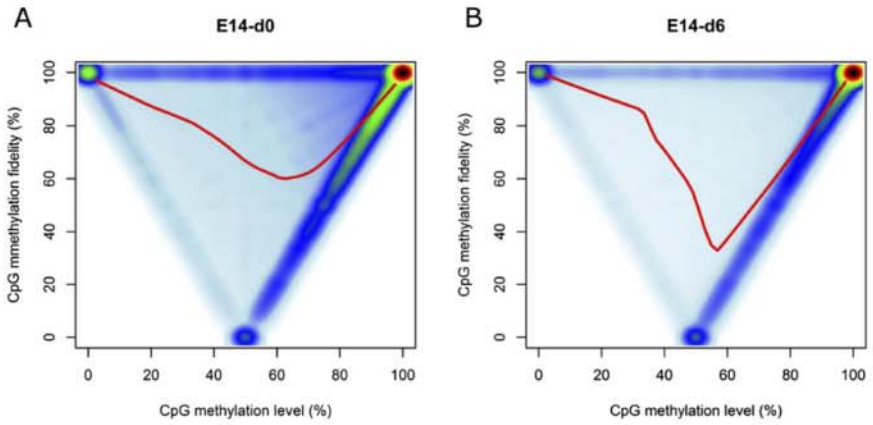

Figure 1.15: Bimodal Distribution of DNA methylation level and fidelity. Methylation fidelity exhibits a bimodal distribution with most and least methylated regions exhibiting the highest fidelity. Additionally, fidelity for methylation levels of 10-50% is considerably better than fidelity for levels of 50-90%. These data are from the mouse cell line ES-E14TG2a which is capable of self-renewal when cultured with Leukaemia Inhibitory Factor (LIF) and spontaneous differentiation upon removal of LIF (denoted as E14-d0 at day 0 and E14-d6 at day 6 after the withdrawal of LIF, respectively). Fidelity represents the percentage of symmetrically methylated or unmethylated CpG dyads for a given position as determined by Hairpin-Bisulfite PCR. (Reproduced from Zhao et al. 2014 [122] (figure 3 a and b).)

Zhou et al. linked the substantial loss of DNA methylation in proliferative cancer cells, even counterintuitively those cancers with high DNMT1/UHRF1, DNMT3A/B levels and physiological levels of TETs, to a systematic mechanism of DNA methylation loss. Partially methylated domains (PMDs) are characterised by low CpG density, larger numbers of lone CpGs flanked on both sides by A/T nucleotides, and low levels of H3K36me3 as well as being lamina-associated late replicating domains. PMDs are prone to becoming hypomethylated in a manner which progressively increases with mitotic division. Copying of the parent strand methylation state by DNMT1/UHRF1 takes time and may not be fully completed with each round of replication leading to progressive DNA methylation loss in these regions, especially if proliferation is rapid. Some late replicating regions have greater methylation fidelity through the action of DNMT3A/B at sites with the H3K36me3 mark [123].

1.4.5 DNA Methylation Assays

There are numerous DNA methylation assays. Some are designed to assay global DNAm levels producing a single global measurement of the amount of 5mC present in a sample, others assay DNAm in a manner traceable to a specific genomic locus. These methods can be further subdivided into targeted methods where the loci to be examined are known in advance, and those where

the loci covered are quasi-randomly sampled. Both the targeted and untargeted methods vary in their granularity, from individual CpGs to large regions of the genome, on the order of megabases. These assays also vary in their coverage of the genome from locus-specific approaches looking at small numbers of individual loci to whole-genome methods providing methylation information on all ~28 million CpG sites. Between these two extremes are 'epigenome-wide' methods which are focused with varying degrees of specificity on genomic regions of interest.

These DNA methylation assays generally exploit one of four following principles:

1. **Methylation sensitive restriction digestion** A variety of methods exploit enzymes which differentially cut methylated and unmethylated DNA. An example of such enzymes is: HpaII and MspI which are isoschizomers for the sequence: 5'-C^CGG-3' but only HpaII is methylation-sensitive and unable to cut the sequence if the internal cytosine is methylated. The methods which use such enzymes include: Restriction landmark genomic scanning, which gives a roughly megabase resolution indicators of DNAm levels by 2D electrophoresis [124]. MRE-seq (Methylation sensitive restriction enzyme sequencing), which makes use of methylation-sensitive restriction digests to enrich for unmethylated DNA [81]. This is then sequenced to provide the genomic location of this unmethylated signal. Reduced Representation Bisulfite Sequencing (RRBS), which uses methylation-sensitive restriction digestion to enrich for unmethylated sequences which are then subject to bisulfite sequencing, (see point 3 below).

2. **DNA methylation-sensitive binding of DNA by antibodies or other proteins selectively which bind methylated DNA** Anti-5mC antibodies were first used to isolate methylated DNA in 1985 and were subsequently paired with array-based technologies to assay DNAm levels at specific loci in 2005 [125,126]. This was followed by MeDIP-seq (methylated DNA immunoprecipitation and sequencing) [127,128], which enriches methylated DNA that is then sequenced and the number of reads mapping to a locus is indicative of the relative methylation level. MBD-seq uses a recombinant Methyl Binding Domain (MBD) protein to enrich for methylated double-stranded DNA, prior to sequencing [129]. Inferring the absolute methylation levels form the number of reads mapping to a locus enriched by one of these pull-down methods is complicated by variation in CpG density in the genome and the fact that CpG dense regions tend to have low methylation levels, and thus tend not to be pulled down by antibodies binding 5mC or MBD proteins. The best results for estimating absolute methylation levels with these methods come from pairing them with a complementary method to enrich for unmethylated sites such as MRE-seq (described above) [130]. Though relative methylation levels remain effective in identifying differentially methylated regions when using these methods alone [131,81].

3. **Sodium bisulfite conversion of unmethylated cytosines to uracil** Conversion of cytosine to uracil changes the base complementary to this site from G to A [132]. The

basic chemistry of the conversion process is illustrated in Figure 1.16. This conversion can be detected with a variety of technologies such as: Sanger sequencing, The Illumina bead chip methylation arrays; WGBS; and Targeted bisulfite PCR or sequencing to examine a small number of selected loci. It can also be performed on a larger number of loci with microfluidic multiplexing such as that provided by the Fluidigm access array [133]. It is worth noting that bisulfite conversion cannot differentiate between methylated and hydroxymethylated cytosines. Because 5hmC represents a small fraction of modified bases compared to 5mC many analyses have made the working assumption that unconverted bases are methylated, but it is beginning to be recognised that distinguishing between the two may be biologically important especially in tissues where 5hmC is more prevalent such as neurons. A variant of bisulfite conversion, oxidative bisulfite conversion, exists which can permit 5hmC to be distinguished from 5mC [134]. Additionally, New England Biolabs has recently developed an enzymatic alternative to chemical bisulfite conversion [135]

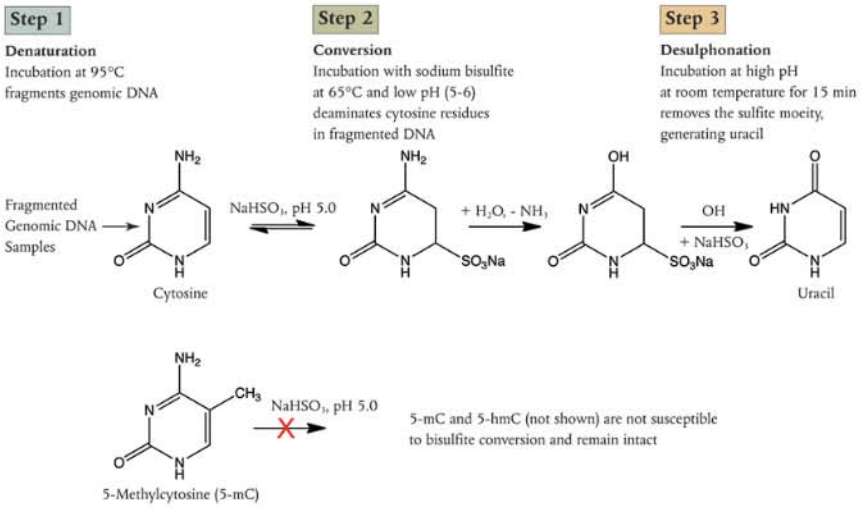

Figure 1.16: Sodium Bisulfite conversion of C to U, 5mC and 5hmC are unaffected. (Image Credit New England Biolabs).

4. **Physical differences in the methylated base** These are exploited by the not yet widely used '3rd generation' sequencing technologies such as nanopore and single-molecule real-time sequencing. These methods respectively exploit the effects of modified bases on ion flow through the nanopore and impact on incorporation time of new bases whilst in the polymerase [136,137].

1.5 DNA Methylation and Ageing

Returning now to the relationship between DNA methylation and ageing we will review the major discoveries in that field, the roles thought to be played by DNAm in ageing and the utility of DNAm clocks as biomarkers of ageing.

Early work on DNA methylation and ageing used biochemical, chromatographic and radiolabeling techniques to assay global changes in the amount of 5mC. These studies found differences in 5mC composition with cell type [82] but did not see changes with age. Other work around that time however found differences in 5mC levels with the age of cells in culture [138,139]. Wilson et al. noted that DNA methylation decreased across several tens of generations of cell lines in culture, but not in immortal cell lines. Immortal cell lines had lower absolute levels of 5mC to start with but remained constant over time. Wilson et al. also cite earlier work by Romanov at al. 1981 [140] and Berdyshev el al. 1967 [141] which documented decreases in 5mC with age in cows and salmon, respectively.

Later work by Wilson et al. 1987 [142] in mice *P. leucopus* and *M. musculus* found decreases in 5mC with age and that the rate of decline was less in the longer-lived *P. leucopus* than in *M. musculus*. Interestingly, a recent study using Reduced Representation Bisulfite Sequencing (RRBS) by Cole et al. found no global differences between young and old mice in short and long-lived strains, however, their other observations would seem to corroborate these initial trends. For the assayed sites long-lived mice exhibited 10x more hypermethylation than wild-type (WT), and WT mice had 3x more Differentially Methylated Regions significantly associated with age (DMRs/aDMRs) than did long-lived mice. Notably, the WT and long-lived mice shared many of the same aDMRs which differed in their degree of methylation as opposed to affecting different sites in the genome [143]. The lack of apparent global changes may be due to the biases of the reduced representation bisulfite sequencing (RRBS) method used. RRBS uses a restriction enzyme-based approach to enrich for regions with high GC content such as CGIs which tend to have low levels of methylation, and consequently may not be sensitive to loss of methylation in generally hypermethylated regions which could contribute to a global trend [144].

Wilson et al. also noted that the mitotic index of tissues did not relate to the loss of DNA methylation with age in tension with their earlier observations *in vitro*. The persistent loss of 5mC over time and the dramatic changes in methylation seen in cancer cells led Wilson et al. to suggest that dysregulation of DNA methylation may have a substantial role to play in the age-dependency of cancer risk and ageing more generally [145]. It has been suggested that "Epimutations" may be able to substitute for mutations in the multi-hit model of carcinogenesis [146]. For example; hypermethylation of the *BRCA1* promoter [147], Or, the development of Wilms' tumour due to Beckwith–Wiedemann syndrome, a disorder arising from loss of imprinting of the gene encoding insulin-like growth factor 2 (IGF-2) leading to a double dose of IGF-2 protein

[148]. Furthermore, the loss of methylation and increased entropy of closed chromatin regions may lead to increased susceptibility to the types of structural mutations commonly found in cancer cells. This could result from the exposure of homologous sequences presenting increased opportunities for erroneous breakpoint repair or recombination between similar sequences in different parts of the genome [145,118].

Methods which allowed the examination of changes in DNA methylation at known loci in the genome permitted a more nuanced picture of changes in DNA methylation over time to develop. Fraga et al. [119] introduced the concept of "epigenetic drift" being the divergence of DNA methylation and other epigenetic modification patterns with time. Fraga et al. looked at the divergence in DNA methylation along with global Histone H3 and H4 acetylation patterns between monozygotic (MZ) twins over a wide range of ages. They observed that older twins had greater epigenetic differences with time, in DNA methylation as well as H3, H4 Acetylation. Fraga et al. also noted that the divergence in epigenetic state was greater in twins who had lived longer apart and had different medical histories. Figure 1.17 is a useful visual encapsulation of Fraga et al.'s DNA methylation results. Slieker et al. [120] identified 6366 CpGs whose methylation variability increased with age (aVMPs) using the Illumina 450k array platform on whole-blood from 3295 individuals. Variably upregulated aVMPs were associated with the expression of DNA repair and apoptosis genes and variably downregulated aVMPs with cellular metabolism. Furthermore Both the increase in variability with age and the increasing divergence of twins support a narrative of epigenetic dysregulation and increasing entropy with age.

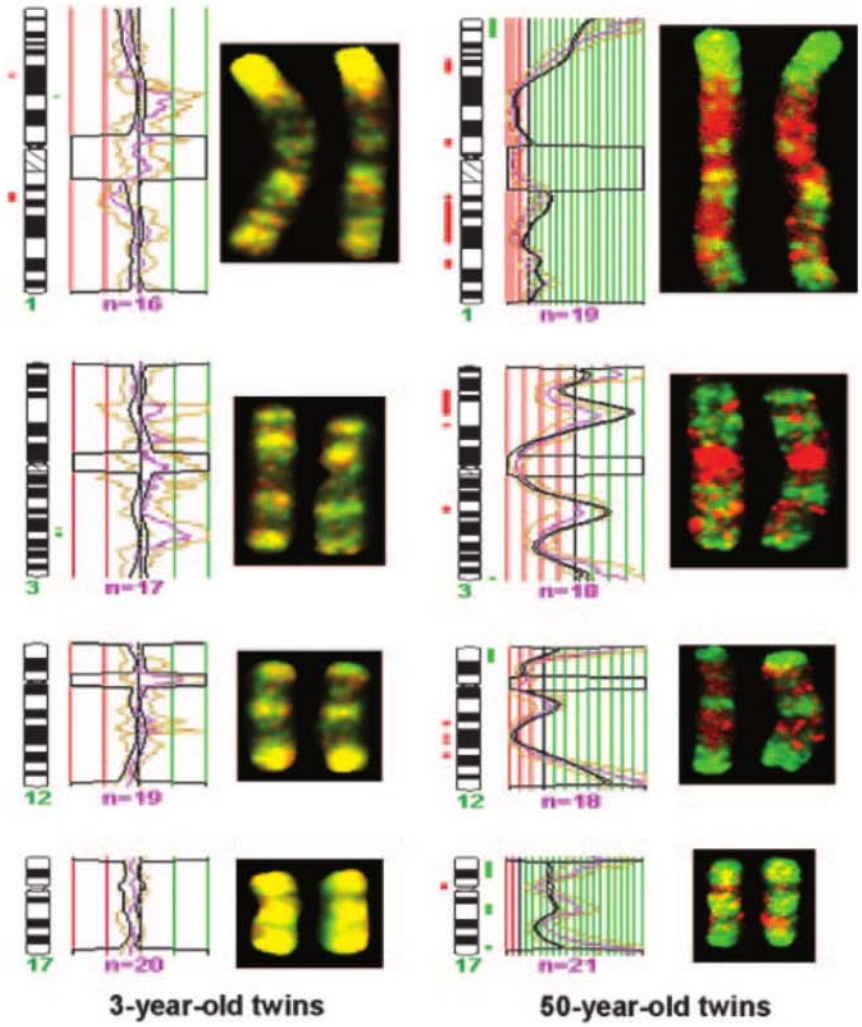

Figure 1.17: Mapping Chromosomal Regions of Differential DNA Methylation. Examples from Chromosomes 1, 3, 12 & 17 are shown for a 3 and 50-year-old twin pair. Methylation status is shown by competitive hybridization of AIMS (amplification of inter-methylated sites) products to metaphase chromosomes. Green and Red signals indicate hypermethylation and hypomethylation events between twins, Yellow indicates little difference. Red and Green blocks adjacent to ideograms indicate areas of significant DNA methylation change. (Reproduced from Fraga et al. [119] (figure 3).)

1.5.1 Age-Related Differential Methylation and Epigenetic Clocks

With the advent of array-based techniques which permitted the measurement of the methylation status of specific CpGs placed strategically throughout the genome, researchers were able to undertake a more fine-grained analysis of the changes in DNA methylation with age. This has allowed the prediction of chronological age from a metric of "DNA methylation age" (DNAm age), and interpretation of the differences in predicted and chronological age in terms of the

pathophysiology of ageing.

The Illumina Golden gate array [149] with 1505 probes targeted to cancer-related genes was the first of these. Bjornsson et al. observed DNAm changes with age as well as noting substantial intra-individual differences. In contrast earlier work by Eckhardt et al. [150] and Ehrlich et al. [82] found no age-related changes using averages across individuals. Bjornsson et al. [151] also found that the intra-individual differences in differential methylation with ageing were highly heritable by use of familial clustering. Work by Boks et al. [103] also using the golden gate array corroborated Bjornsson et al.'s findings in monozygotic twins.

The Bjornsson and Boks studies used the golden gate array on peripheral blood samples, whereas Christensen et al. [152] used the golden gate array to look for age-related changes in DNAm in several tissue types. CpG sites that are differentially methylated with age when searching across tissues were frequently specific to a small number of tissue types. In addition, CpGs in CGIs tended to be hypermethylated with age and *vice versa*. Christensen et al. also found that the methylation profiles for different tissues were highly predictive of tissue type.

Rakyan et al. [153] looked at age-related changes in DNAm using the Illumina 27k array [154] with whole blood samples in a discovery set. They replicated their findings in sorted cell fractions to see if they could capture changes in DNAm that could be attributed to changes in blood cell-type composition over time, which they did not find. Teschendorff et al. [155] found that promoters of targets of the polycomb group proteins (PCGTs) were more likely than non-PCGTs to become methylated with age [156]. Repression of PCGTs is required for Stem cells to differentiate. PCGTs whose methylation status was associated with age were associated with pre-neoplastic conditions in a large cohort using the 27k array with blood and epithelial cell samples.

Whilst Bell et al. [157] observed a correlation between a major component of variation in DNA methylation in blood and age, Bocklandt et al. [158] also using the 27k array but with saliva samples created the first "epigenetic clock" used to predict the chronological age of donors based on the DNAm landscape of their cells, the mean error was 5.2 years. Bocklandt et al. were able to achieve similar predictive accuracy with as few as 3 CpG sites in their model. Koch et al. [159] used publicly available 27k array datasets from a variety of different tissues to train a model using a different statistical method but only achieved an error of 11 years. They did, however, identify a number of CpGs also identified by Bocklandt et al. Bell et al. [160] used 27k array data to look for correlations between differential methylation and age-related phenotypes as well as chronological age. They found that few age-related phenotypes were correlated with differential methylation, only 5 CpG sites were identified for the 16 age-related phenotypes examined. Whereas 490 significant CpGs were associated with chronological age. Many of the CpGs they identified persisted across tissue types and replicated in a second cohort, several had been identified previously by Rakyan et al. [153] and Bocklandt et al. [158].

Heyn et al. [161] examined the DNA methylation status of a newborn and a centenarian in CD4+ T cells using whole-genome bisulfite sequencing (WGBS), along with a group of newborns and nonagenarians using the Illumina 450k array [162]. Heyn et al. observed a global decrease in methylation from newborn to centenarian, as well as an intermediate level of methylation on a sample of intermediate age (see figure 1.18). This observation was replicated in 450k data. In addition, Heyn et al. noted that adjacent CpGs normally exhibit a substantial degree of correlation in methylation status and that they were less well correlated with their neighbours with increasing age. Interestingly Jenkinson et al. [118] found that correlation among nearby CpGs increased in cancer tissues. Modelling work done by Affinito et al. [163] agrees with this correlation between neighbouring CpGs under physiological conditions, which is particularly pronounced in CpG dense regions like CGIs. The greater physical proximity of CpGs in CpG dense regions means that greater correlation is expected. Due to the fact that these CpGs are more likely to be affected by the same proteins and regulatory features than CpGs with greater distance separating them [164]. Unmethylated CGIs with H3K4me3 histone modifications can act as functional promoter units, this unmethylated state is needed for their interaction with CXXC domain containing proteins like Cfp1 [165]. Garagnani et al. used 450k array data from a cohort of 64 subjects aged 9-83 to identify those CpGs most well correlated with age as had been done with previous array technologies [166].

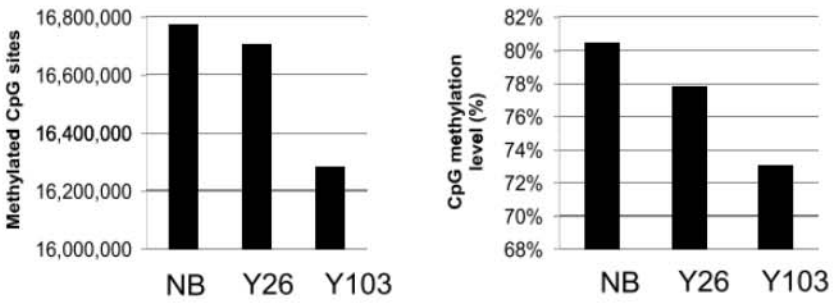

Figure 1.18: Global Hypomethylation with Age. NB = Newborn, Y26 = 26 year old, Y103 = 103 year old. (Reproduced from Heyn et al. [161] figure 1 b.)

In January of 2013 Hannum at al. [167] published a study using the Illumina 450k array and whole blood samples for 656 participants aged 19-101. Their optimised model was able to predict age with an error of 3.9 years. In December of the same year, Horvath [168] published a study using multiple datasets, including Hannum's data, some produced on the 27k and some the 450k array. Both used 'elastic net' penalised multivariate regression models to identify CpGs which collectively provide the best predictor of DNAm age rather than the CpGs which are individually most highly correlated with age [169].

Florath et al. [170] identified some additional age-associated CpGs from two large cohorts

totalling N = 898 and an N = 67 8yr follow-up longitudinal cohort. More than 3/4 of the CpG sites they identified began as hypomethylated and increased in methylation with age, a proportion likely skewed by the large number of 450k probes in CGIs and other typically low methylation regions. Bacalini et al. [171] performed a meta-analysis of existing DNAm age datasets and employed a "region-centric" approach to try to identify loci larger than single CpGs which they anticipate will be more biologically meaningful than lone CpGs, despite being limited to just the CpGs for which there are array probes. They found that their approach increased the number of common features identified using the Hannum et al. [167] and Heyn et al. [161] datasets. Zaghlool et al. [172] performed 450k array DNAm age study in Qatari population they identified 12/88, 23/490 and 102/162 of the CpGs found by Bocklandt et al. [158], Bell et al. [160] and Florath et al. [170] respectively. Zaghlool et al. [172] also found that Horvath's age predictor [168] had an error of 3.7 years in their dataset. This is in agreement with existing findings and indicates that ethnicity has no major effects on the DNAm age signature. Benton et al. [173] looked at changes in DNA methylation with age in a genetically isolated population on Norfolk Island with similar results to previous studies but identifying some novel age-associated CpGs.

Most DNAm relationships with age noted in previous studies have been linear, Johnson et al. [174] used the 450k array in peripheral blood samples. They identified 21 CpGs whose DNAm changes at a rate that changes with age from an initial pool of 27,723 CpGs which were differentially methylated with age. Two sites exhibited an increasing rate of increase in DNAm with age, and 18 sites a decreasing rate of increase. However, these effects could be impacted by changes in blood cell-type composition with age.

1.5.2 Genetic Influences on DNA Methylation

Epigenetic variation falls on a continuum of genetic influence that can be summarised by three categories:

1. **pure** - DNA sequence has no predictive value for epigenetic state.
2. **facilitated** - DNA sequence biases the epigenetic state enabling epigenetic variation.
3. **obligatory** - DNA sequence permits exact prediction of epigenetic state.

Regional methylation state is strongly influenced by genotype, by single nucleotide polymorphisms (SNPs) [175] and by structural variants [176,177]. An example of an obligatory effect on methylation is a point mutation at a CpG site, a C to T transition precludes methylation at that site in future. The effect of structural variants on methylation can be hard to determine as changes in sequence dosage often lead to measurement artefacts. Efforts have been made to correct for the influence of genetic factors in EWAS, when searching for purely epigenetic effects

but the potential interaction of the somatic mutations known to accumulate with age (Figure 1.19) and the changes in DNAm with age remain largely unexplored.

Increases in frequency of particular mutations can also be driven by clonal expansion of cellular sub-populations, particularly blood, and possibly in other tissues with a high level of turnover. The differentiation of blood cells from heamatopoetic stem cells is less linear and hierarchical than initially modelled. The hematopoetic stem cell compartment is comprised of a pool of cells with characteristics of multiple lineages and are to a greater or lesser extent primed for a particular branch of differentiation [178]. The somatic mutations present in these hematopoetic stem cells [179], provide the variation necessary for selection processes to act, and the changes in variant allele frequency in blood cells indicates that there are positive selection effects driving the increase in frequency of certain alleles [180]. These selection effects favour more highly proloferative cells and may lead to the exhaustion of stem cells able to produce particular lineages and, imbalanced production of different cell types in undesirable proportions [181]. This can be a pre-cursor to hematalogic malignancies [179].

This is of particular relevance to DNAm as the profile of the types of mutation which accumulate with age (Figure 1.20) distinctly favours C to T transition mutations which can disrupt CpG dinucleotides [182]. CpG dinucleaotides are particularly prone to these mutations through the deamination of 5mC to T.

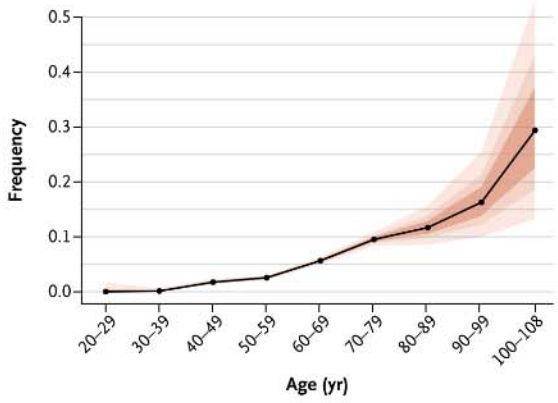

Figure 1.19: Somatic mutations increase with Age. (Reproduced from Jaiswal et al. [182], figure 1)

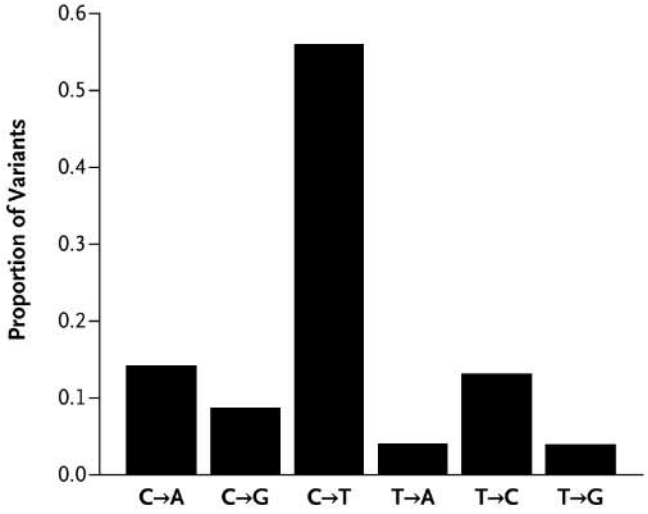

Figure 1.20: C to T transitions are the most common substitutions that occur with age. (Reproduced from Jaiswal et al. [182], figure2 c)

1.5.3 DNAm Age as a Biomarker of Ageing

We have seen that DNAm can be a fairly accurate predictor of age, however, there is more to being a good biomarker of ageing than predictive accuracy. Weidner et al. [183] conducted a study with the specific intent of identifying CpG sites that would serve as the best biomarkers of ageing. Below is a list of specific criteria for a high-quality biomarker of ageing laid out by Johnson et al. [184].

1. It must predict the rate of ageing. In other words, it would tell exactly where a person is in their total lifespan. It must be a better predictor of lifespan than chronological age.
2. It must monitor a basic process that underlies the ageing process, not the effects of disease.
3. It must be able to be tested repeatedly without harming the person. For example, a blood test or an imaging technique.
4. It must be something that works in humans and laboratory animals, such as mice. This is so that it can be tested in lab animals before being validated in humans.

We have seen that DNA methylation clocks can be good predictors of chronological age, we will revisit the question of whether or not they are superior to chronological age in their ability to predict lifespan below. Bell et al. [160] found DNAm did not generally correlate well with ageing phenotypes but did correlate well with chronological age and Teschendorff et al. [155] found that their age-associated DNAm signature remained fairly constant across several disease states including ovarian cancer and type 1 diabetes. Suggesting that DNAm age prediction does indeed: "monitor a basic process that underlies the ageing process, not the effects of disease," meeting criterion number 2. Several of the DNAm clocks discussed so far have been based

on whole peripheral blood samples meeting criterion number 3. DNAm age clocks have been shown to work in Chimpanzees [168], Dogs/Wolves [185], Mice [186], Naked Mole Rats [187], Rhesus Macaques, Humpback Whales [188] and are likely to work in other mammalian model organisms. Furthermore, a cross-species clock based on conserved ribosomal RNA genes has been developed [189]. It should be noted that Horvath's Human clock uses array data and that the Dog/Wolf and Mouse clocks use RRBS. Model organisms such as *C. elegans* which essentially lack DNA methylation would not, however, be likely to be tractable for DNAm based age prediction. Consequently DNAm age prediction at least partially meets criterion number 4.

Is DNA methylation a better predictor of lifespan than chronological age? Both Hannum and Horvath identified the difference between DNAm age and chronological age or Δ_{Age} as a potential indicator of disproportionate biological ageing. Marioni et al. [190] explicitly set forth to test the value of DNAm age as a predictor of mortality. Adjusting for age, sex, childhood IQ, education, social class, hypertension, diabetes, cardiovascular disease, and *APOE* e4 status a Δ_{Age} of +5yr is associated with a 16% increased mortality risk (See figure 1.21). A longitudinal study of twins found 3.2 fold increase in the risk of dying first per 5yr Δ_{Age} within twin pairs [191]. A meta-analysis of DNAm age predictors by Chen et al. [192] also found that measures of age acceleration based on Hannum and Horvath clocks were superior predictors of mortality than chronological age before and after correction for various potentially confounding factors. Chen et al. also noted that correction for blood cell composition improved predictive power. Indicating that DNAm age is indeed superior to chronological age as a predictor of lifespan meeting criterion number 1 for an ageing biomarker.

A new DNAm age clock explicitly designed to capture phenotypic age and outperform age acceleration as a biomarker of ageing has recently been produced by Levine et al. [193]. The 'PhenoAge' metric was created by selecting nine biomarkers of ageing from 42 possible metrics using a proportional hazards penalised regression model and combining these with chronological age. (The nine biomarkers are: Albumin, Creatinine, Serum glucose, C-reactive protein, Lymphocyte per cent, Mean red cell volume, Red cell distribution width, Alkaline phosphatase, White blood cell count.) DNAm data was then regressed against PhenoAge using elastic-net regression to produce a DNAm based PhenoAge predictor which made use of 513 CpGs. The PhenoAge predictor outperformed the Horvath and Hannum clocks at predicting all-cause mortality, comorbidities, coronary heart disease risk, and measures of physical functioning. 41 of the 513 CpGs in the PhenoAge clock were present in the original 353 CpG sites used in the Horvath clock.

The 'GrimAge' clock [194] takes a different approach to previous epigenetic clocks in that it has a two-tiered model predicting first seven surrogates for a variety of other biomarkers of ageing then fitting a model to predict time to death using these surrogates. 'GrimAge' outperforms age accelerations from the Horvath, Hannum and PhenoAge clocks at predicting time to death [195].

Belsky et al. developed a DNA methylation based estimate of the rate of biological ageing called 'DunedinPoAm' [196].

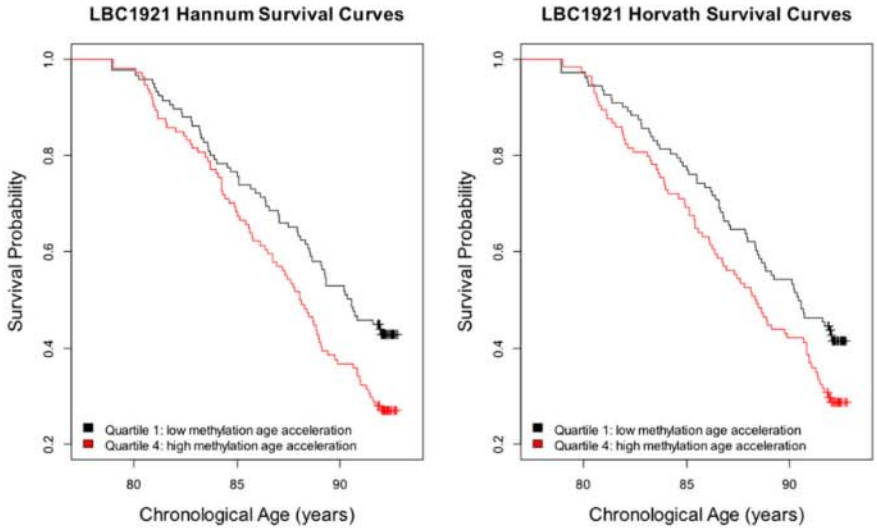

Figure 1.21: Survival probability by quartiles of Δ_{Age} in LBC 1921 adjusted for sex, and chronological age. LBC = Lothian Birth Cohort. Using the Hannum and Horvath predictors' values for Δ_{Age}. (Reproduced from [190] figure 3.)

1.5.4 DNA methylation clock measurements as outcomes for anti-ageing interventions

Fahy et al. showed in a small exploratory study that they could reduce epigenetic age, as measured by four different epigenetic clocks [197]. The clocks used were: Horvath et al. [168], Hannum et al. [167], 'PhenoAge' Levine et al. [193] and 'GrimAge' Lu et al. [198]. The study called TRIIM (Thymus Regeneration, Immunorestoration, and Insulin Mitigation) consisted of ten men aged 51-65 who were treated with a regimen of recombinant human growth hormone (rhGH) in combination with dehydroepiandrosterone (DHEA) and metformin. The aim of this treatment was to counteract the decline in immune function attributed to the decrease in diversity of t-cell receptors that occurs in this age group [199]. rhGH was found to stimulate the reconstitution of the thymus and thereby T-cell levels in HIV patients, hence it was reasoned that it might also counteract the age-related deterioration of the thymus. DHEA and metformin were included in the regimen to counter the hyperinsulinemia which can be induced by growth hormone [200]. The mean decrease in epigenetic age after 12 months treatment was 2.5 years so a reduction of 1.5 years from baseline, the rate of this decrease appeared still to be increasing in the final 3 months of treatment (9-12 months). There was incomplete regression towards pre-treatment levels of age acceleration at 18 months from onset, 6 months after cessation of treatment, in 3 of the 4 epigenetic clocks. There were however no sign of change in the GrimAge [198] acceleration

estimate which remained at, on average, 2 years lower than age acceleration at onset. This study sets an important precedent for for the use of DNAm age acceleration as a readout of the effectiveness of anti-ageing interventions, opening the door for their use in future studies. The challenge faced by studies where the intended outcome is increased lifespan is that this requires a long, expensive, and impractical follow-up period to measure directly. If DNA methylation based measures of age acceleration can be used as effective proxies for measuring lifespan then this dramatically decreases the cost and inconvenience of studying interventions which effect lifespan. Clearly however it needs to be established how good a proxy for actual health/lifespan increases DNA methylation age metrics actually are before they can be relied upon.

Ascertaining the effectiveness of using DNA methylation based age acceleration metrics as an outcome in ageing and longevity interventions in model organisms is likely to be informative about how well they can be expected to perform in humans. Promisingly, differences in interventions known to effect longevity in model organisms such as caloric restriction, rapamycin treatment, ovariectomy in female mice, and genetically long lived mouse strains [201,202,186] have consistently shown decreased epigenetic ageing under the pro-longevity conditions. Horvath et al. recently produced a single epigenetic age predictor which works in both humans and rats using the same assay of conserved CpG sites and the same age prediction algorithm [203]. This unified metric provides greater confidence that an intervention which improves ageing outcomes in rats will also do so in humans. Horvath et al. also demonstrate the effectiveness of Heterochronic Plasma Exchange (HPE), the replacement of the plasma of old rats by that of young rats at rejuvenating the rat tissues. They observed an average of a 54.2% reduction in epigenetic age across liver, blood, heart, and hypothalamus with their pan tissue clock in old rats which underwent HPE [203].

These results have laid the groundwork for the first large scale trial of an anti-ageing intervention. Metformin has a relatively small but well documented effect on lifespan and reduced risk of a number of ageing related conditions in model organisms and in humans with type two diabetes, for which it is widely prescribed [204]. The TAME (Targeting Aging with Metformin) trial plans to include over 3,000 individuals aged 65-79 in a double-blind placebo controlled trial to examine the effects of metformin treatment on a composite primary endpoint consisting of stroke, heart failure, dementia, myocardial infarction, cancer and death [205]. The biomarker work group indicated that DNA methylation ageing assays may not be among the primary set of biomarkers used in the trial [206]. However, TAME will establish a repository of bio-specimens including blood which will likely permit the examination of the effects of metformin on DNA methylation based ageing metrics in humans at a large scale.

1.6 Aims

Going forward this work covers several aspects of change in DNA methylation related to ageing phenotypes. Starting at the beginning of the human life-course EWAS will be performed for *in utero* markers and interventions thought to impact on long-term bone health. An examination will be undertaken of the age-related changes in the methylation state of a key functional region of the genome, whose methylation state has not previously been characterised in detail, namely the tRNA genes. Lastly, new DNA methylation clocks based on a family of repetitive elements, specifically the Alu repeat family, will be generated. This will make use of another region of the genome the DNA methylation state of which has not previously been characterised in detail, but which represents a substantial proportion of all CpG sites, the derepression of which has long been considered to play a role in ageing.

1. Identify epigenetic associations with bone health outcomes and of vitamin D supplementation during pregnancy as an intervention to improve bone outcomes [207,208]. Provide some groundwork for mechanistic studies to extend understanding of what systems influence bone health by identifying what genomic features change epigenetically with bone traits and interventions intended to affect these traits.

2. Identify Age-related changes in DNA methylation in regions of the genome characterised in the TwinsUK MeDIP-seq dataset and poorly covered by or covered in small samples by other technologies. The regions to be examined are:

    a. The tRNA genes, following up on the previous finding of age-related DNA hypermethylation at tRNA-iMet-CAT-1-4 [209]. tRNA genes have a core role in cellular metabolism, many emerging regulatory functions both structural and as signalling molecules. tRNA genes also interface with many systems the modulation of which impact longevity, making any age-related changes in their epigenetic state potentially very consequential.
    b. Alu repeat elements, a primate-specific family of SINEs present in over 1 million copies in the human genome. The aim is to construct a DNA methylation clock using only these elements. Age acceleration based on this clock may yield different information about biological ageing than from previous clocks as the effects of DNA methylation changes differ between repressed repetitive elements and enhancers or promoter proximal regulatory sequences.

Chapter 2

Methods

2.1 Illumina DNA Methylation arrays

The Illumina methylation bead chip arrays make use of 50 bp long probes with sequences corresponding to the loci of interest in the genome, with an additional 23 base sequence for positional addressing on the array [149,154,162,210]. The CpG of interest is located at the 3' end of the probes. These probes are attached by their 5' ends to $3\mu m$ diameter silica beads which are embedded in indentations acid etched into the surface of planar silica slides. Each locus is represented by an average of ~30 beads and a minimum of ~5 beads which are distributed randomly across the slide. The number of beads for a given locus is approximated well by the poisson distribution, so this can be used in conjunction with the number of loci to ensure that the probability of having fewer than 5 beads per locus is kept negligibly small by adjusting the ratio of loci to possible bead sites. Thus each BeadChip array has a unique random distribution of probes which must be decoded to yield the mapping between positions on the array and target loci being assayed. The positions of the beads are decoded through the sequential hybridisation of fluorescent oligonucleotide probes within the 23 base addressing sequence, with probes for each locus having a unique combination of hybridisation events to identify them [211,212]. An optical scanner such as Illumina's iScan system with a charge coupled device (CCD) to detect the light and a laser to excite the flourescent dyes is used the read the arrays. There are two fluorophore channels the respective fluorescence intensities of which are are recorded at each spot. Initially the probe decoding oligos are used to map the probe locations and then flourescently tagged DNA bases are cycled to assay the DNA methylation as described below.

Samples are prepared for the array assay by fragmentation, bisulfite conversion and amplification. bisulfite conversion creates a 'pseudo-SNP' by selectively deaminating unmethylated cytosines to uracil and thus changing their base pairing properties and permitting this technology, originally developed to detect SNPs, to assay DNA methylation [149]. These sample fragments anneal

to complementary probes on the array and unbound DNA is washed away. A single DNA polymerase-mediated elongation step is carried out to permit a fluorescently labelled base to be incorporated at the end of each probe. C and G are tagged with one fluorophore and A and T with another. There are two probe designs:

1. The type I probes, which make use of two distinct bead types for each locus.
2. the type II probes, which make use of a single type of bead at each locus.

The 'colour' of the fluorophore incorporated signifies the methylation state and the intensity of the proportion of sites which are in that state. The logic of which 'colour' corresponds to methylated or unmethylated varies with the probe type and Figure 2.1 illustrates and explains this in-depth.

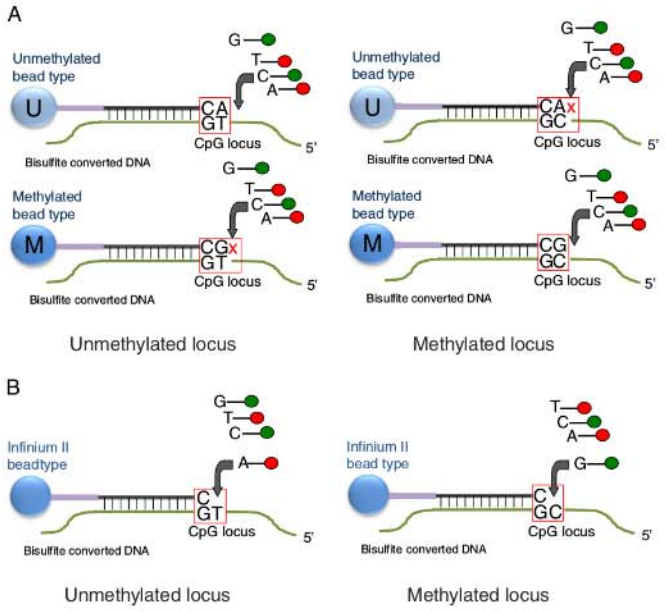

Figure 2.1: **Illumina Methylation Bead Chip Array Probe Designs** The probes on the array are 50bp in length. In type I probes, unlike type II probes the base that is incorporated (or not) in the elongation step is the base that **follows** the CpG site. The probes on the array are bound by the amplification products derived from the bisulfite converted DNA, not direct conversion products. Thus, Cs in the sample DNA represent methylated positions and Ts unmethylated positions, as Ts take the place of Us created by bisulfite conversion during the amplification process. **Type I**: Panel A shows the type I probe design. Each locus is represented on two beads, a methylated and an unmethylated bead. *Methylated Sample*: When sample DNA binds a methylated probe the G at the second position in the CpG on the methylated probe will be complementary to the C marking the location of the 5mC. Extension will then be able to occur incorporating a fluorescently labelled base complementary to the base just 5' of the 5mC position in the probe. On the unmethylated probe, the C marking the position of the 5mC will not be complementary to the A base in the probe and thus extension cannot occur. No fluorescent base will be incorporated at the unmethylated probe. *Unmethylated Sample*: When sample DNA binds the methylated probe the T marking the position of the unmethylated C is not complementary to the G at the second position in the CpG site on the probe. Thus, no fluorescent base is incorporated. Whereas, on the unmethylated probe the T marking the position of the unmethylated C will be complementary to the A at the second position in the CpG site on the probe. Thus, a fluorescently labelled base can be incorporated. **Type II**: Panel B shows the Type II probe design. For the Type II probes only one probe represents each locus. In the Type II probes the incorporated base is at the second position in the CpG site. A fluorescently labelled A is always incorporated opposite a T in the sample DNA marking the position of an unmethylated C, and a fluorescently labelled G is always incorporated opposite a C marking the position of an unconverted 5mC in the sample DNA. For the Type II probes, in contrast to the Type I probes, methylation is always signalled in the green channel and unmethylation in the red channel. (figure reproduced from Bibikova et al. [162].)

These arrays have now undergone several iterations [154,162,210]. The current array probe design was preceded by the 'golden gate' array which used a variant on the SNP probe design with methylation-specific PCR. This array contained ~1.5k probes covering 371 genes with a strong focus of cancer-related genes [149]. The first array using the probe design outlined above was the 'Infinium' array which contained 27k type I probes focused on the promoter regions for 14,475 consensus coding genes with 110 miRNA promoters [154]. The 450k array which had ~480k mixed type I and II probes covering 21,231 RefSeq genes, 26,658 CpG islands, 80,538 predicted enhancer regions and a variety of other features, including the MHC region [162]. The 450k array systematically underestimates the methylation level in highly methylated regions [213]. A Markov chain model of DNA methylation state developed by Affinito et al. [163] suggests that once a threshold level of DNAm is reached on a molecule further methylation becomes less likely. They reason that DNMTs recruited by the high methylation state have greater difficulty accessing the remaining unmethylated sites. Thus sampling a single site is likely to produce a systematic underestimate as once neighboring sites are methylated methylation at a given CpG is less likely to increase further. The 'EPIC' array has ~850k probes including >90% of those on the 450k with greatly expanded coverage of loci with more dynamic methylation states than promoters and CGIs [210]. Promoters and CGIs tend to have relatively low methylation variability and the EPIC array aims to capture more functional methylation variation at regions such as enhancers [86]. The 'EPIC' array also contains both Type I and Type II probes, with many of the new sites being type II. The type I probes have greater dynamic range than the type II but take up twice as much space on the array so there is some trade-off between maximising the number of sites covered and the quality of data at those sites.

Methylation at each site on the array is commonly reported as a β value, which corresponds to the proportion of the sample DNA which was methylated at that site. This is computed from the intensity values extracted from Illumina's IDAT format files which in the case of the methylation arrays are a binary format which had to be reverse engineered to permit analysis outside of Illumina's 'GenomeStudio' [214].

$$\beta = \frac{intensity[M]}{intensity[U] + intensity[M]}$$

Where M = methylated and U = unmethylated.

The type I and type II probes have slight systematic differences not accounted for by their genomic context [215]. Type I probes have a wider range and are more reproducible than the type II probes. This is likely due to the dual complementary probes with fluorescence in different colour channels providing additional information for methylation level calling leading to more robust estimates at extreme values near 1 and 0. This difference means that to operate uniformly on this mixed data a normalisation procedure is needed to correct for these differences.

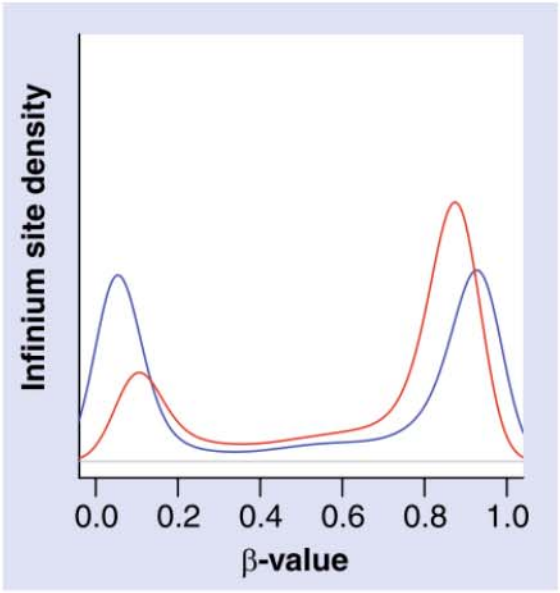

Figure 2.2: **Type I probes have a wider range and less variability than type II**. Type I probes shown in blue, Type II in red. Plot shows kernel density of beta values with gausian smoothing. (Reproduced from Dedeurwaerder et al. [215])

There exist a number of normalisation methods to account for probe-type bias [216] such as beta-mixture quantile normalization (BMIQ) [217]. This method was applied as a part of the pre-processing of the Twins UK 450k array datasets used in (Chapter 4) in this thesis. In addition, there are other techniques capable of correcting for probe bias as well as other sources of technical variation such as batch effects [218]. Functional normalisation [219] which makes use of the technical control probe data on the arrays to inform its transformation of the DNA methylation values was used to normalise EPIC array data from the MAVIDOS and Southampton Women's Survey (SWS) data in (Chapter 3) of this work.

Bisulfite conversion efficiency can be estimated by looking at the rates of conversion at non-CpG cytosines as these are very rarely methylated they should be bisulfite converted at a very high rate. If this is not the case it can lead to the overestimation of DNA methylation levels, bisulfite conversion also fragments DNA so there is a trade-off between achieving sufficient conversion efficiency and not producing overly fragmented DNA [220]. On the Illumina methylation arrays control probes ([221] p226) are located at non-CpG cytosines to assay the conversion efficiency for use in quality control assessment of array data.

In addition to probe-type bias, there are other sources of potential problems for individual probes on these arrays [222]. Zhou et al. divide these into three categories [223]:

1. Probes with SNPs internal to the probes especially if located near the 3' end. These can cause issues by:
    a) preventing extension through mismatches near the 3' end.

   b) Altering the ability of the target cytosine to be methylated through effects on the sequence context, including the case where a C/T polymorphism in place of the C/T bisulfite induced conversion.
   c) Inducing a colour change in a type I probe by altering the extension base.
2. Probes which map to multiple possible locations in the bisulfite-converted genome. Methylation values from multi-mapping probes are an amalgam of the methylation levels at the different loci to which the probes can map and cannot be disentangled to yield reliable results for the different loci.
3. Probes with partial off-target matches to other loci. Probes with partial matches at the 3' end can lead to extension and an erroneous readout of methylation levels for that locus or partial matches elsewhere can lead to competition for binding with the probe intended for that locus.

To account for these potential problems probe 'masks' have been produced which identify probes which commonly exhibit one or more of these problematic properties so that findings at these loci can be excluded or followed up for validation with other methods [223,222]. These masks have been used here to identify potentially problematic probes.

When looking for a purely epigenetic effect the presence of a genetic variant which influences the DNA methylation state of a CpG is known as genetic confounding. This is not necessarily a problem if genetic effects on a phenotype of interest are a useful outcome but it is an issue if looking for effects mediated purely through epigenetic mechanisms. Genetic variants which bias the DNA methylation in a particular direction, or modulate the plasticity of DNA methylation state to environmental influences may be important to mechanistic understanding the biological system under examination [224,225]. Simply excluding all probes with a SNP above a certain minor allele frequency and with a certain distance of a target CpG can both miss rarer genetic confounders and throw away actual non-genetic differences in DNA methylation. Andrews et al. developed a method called 'gap-hunting' to identify probes which commonly have multi-modal DNA methylation distributions which may be associated with different genotypes influencing a probe locus [226]. 'gap-hunting' was also employed in this work to search for DNA methylation changes which may be under genetic influence (Chapter 3). In addition to SNPs causing sequence context changes that affect the DNAm of probes, structural mutations can also have this effect on DNA methylation [107]. Thus probes in the vicinity of non-SNP mutations should also be subject to additional scrutiny.

2.2 MeDIP-seq

MeDIP-seq uses a monoclonal anti-5mC antibody to bind denatured fragmented genomic DNA at methylated CpG sites. This antibody-bound fraction of DNA is isolated and sequenced [127],

Figure 2.4 outlines the workflow. Unlike bisulfite conversion approaches this method permits 5mC to be differentiated from 5hmC, as the antibody binds specifically to 5mC. The methylation level across the CpG sites in the region of the genome to which the resultant sequencing reads map can subsequently be estimated by counting the number of reads and accounting for the CpG density with software tools such as the `MEDIPS` R package [128].

MeDIP-seq can also cover much more of the methylome than do the 450k or EPIC arrays. The 450k array captures ~1.6% of the 28 million CpGs in the genome. At near saturation coverage of ~40 million reads per sample, MeDIP-seq can cover ~60% of all CpGs with at least 1 read, this is almost all methylated CpGs. At half that total read count MeDIP-seq still covers ~50% of CpGs at $\geq 1\times$ and ~20% at $\geq 5\times$ [227]. Other estimates place the threshold for saturation coverage by MeDIP-seq lower at between 20 and 30 million reads and suggest that the maximum number of CpGs covered at least once could extend up to 90% [228]. MeDIP-seq provide a particular advantage by comparison with the Illumina beadchip array technologies with respect to the coverage it provides for repetitive regions of the genome. This is illustrated by Clark et al. who show that MeDIP-seq covers an estimated 91.4% of repetitive sequences and the 450k array only 3.4% [213] (Figure 2.3). This makes it particularly suited to the aims of this project (Chapters 4 & 5).

MeDIP-seq does however have the disadvantage that it is less good at absolute quantitation of DNA methylation levels than bisulfite conversion base methods. This is especially the case without input controls i.e. sequencing data without antibody pull-down to account for the 'sequencability' or background levels of reads that are generated for a given region in the genome which can be corrected for when estimating methylation levels. Better even that input controls is to couple MeDIP-seq with MRE-seq which makes use of methylation-sensitive restriction digests to enrich for unmethylated DNA [81]. Signal in MRE-seq thus corresponds to the genomic location of unmethylated regions and can thus be used to differentiate gaps in MeDIP-seq coverage from truly unmethylated regions. In conjunction these complementary sequencing technologies provide an improved ability to discern absolute methylation levels compared to MeDIP-seq alone [131]. The data generated on behalf of TwinsUK and used in this thesis did not make use of input or MRE-seq controls. Additionally Data for Chromosome X were not available for use in the analyses performed in this thesis. Interpretation of changes in chromosome X DNA methylation with age would be interesting to examine especially in this almost exclusively female cohort, despite the challenges in interpreting such changes given X inactivation.

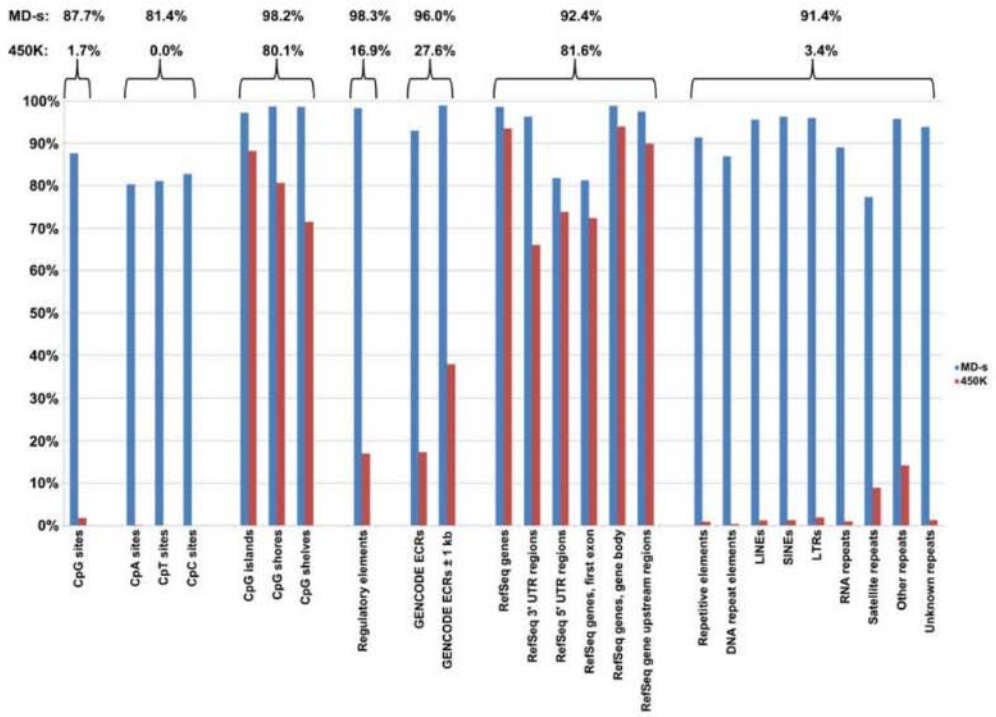

Figure 2.3: MeDIP-seq provides superior coverage of repetitive regions of the genome than the Illumina array platform. This presumes saturation coverage for the MeDIP-seq assay. MD-s = MeDIP-seq, 450k = Illumina 450k DNA methylation bead chip array (Reproduced from Clark et al. [213]) The CpG islands, shores, and shelves major RefSeq gene features are covered almost as well in the 450k array as they are by MeDIP-seq but repetitive regions and other regulatory elements are where MeDIP-seq provides substantially greater coverage.

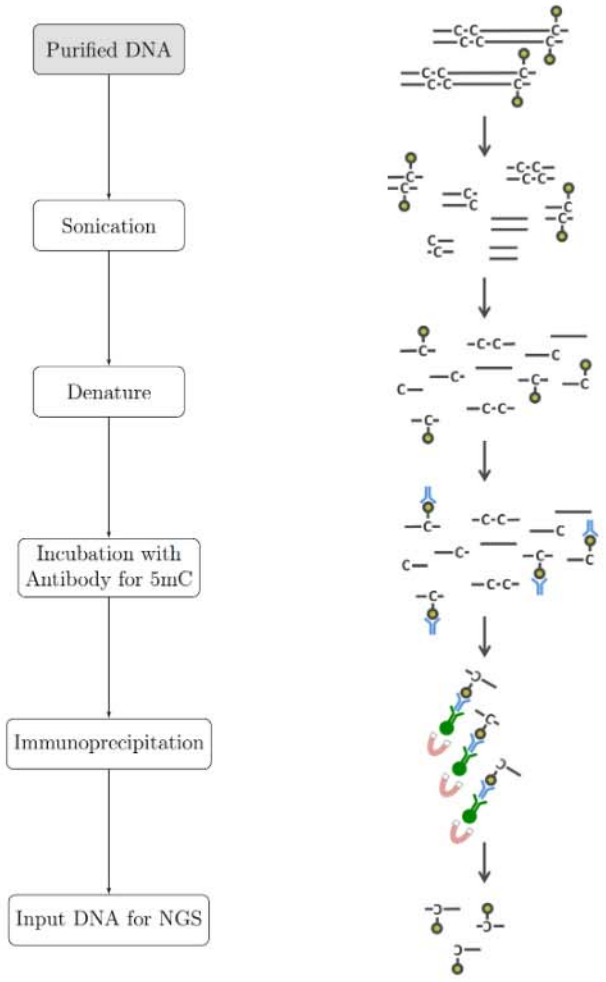

Figure 2.4: **Graphical Summary of the MeDIP-seq process** Purified DNA is fragmented by sonication, denatured and incubated with anti-5mC antibodies. It is then immunoprecipitated resulting in fragments containing methylated CpGs which are subsequently sequenced.

DNA sample preparation, MeDIP reaction, and Illumina sequencing were performed by BGI-Shenzhen (Shenzhen, China) on behalf of Twins UK, prior to the analytical work performed on the resulting data in this thesis. Fragmentation of genomic DNA from whole peripheral blood samples from the Twins UK cohort was performed by sonication using a Covaris system (Woburn, MA, USA). $5\mu g$ of DNA was used in the Illumina Single-End DNA Sample Prep kit followed by end repair, base addition and adaptor ligation. The MeDIP reaction was performed, according to the protocol for the MagMeDIP kit (mc-magme-048), with a monoclonal antibody for 5-methylcytosine (5mC) (Cat. No.: CO2010021 mc-magme-048 from Diagenode (Liège, Belgium)). The antibody was incubated with Adaptor-ligated DNA combining $0.5\mu l$ antibody + $0.5\mu l$ water; then $0.6\mu l$ MagBuffer A, $1.4\mu l$ water and, $2\mu l$ MagBuffer C; yielding a final volume of $5\mu l$ for the immunoprecipitation reaction. The MeDIP reaction was validated with quantitative PCR. The product of the MeDIP reaction was purified with Zymo DNA Clean & Concentrator-5 (Zymo

Research), and amplified with adaptor-mediated PCR. Size selection of fragments (200–500 bp) was performed by gel excision and quality assessed by Agilent BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA). The resultant libraries were subjected to highly parallel 50bp single-end sequencing on the Illumina HiSeq2000 platform.

Quality control (QC) and alignment were performed by Twins UK, also prior to the work done in this thesis. Sequencing data were subject to initial QC for call quality and base composition using `FASTQC` (v0.10.0) (https://github.com/s-andrews/FastQC) [229]. Duplicates were removed with `SAMTools` (https://github.com/samtools/samtools) [230]. The alignment was performed using the Burrows-Wheeler Aligner (`bwa`) (https://github.com/lh3/bwa) [Li2009] with a minimum mapping quality score of Q10. `MEDIPS` (v1.0) was used to perform MeDIP-specific QC, as well as to generate reads per million (RPM) and Absolute methylation score (AMS) values (https://bioconductor.org/packages/release/bioc/html/MEDIPS.html) [128]. AMS and RPM values are binned into 500bp windows with a 250bp slide in the BED format, resulting in ~12.8 million windows on the genome (Build: hg19/GRCh37). Additional quality control checks were performed by Twins UK in the R language: correlation matrix, hierarchical clustering, dendrogram, heatmap, density plot, and batch effects inspection by principal component analysis. Analyses in this thesis are based on these binned DNA methylation level estimates and the accompanying individual metadata, such as age, produced by Twins UK. There was a mean of ~16.89 million (SD 3.04M) reads per sample. This number of reads per sample is consistent with covering ~50% of all CpGs at 1x, ~15% at 5x and ~10% at 10x (estimated using the results of Taiwo *et al.* 2012 [227].) MeDIP-seq data from regions of interest was extracted using `bedtools` v2.17.0 [231].

The above-described processing steps for Methylated DNA Immunoprecipitation sequencing (MeDIP-seq) data used here have been previously described in Bell et al. [209,107]. These processed data are available, subject to approval, in the European Genome-phenome Archive (EGA) (https://www.ebi.ac.uk/ega) under study number EGAS00001001910 and dataset (https://www.ebi.ac.uk/ega/datasets/EGAD00010000983).

2.2.1 Participants

Peripheral blood samples for DNA extraction were provided by the adult volunteers from the UK Twin Register (TwinsUK Resource [232]). The participants are rigorously phenotyped at visits at St Thomas' Hospital, London. A twinning questionnaire determines twin zygosity and is confirmed by genotyping. Ethics were approved by Guy's & St Thomas' NHS Foundation Trust Ethics Committee (EC04/015—15-Mar-04) and written informed consent was obtained from all subjects in accordance with this. Samples are stored at −80 °C in EDTA tubes before extraction via the Nucleon Genomic DNA Extraction Kit. DNA is subsequently stored in TE

Buffer at −20 °C.

The MeDIP-seq dataset used in this work consists of 4350 whole blood methylomes with age data. 4054 are female and 270 male. 3001 have full blood counts. There are 3652 individuals in this data set. These individuals originate from 1933 unique families. There are 1234 monozygotic (MZ) twin pairs (2468 individuals), and 458 dizygotic (DZ) twin pairs (916 individuals).

2.3 Targeted Bisulfite Sequencing

Targeted Bisulfite sequencing [233] is very similar in principle to ordinary targeted DNA sequencing but preceded by a bisulfite conversion step as illustrated in Figure 1.16 [131]. However, bisulfite conversion has several implications for downstream processing. Bisulfite library prep can be either directional, sequencing reads correspond to a bisulfite-converted version of either the forward or reverse strand, or non-directional in which sequencing reads correspond to bisulfite-converted versions of either strand giving a total of four bisulfite-converted sequences with the strand of origin unknown. If a library is directional there are only two bisulfite-converted sequences. Other than ability to trace strand of origin there is no particular advantage to either approach. In both cases very rigorous adapter trimming is needed to avoid biases arising from the reads that are longer than the fragments and thus might be inferring methylation state for cytosines introduced during the library preparation process. Library preparation in this work was non-directional.

Mapping can be performed with customised software wrapper to implementations of existing alignment tools. `Bismark` is a popular example of such as wrapper which uses the bowtie alignment tool and calls site methylation levels [234]. Alignment is performed with *in silico* bisulfite converted versions of the genome and reads, with G to A conversions for reverse strand reads. This requires four parallel instances of bowtie for forward and reverse G to A and C to T conversions, which are combined to determine the unique best alignment. (Figure 2.5 panel A). Bisulfite conversion results in loss of sequence complexity as many Cs effectively become Ts meaning short reads can be challenging to map. For targeted sequencing approaches it is advisable to align to both the targeted sites and their flanking regions as well as to the whole genome, especially if any of the targeted regions contain repetitive sequences [235]. Substantial mapping to areas other than those targeted could be indicative of off-target amplification, this step forms a part of the quality control process. `Bismark` produces methylation calls either combined or by strand and which can be filtered by methylation context CpG, CHG or CHH.

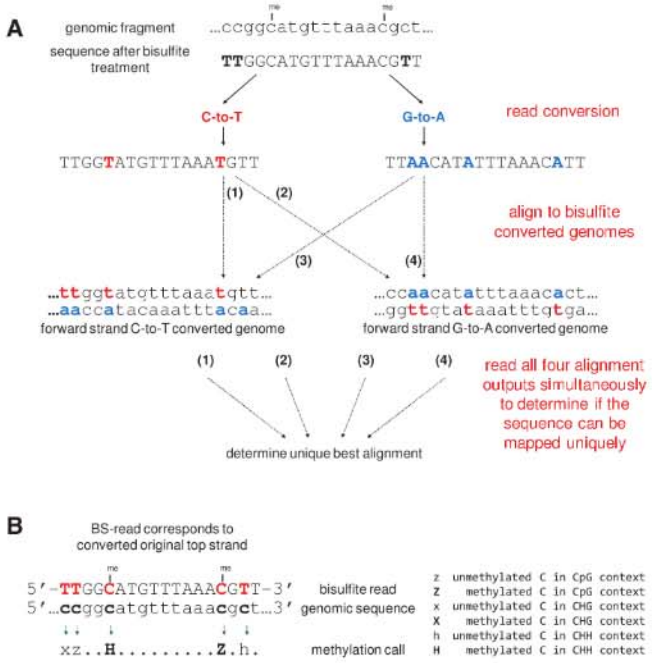

Figure 2.5: **Bisulfite sequencing alignment process A)** Unmethylated cytosines are converted to uracil in the bisulfite conversion process these are rendered as thymidine as T is used in the PCR amplification following bisulfite conversion. Following the conversion of a C to A T its corresponding G is swapped for an A in the complementary strand. The Genome is converted in silico to permit the alignment of these altered bases. Thus there is the C to T converted and a G to A converted genome each to which has two strands totalling 4 possible alignment targets for the converted reads. **B)** The unique best alignment for a read is determined from this set and mismatches of C in this alignment are used to call the methylation state. (Reproduced from Krueger et al. [236])

Targeted sequencing methods where the aim is to sequence a number of amplicons (N_a) from a number of samples (N_s) means the total number of PCR reactions in the pairwise matrix is ($N_a \times N_s$). This can rapidly become difficult to manage with conventional bench-top PCR methods especially where consistency between reactions is important for quantitation. Microfluidic systems for conducting multiple PCR reactions in parallel such as the Fluidigm 48.48 access array employed here improve the ease with which a large number of reactions can be performed.

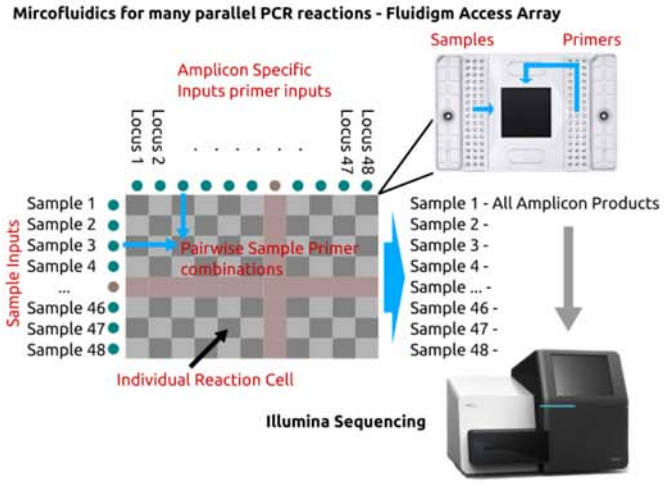

Figure 2.6: **Microfluidic system for multiple parallel pairwise PCR reactions.** The Fluidigm 48.48 access array permits the combination of up-to 48 samples with up-to 48 different set of reagents, without multiplexing. In this case the reagents are locus specific primers for targeted sequencing. Samples are added to one side of the array and reagents to the other the flow cell permits all combinations of the two to be mixed. Fluidigm's controller flow cell and thermocycler allow all the PCR reactions to be carried out simultaneously and their products be extracted for sequencing.

A four primer step-out PCR process is performed on the bisulfite-converted genome to generate the final PCR product used in the Illumina sequencing reaction (Figure 2.7). After the loci of interest are selected locus-specific primers are designed to be complementary to the bisulfite converted regions flanking the area of interest and avoiding CpG sites in those flanking regions. A number of tools for designing such primers are available, the tool used in this work was: 'methPrimer' [237].

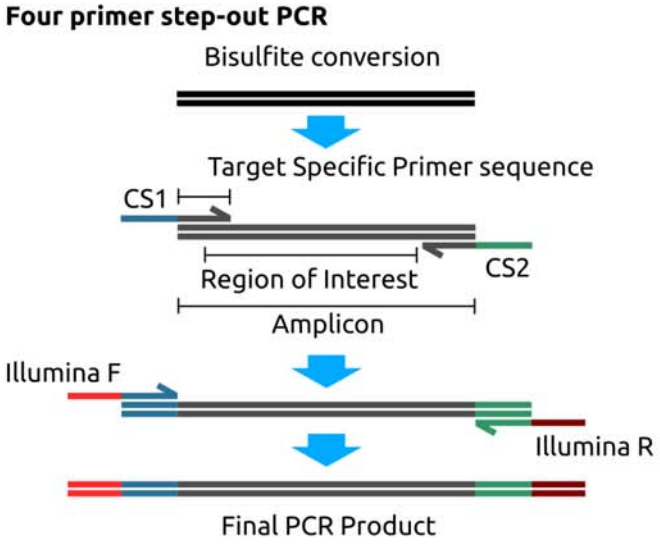

Figure 2.7: **Four primer step-out PCR for targeted bisulfite sequencing on the Fluidigm access array.** The sample DNA is bisulfite converted. The region of interest amplified using site specific primers with the Fluidigm CS1 and CS2 primers in the first PCR step. The CS1/2 primers have regions complementary to the Illumina sequencing primers permitting them to be added in a second PCR step to generate the final PCR product ready for sequencing.

After designing the target-specific primer sequences the CS1 and CS2 adaptor sequences are added to the forward and reverse primers respectively to permit the addition of Illumina sequencing primers in a second PCR step. The subsequent analysis is similar to that for whole-genome bisulfite sequencing. The PCR products are sequenced and the resulting reads aligned and methylation levels called with `Bismark`.

Chapter 3

Epigenome Wide Association Studies for Bone Health Outcomes in Umbilical Cord Blood and Tissue

3.1 Abstract

Long term bone health and fracture risk is strongly influenced by the peak bone mass that is attained in early adulthood, this has its origins in development from late pregnancy through early childhood. The availability of vitamin D to facilitate the uptake of calcium and phosphate is essential to healthy bone development, in addition to the role of vitamin D as a hormone mediating this process. Greater mechanistic understanding of this process is desirable to permit the design of more effective interventions to reduce the substantial health burden of osteoporotic fractures in the elderly. The risk of osteoporosis increases with age, and may be mitigated by improvements in bone health. Maternal vitamin D levels during pregnancy have previously been shown to impact on the methylation status of specific genes associated with vitamin D signalling and metabolism, such as *RXRA*. Long-term bone health outcomes have also been associated with both maternal vitamin D during pregnancy and methylation of the *CDKN2A* gene. To attempt to identify other loci, the epigenetic regulation of which is pertinent to the impact of maternal vitamin D levels on long-term bone health, Epigenome-Wide Association Studies (EWAS) were performed with DNA methylation data from the Illumina EPIC & 450k methylation bead chip arrays. We provisionally identify two CpGs whose DNA methylation state is associated with

bone mineral content at 6 years of age ($p < 2.52 \times 10^{-8}$, n = 402) and periosteal circumference at 6 years of age ($p < 4.24 \times 10^{-8}$, n = 141) respectively. These associations are in need of replication in an independent cohort but present a starting point for further investigations.

3.2 Introduction

In England and Wales a survey of general practice data indicates that 53% of women and 21% of men will experience a bone fracture in their remaining lifetime from the age of 50 years [238] (Figure 3.1). Hip fractures in particular are associated with a reduction in survival of an estimated 10-20% with most deaths occurring in the first 6 months, vertebral fractures though approximately an order of magnitude less common carry a similar mortality risk [238,239]. The national osteoporosis foundation estimated the societal cost of osteoporosis at 58 Billion US$ for 2018 in the USA alone [240] with serious negative implications for the quality of life of those with this condition. Thus interventions to increase bone health and reduce the rate of osteoporosis as well as developing a mechanistic understanding of the development of this condition are of considerable medical interest.

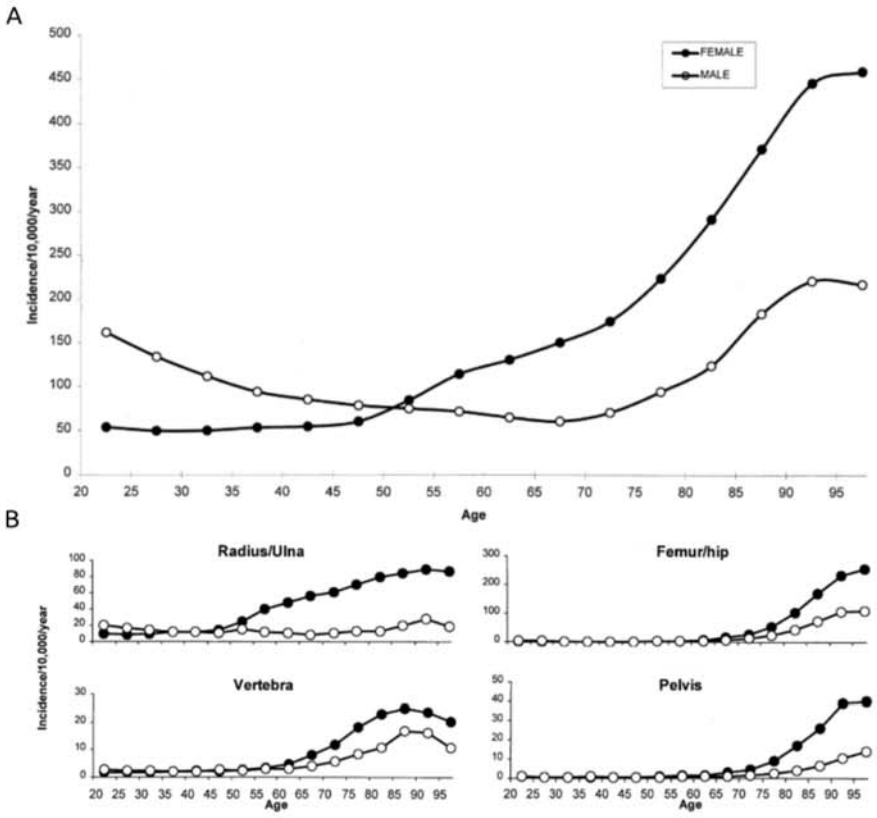

Figure 3.1: **Overall Fracture Incidence Increases With Age** Data from 5 million adults in the General Practice Research Database 1988-1998. **A** Incidence of all fractures at any site by sex and age. **B** Fractures by sex and age at select sites, pelvis, Femur/Hip & Vertebra confer the greatest increased mortality rates and Radius/Ulna has the greatest sexual dimorphism. (Adapted from van Staa et al. [238].)

Modelling work indicates that the largest determinant of osteoporosis risk is peak bone mineral density (BMD) with a 10% change in peak BMD leading to a predicted 13 year delay in the development of osteoporosis [241]. Though the baseline BMD is a better predictor of fracture risk, the rate of loss also independently predicts risk [242–244]. Thus factors altering the peak bone mass will be important for modulating osteoporosis risk. Bone mass increases from intrauterine development through to a peak in early adulthood (Figure 3.2).

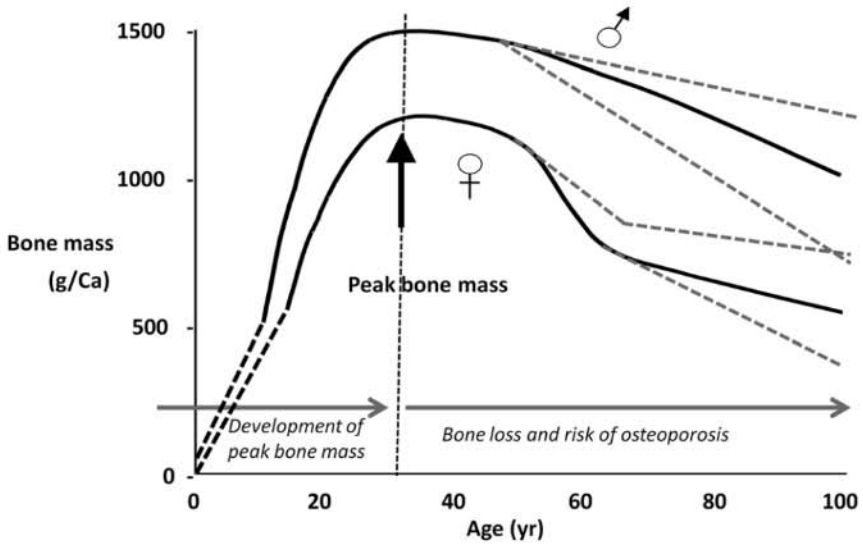

Figure 3.2: **Bone Mass Over The Lifecourse** Bone mass increases from intrauterine development through to a peak in early adulthood. Intervention in early life to modulate growth trajectory and increase peak mass may provide a higher starting point and delay bone loss in later life. (Reproduced from [245].)

Calcium, phosphate and vitamin D are needed for bone development [246] and maintenance. Vitamin D is needed for sufficient intestinal absorption of calcium [247] and phosphate [248] in addition to the hormonal role it plays in the regulation of bone development. Calcium and vitamin D intake are modifiable factors and thus constitute a reasonable starting point for interventions to improve bone health. This raises the question of how and when they can best be deployed to maximise their impact on bone health.

Birth weight predicts bone mass in adulthood [249]. A longitudinal study with long term follow up at Helsinki University Central Hospital linked low childhood growth rate with greater fracture risk in later life [250]. Data from the Southampton Women's Survey (SWS)[251] suggest that intervention in the interuterine period and infancy may afford the greatest opportunity to impact on long term bone health. Rate of growth in late pregnancy (19-34 weeks) and early post natal growth (up to ~2 years) predict bone mass at 4 years of age [252], and estimates of hip strength at 6 years [253]. The proportion of individuals crossing between tertiles of the distribution of growth in length decreases with age, indicating that the variability and thus potential malleability of growth is greatest during the period from 11 weeks gestation to ~2 years.

Observational data indicated a relationship between maternal vitamin D status and bone outcomes in children [254,255]. Prospective studies showed the effects of maternal vitamin D status on bone growth in newborns [256] and at 20 years of age [257]. MAVIDOS (Maternal Vitamin D Osteoporosis Study) is a randomised, double-blind, placebo-controlled trial of oral vitamin D

supplementation (1000 IU cholecalciferol/day or placebo) from 14 weeks gestation in women with initial circulating 25(OH)-vitamin D levels of 25-100 nmol/l [207,208]. The primary outcome of the study was whole body bone mineral content, assessed by Dual Energy X-ray absorptiometry (DXA) within 14 days of birth, and follow up DXA at 4 years of age. MAVIDOS did not find significant differences in neonatal bone outcomes overall but there were significant differences between intervention and placebo for births taking place during the winter months. Winter births saw a mean difference in BMC between treatment groups of 5.5 g [95% CI 1.8-9.1]; p=0.004) and for BMD a mean difference 0.01 gcm-2 [0.00-0.02]; p=0.04 [208].

$1,25(OH)_2-Vitamin\ D$ has been implicated in the control of Plasma membrane Ca^{2+}-ATPases (PMCAs). The expression of the *PMCA3* gene in the placenta has been associated with umbilical cord calcium concentration and intrauterine accrual of bone mineral content [258]. This provides part of a candidate mechanism linking intrauterine vitamin D availability to bone outcomes, but much remains to be elucidated about how maternal vitamin D status may impact bone development. Work in rat models has demonstrated the ability of maternal nutritional exposure during pregnancy to influence the epigenetic state and expression of genes in their offspring, specifically a change in DNA methylation was identified [259–261]. Consequently, the DNA methylation status of target genes of interest for vitamin D and bone development related pathways have been examined [262].

Retinoid-X-receptor-alpha (RXRA), which forms a heterodimer with the vitamin D receptor is essential for the nuclear action of $1,25(OH)_2-Vitamin\ D$. Harvey et al. found that DNA methylation at four of ten CpG sites in the *RXRA* gene were significantly ($p \leq 0.05$) lower in the umbilical cord of offspring from cholecalciferol supplemented mothers compared to placebo with a mean difference in methylation of -1.98% (n=447, 95% CI -3.65 to -0.32, p=0.01). One CpG site in the gene is related to estimated free 25(OH)D levels [263].

Methylation of cyclin-dependent kinase inhibitor 2A (CDKN2A) in umbilical cord tissue has been implicated in bone cell activity mediating skeletal development and homeostasis. The *CDKN2A* locus has quite complex biology encoding two cell cycle inhibitors $p14^{ARF}$ and $p16^{INK4a}$ as well as the long non-coding RNA *ANRIL* which inhibits $p16^{INK4a}$. Targeted analysis of nine CpGs in a 300bp stretch of the *ANRIL* promoter region of *CDKN2A* was carried out by Curtis et al.. Methylation at several of these sites showed an inverse correlation with bone size, mineral content, and mineral density at age four years [264].

Epigenetic states are a product of both genetics and environment. They exist on a continuum from complete or obligatory genetic determination through facilitative variation influenced by genetics to the agnostic to genetic state. When epigenetic state is obligatory it is effectively determined by the genetic state. When epigenetics is facilitative it has a genetic bias for a particular state but which is subject to environmental influences including the intracellular environment [265]. At the opposite extreme to a genetically obligatory epigentic state, the epigenetic state may

be essentially agnostic to the genetic state and be determined entirely by environmental factors. This picture is complicated by the observation that genetics can not merely influence epigenetic state but also the extent to which that state will vary depending on environmental factors [266]. Epigenome-wide association studies have been performed on DNA methylation data for many traits including age (Section 1.5.1), smoking status [267], exposure to atmospheric particulate matter [268], and obesity [269]. Some robust and reproducible DNA methylation changes have been identified with EWAS, particularly for traits such as age and smoking. Unlike genome-wide association studies (GWAS) associations found by EWAS can be due to reverse causation [224]. That is to say, a genetic variant associated with a disease is very unlikely to have been caused by the disease state whereas an epigenetic association can be the consequence of a disease state. Whilst this can complicate the dissection of disease aetiology it can also be a boon in unpicking the biology that follows on from a particular disease or environmental exposure by revealing what biological networks are affected.

The previous work identifying DNA methylation changes at the *RXRA* [263] and *CDKN2A/ANRIL* [264] loci associated with childhood bone outcomes lead to an interest in establishing if DNA methylation changes were occuring at other loci. EWAS using the Illumina 450k and EPIC array technologies were performed for maternal Vitamin D status and childhood bone outcomes in data from the MAVIDOS and SWS cohorts.

3.3 Primary Questions

Are there DNA methylation changes associated with maternal vitamin D levels?

Are there DNA methylation changes associated childhood bone health markers?

What the DNA methylation changes with vitamin D and Bone outcomes at *RXRA* and *CDKN2A* can be observed?

3.4 Methods

3.4.1 Outline of EWAS

Three sets of EWAS analyses were performed: Phase I of the MAVIDOS EPIC arrays and comparison with 450k array results. Phase II of the MAVIDOS EPIC analysis with additional samples. Analysis of the SWS cord blood EPIC array data.

1. MAVIDOS phase I, cord tissue, results (EPIC n = 140, 450k n = 60), EWAS outcomes:

· Intervention Vs. Placebo

- Bone Mineral Content at Birth (g), measured by DXA
- Maternal circulating 25(OH)-vitamin D levels (nmol/l) at 34 weeks
- Change in Maternal Vitamin D form 11 to 34 weeks

2. MAVIDOS phase II, cord tissue, results (EPIC n = 237), EWAS outcomes:

- 4 year Total BMC (g), without heads* , measured by DXA

3. SWS cord blood, EWAS outcomes:

- 8 year Total BMC (kg), without heads*, adjusted for sex and age, measured by DXA (n = 408)
- 6 year Total BMC (kg), without heads*, adjusted for sex and age, measured by DXA (n = 402)
- 6 year Periosteal circumference of the Tibia at 38% from the distal end (mm), measured by Peripheral Quantitative Computed Tomography (PQCT) (n = 141)
- 6 year Cortical density of the Tibia at 38% from the distal end (mg/cm3), measured by Peripheral PQCT (n = 141)

*'without heads' indicates that bone mass contributed by the head is not included in the total as the head is not scanned during the DXA imaging.

DNA was extracted from cord tissue for the MAVIDOS studies by Dr Nevena Krstic at the Institute of Developmental Sciences, University of Southampton. Phenotypic data and batch information was provided in spreadsheets by Dr Millie Parsons at the MRC Lifecourse Epidemiology Unit, University of Southampton. All array data was provided as raw IDAT files. Analysis of the Illumina EPIC and 450k methylation arrays was carried out in the `R` statistical programming language (v3.5.2) using the `meffil` package [270], which was chosen as it is capable of performing functional normalisation in a more memory efficient fashion than alternatives such as `minfi`.

3.4.2 Functional Normalisation

Functional normalisation is an approach to removing unwanted variation associated with 'batch' effects such as the date on which a sample was analysed or which slide a sample is on which was developed by Fortin et al. [219]. This noise in the data masks the signal associated with the underlying biological effect of interest.

Functional normalisation makes use of control probes on the arrays which are designed to capture only technical variation as surrogates for the sources of unwanted variation. The control probes are processed into 42 summary measures, and principal components analysis is performed on the control probe summary matrices for all samples. The top m principal components (PCs) are used as the surrogates for technical variation going forward. The number, m, of PCs from the

control probe summary matrices used is informed by the amount of residual variation remaining after normalisation. Picking the number of PCs which correspond to the last steep drop in residual variation is the approach recommended by the implementers of `meffil`, and Fortin et al. [219] recommend an m of 2 as performing well across a variety of analyses. (See supplementary material from [219] for details of the control probe summary process.)

The process used by functional normalisation is a variant on quantile normalisation. Instead of forcing the empirical marginal distributions of the samples to be identical at each site across arrays. To produce the normalised values it constructs a quantile function which only removes variation arising from surrogates for batch variation. This approach is effective even when comparing samples with large global differences in methylation levels such as between normal and cancer samples, but cannot overcome high degrees of confounding [219].

3.4.3 Genetically Confounded and Multi-mapping Probes

Probes from different locations in the genome with similar sequences, especially following the reduction in sequence complexity associated with bisulfite conversion, can be cross-reactive on the arrays leading to erroneous signals and are thus commonly excluded from analysis. DNAm is often strongly influenced by genetic factors, the effect is especially pronounced when variants alter the sequence at CpG sites themselves as the site can then no longer be methylated. Thus, sites that have common genetic variants at the probe site are also excluded from the analysis, as well as some sites whose methylation is known to be under strong genetic influence by common genetic variation.

43,254 probes on the EPIC array have been identified as multi-mapping, DNA binding to the probes may be derived from other locations in the genome invalidating these probes as a measure of methylation at their intended loci [222]. 12,510 probes were identified as having genetic variants at the CpG locus they are intending to assay, this can produce misleading results as mutant bases can resemble the products of bisulfite conversion [222]. 1,812 probes where found t overlap regions exhibiting haplotype-specific methylation associated with common non-SNP genetic variants (CNVs, Indels, STRs) and regional in-phase clustering of CpG-SNPs [107]. Zhou et al. [223] provided a list of probes which they recommend 'masking' from the 450k array due to multi-mapping issues, genetic variants overlapping the CpG sites, and other factors which may render results from these probes problematic.

In order to identify any potential additional sources of genetic confounding in the phase I MAVIDOS analysis, Probes with methylation values which cluster into distinct groups were identified using the 'gap hunting' method developed by Andrews et al. [226]. Such distinct clusters of methylation can arise from genetic variants which influence methylation levels being present in homozygous and heterozygous forms in the study population, see Figure 3.15. As the sensitivity

and specificity of 'gap hunting' is limited, it is the advice of the authors not to exclude probes flagged by `gaphunter()` prior to performing EWAS. It is instead advised to check if any of the results appear in this list after the fact and examine the possibility of a genetic effect if they are. In phase II of the MAVIDOS analysis and in the SWS analyses only probes with specific technical QC issues were excluded prior to normalisation and EWAS. Flagging of problematic probes occurred after EWAS on any significant results.

3.4.4 EWAS Models

3.4.4.1 Cell-Type Correction

Cell-type composition is a known confounder in epigenetic studies [271,272]. Observed variation DNA methylation can be cell-type intrinsic, where changes in DNA methylation are not driven by changes in the cell-type composition of the tissue; or cell-type extrinsic, where changes in DNA methylation are due to changes in the cell-type composition of the population sampled (Figure 3.3). An established approach for addressing this potential source of confounding is to add terms to the regression model which reflect the cell-type composition of each sample. Cell-type composition can be ascertained through three main approaches: Direct cell count data; Estimating cell counts using the experimental data and models fitted on DNA methylation data from reference panels of known cell-type composition [272]; or reference free approaches which make use of mathematical methods to identify sources of confounding variation such as cell-type heterogeneity. Whilst there are several reference samples available for cord blood [273–276], none were available for cord tissue at the time of the initial analysis, and consequently a reference free technique was used in phase I of the MAVIDOS analysis. In the analysis of the SWS cord blood data the "andrews and bakulski cord blood" reference panel [275] was used in `meffil` for cell-type composition estimation. Models were fitted using Surrogate Variable Analysis (SVA) [277] and independent Surrogate Variable Analysis (iSVA) [278]. The focus of these results is on SVA as minimal differences between the two methods were observed and SVA was recommended based on comparisons of the performance of various cell-type heterogeneity correction methods [279,280].

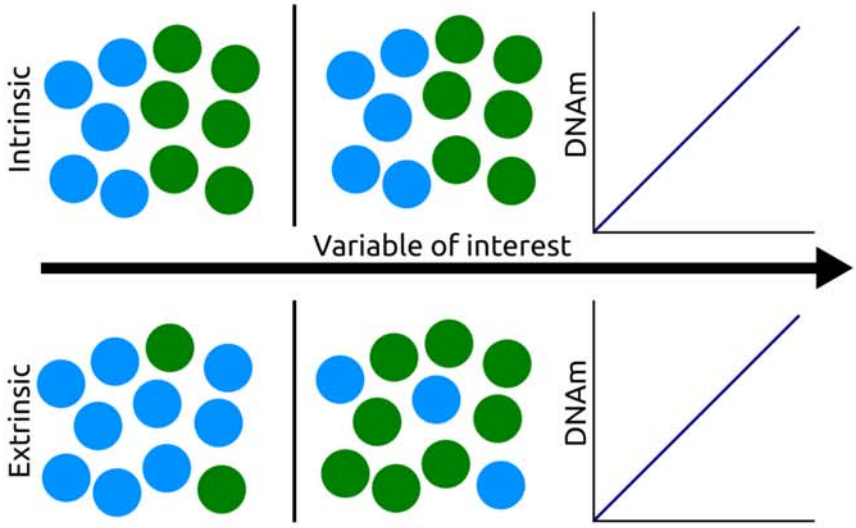

Figure 3.3: Diagrammatic representation of DNAm change arising from extrinsic or intrinsic changes in DNA methylation. Extrinsic changes are due to shifts in cell-type composition. Intrinsic changes in DNA methylation occur without changes in the proportions of cell-types. These two modes of change are of course not mutually exclusive and both can be occurring.

A reference panel for cord tissue samples was recently published [281]. This cell-type reference had not yet been integrated into the `meffil` R package used to perform the EWAS analyses so I created a fork of `meffil` including this reference panel in the required format, this can be installed directly from github. The code I used to add this reference to the package data before building the R package can be found in this gist. Cell-type correction based on this reference was used in phase II of the MAVIDOS analysis.

3.4.4.2 Structure of models fitted for each EWAS

By default EWAS in `meffil` are run with four different models:

1. No covariates, attempting to predict methylation with the variable of interest.
2. All covariates*, attempting to predict methylation with the variable of interest plus a user-supplied list of covariates.
3. Surrogate Variables + all covariates, attempting to predict methylation with user-supplied covariates and surrogate variables generated from SVA.
4. Independent Surrogate Variables + all covariates, attempting to predict methylation with user-supplied covariates and independent surrogate variables generated from iSVA.

*All covariates varies depending on the specific model and is detailed in the individual sections detailing each EWAS. Running EWAS with multiple models permits the effects of adding the

Table 3.1: Summary statistics for variables used in the SWS EWAS. Blood cell count information is estimated based on the methylation data as described in the methods.

Description	mean	median	min	max
Woman's Age at Time of Birth	30.67	31.40	-58.30	41.95
EP: Woman's BMI	26.28	25.04	17.54	41.68
True if Yes in Late (dcrsmok) or early (acrsmok) pregnancy , NA if NA in both, FALSE if both No or No & NA	0.08	0.00	0.00	1.00
Obst: Gestational age (weeks)	40.16	40.29	34.57	42.57
B cells	0.03	0.03	0.01	0.11
CD4+ T-cells	0.01	0.00	0.00	0.07
CD8+ T-cells	0.01	0.01	0.00	0.08
Granulocytes	0.06	0.04	0.00	0.51
Monocytes	0.01	0.00	0.00	0.17
Natural Killer Cells	0.02	0.02	0.00	0.11
Endothelial	0.13	0.13	0.05	0.20
Epithelial	0.17	0.16	0.02	0.42
Stromal	0.58	0.62	0.14	0.86
cam: 34 week VitD	54.54	55.03	10.70	119.00
cam: 11 week VitD	44.83	44.95	14.40	90.20
4yr DXA: Subtotal BMC (g)	356.26	352.52	237.30	455.94
4yr DXA: Subtotal BMD (g/cm sq)	0.47	0.47	0.38	0.57

various covariates on the results of the analysis to be seen.

3.4.5 Concordance of EPIC and 450k EWAS Results

In order to ascertain if the results from the EWAS in the 450k and EPIC arrays were producing similar sets of probes in the top ranking positions when ordered by p-value, the concordance (% overlap) between the top k probes, where $k = 1..100,000$ was calculated. Only probes in common between the two arrays were used and k was incremented in steps of 50.

3.5 Results

3.5.1 Combined MAVIDOS Summary data

Summary statistics for the variables used in the MAVIDOS EWAS models are provided in tables 3.1 & 3.2.

3.5.2 MAVIDOS phase I

DNA methylation at none of the CpGs were significantly associated with any of the four variables of interest for each EWAS performed in either the EPIC or 450k datasets. The concordance between the probes with the top-ranked p-values in common between the EPIC and 450k data was at the level of chance.

3.5.2.1 Whole array QC

Table 3.2: Summary statistics for variables used in the MAVIDOS EWAS divided by the sex of the child. Blood cell count information is estimated based on the methylation data as described in the methods. n=375, Female n=170, Male n=205.

Sex	Description	mean	median	min	max
F	Woman's Age at Time of Birth	31.14	31.36	19.25	41.78
F	EP: Woman's BMI	26.41	24.85	18.80	41.16
F	True if Yes in Late (dcrsmok) or early (acrsmok) pregnancy , NA if NA in both, FALSE if both No or No & NA	0.11	0.00	0.00	1.00
F	Obst: Gestational age (weeks)	40.08	40.29	34.57	42.14
F	B cells	0.03	0.03	0.01	0.07
F	CD4+ T-cells	0.01	0.00	0.00	0.07
F	CD8+ T-cells	0.01	0.01	0.00	0.07
F	Granulocytes	0.07	0.04	0.00	0.51
F	Monocytes	0.02	0.00	0.00	0.17
F	Natural Killer Cells	0.02	0.02	0.00	0.06
F	Endothelial	0.13	0.13	0.07	0.19
F	Epithelial	0.17	0.17	0.03	0.37
F	Stromal	0.56	0.61	0.14	0.79
F	cam: 34 week VitD	53.87	53.58	11.50	114.50
F	cam: 11 week VitD	45.35	45.53	14.55	79.60
F	4yr DXA: Subtotal BMC (g)	355.30	353.05	247.53	439.21
F	4yr DXA: Subtotal BMD (g/cm sq)	0.46	0.46	0.38	0.54
M	Woman's Age at Time of Birth	30.28	31.41	-58.30	41.95
M	EP: Woman's BMI	26.18	25.14	17.54	41.68
M	True if Yes in Late (dcrsmok) or early (acrsmok) pregnancy , NA if NA in both, FALSE if both No or No & NA	0.05	0.00	0.00	1.00
M	Obst: Gestational age (weeks)	40.22	40.43	35.86	42.57
M	B cells	0.03	0.03	0.01	0.11
M	CD4+ T-cells	0.01	0.00	0.00	0.07
M	CD8+ T-cells	0.01	0.01	0.00	0.08
M	Granulocytes	0.05	0.04	0.00	0.39
M	Monocytes	0.01	0.00	0.00	0.16
M	Natural Killer Cells	0.02	0.02	0.00	0.11
M	Endothelial	0.13	0.13	0.05	0.20
M	Epithelial	0.16	0.16	0.02	0.42
M	Stromal	0.60	0.64	0.14	0.86
M	cam: 34 week VitD	55.07	56.88	10.70	119.00
M	cam: 11 week VitD	44.39	44.10	14.40	90.20
M	4yr DXA: Subtotal BMC (g)	357.01	352.52	237.30	455.94
M	4yr DXA: Subtotal BMD (g/cm sq)	0.48	0.48	0.39	0.57

3.5.2.1.1 EPIC arrays The predicted sex (performed with `meffil`) of the samples generated using sex chromosome probe intensities was checked against that in the sample annotation and two mismatches were found. These were MAVIDOS IDs 206 and 63 and the associated arrays were excluded from further analysis, (Figure 3.4).

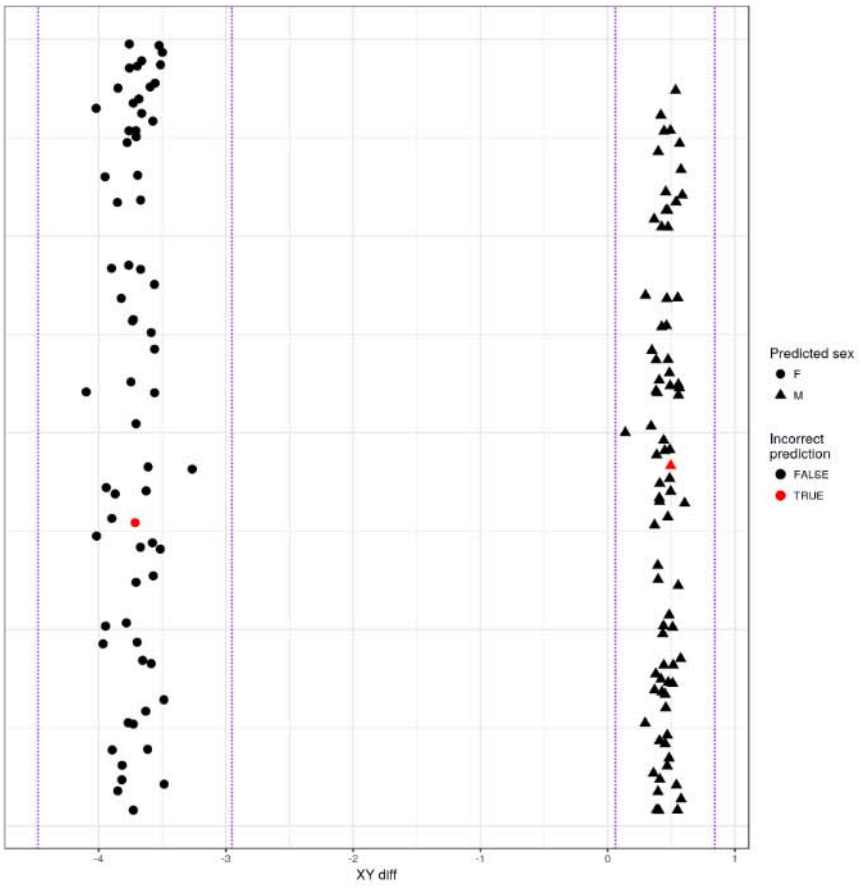

Figure 3.4: Predicted sex of each sample based on the sex chromosome copy numbers inferred from probe intensities for the EPIC array data. Mismatches between the predicted sex and that asserted in the sample annotation metadata are shown in red. Two predicted sex values differ from their annotations. Plot generated by `meffil` QC report.

The dataset also contained four samples for which there were two technical replicates, only the first replicate from each was used (144 arrays run for 140 individuals). Array 201516310023 (MAVIDOS ID 95) was excluded as its median methylated signal was more than 3σ from the expected value, (Figure 3.5). No samples were excluded for having a higher than expected proportion of undetected probes (proportion of probes with detection p-value > 0.01 is > 0.1) (Figure 3.6). No samples were excluded for having a high proportion of probes with low bead counts (proportion of probes with bead number < 3 is > 0.1), (Figure 3.7). In total 3 of the EPIC arrays were excluded from the analysis for failing quality control leaving an n=137.

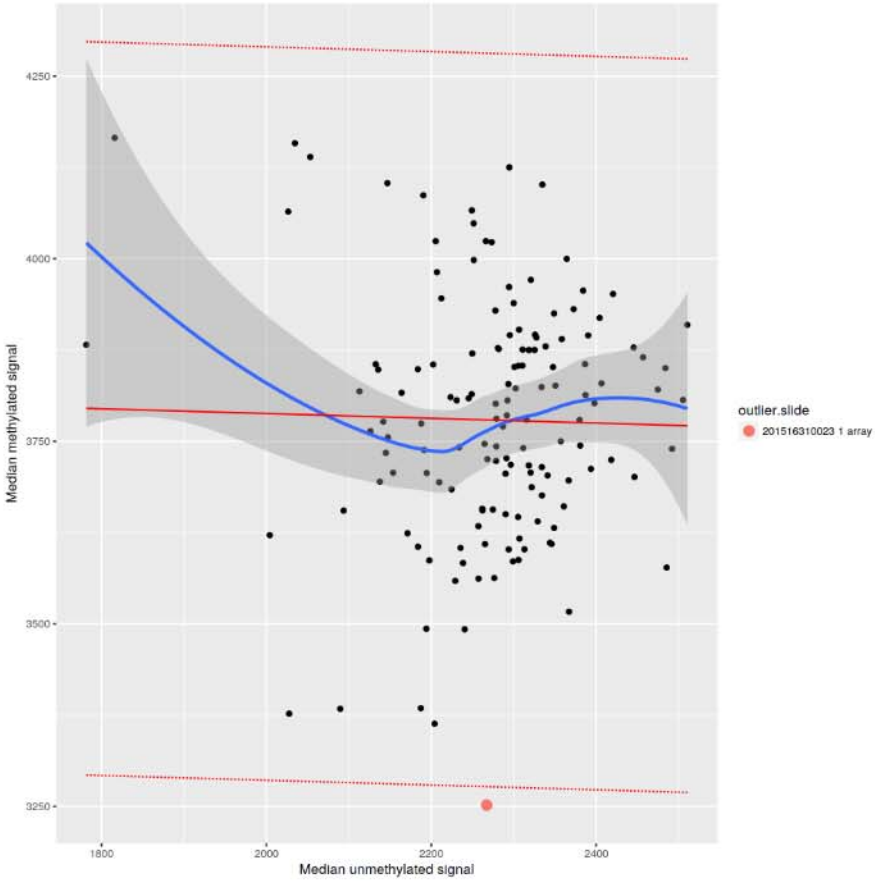

Figure 3.5: Median methylated signal vs unmethylated signal per sample for the EPIC array data, solid red line indicates linear regression of median methylated signal vs median unmethylated signal with dotted red lines representing 3σ from the expected mean. Samples outside the expected range are indicated in the legend. Plot generated by `meffil` QC report.

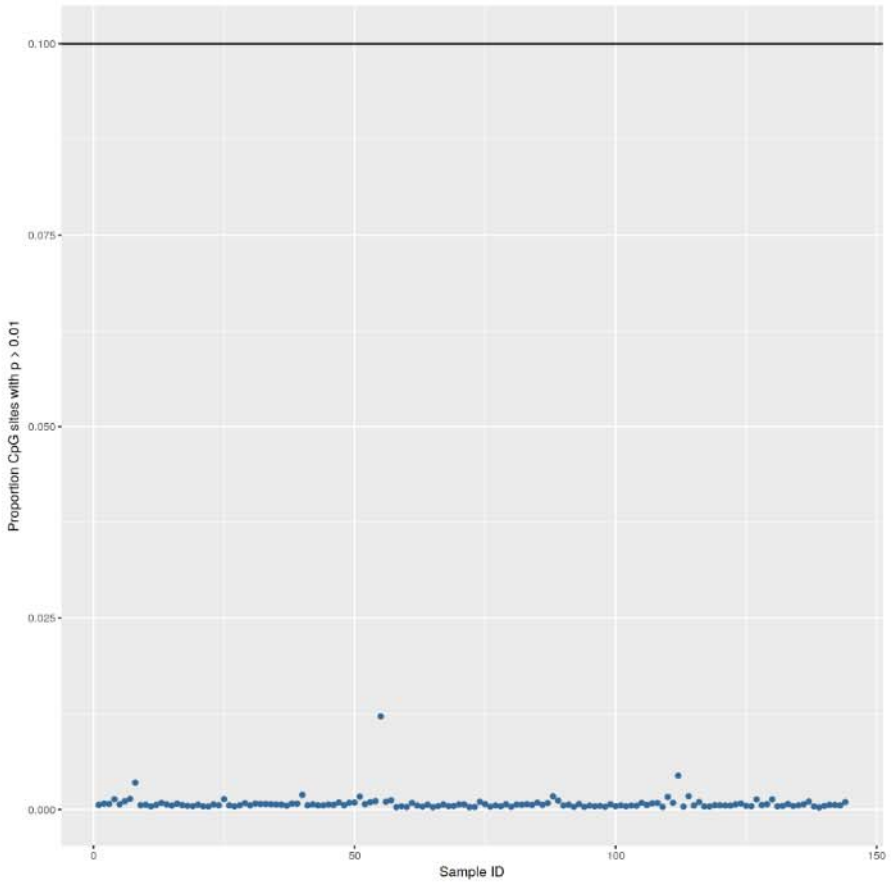

Figure 3.6: Proportion of probes with detection p-values >0.01 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

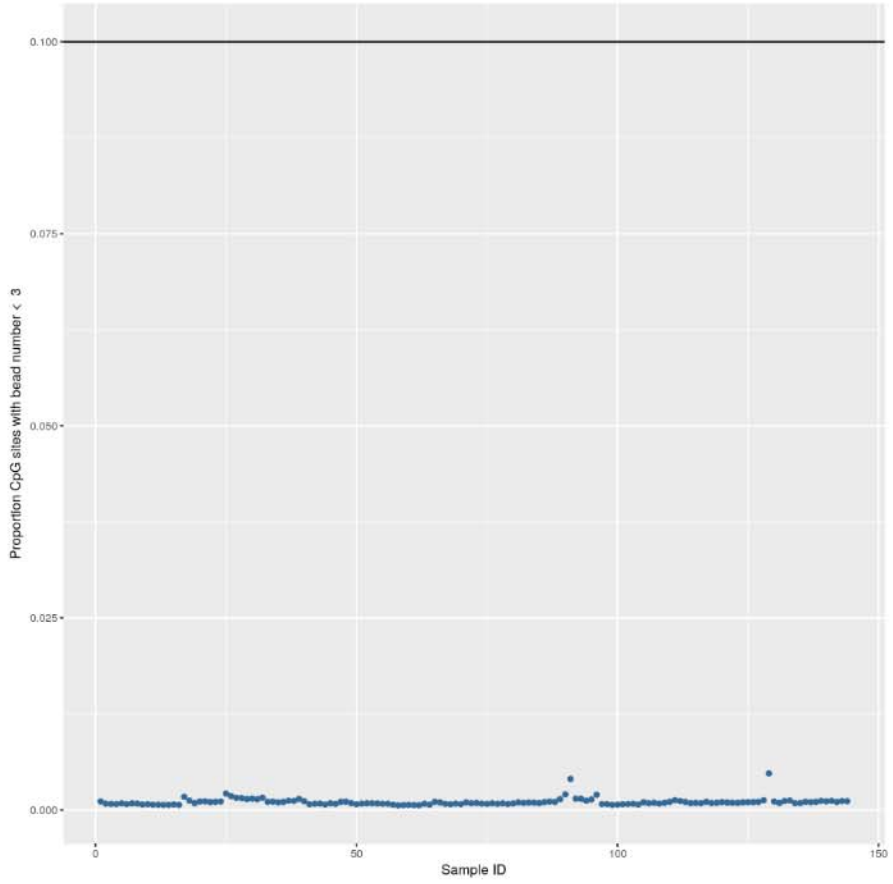

Figure 3.7: Proportion of probes with a bead count of < 3 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

3.5.2.1.2 450k Arrays There were no mismatches between predicted and annotated sex (Figure 3.8). There were not any samples with outliers in their methylated / unmethylated probe proportions (Figure 3.9). No samples were excluded for having a higher than expected proportion of undetected probes (proportion of probes with detection p-value > 0.01 is > 0.1), (Figure 3.10). No samples were excluded for having a high proportion of probes with low bead counts (proportion of probes with bead number < 3 is > 0.1), (Figure 3.11).

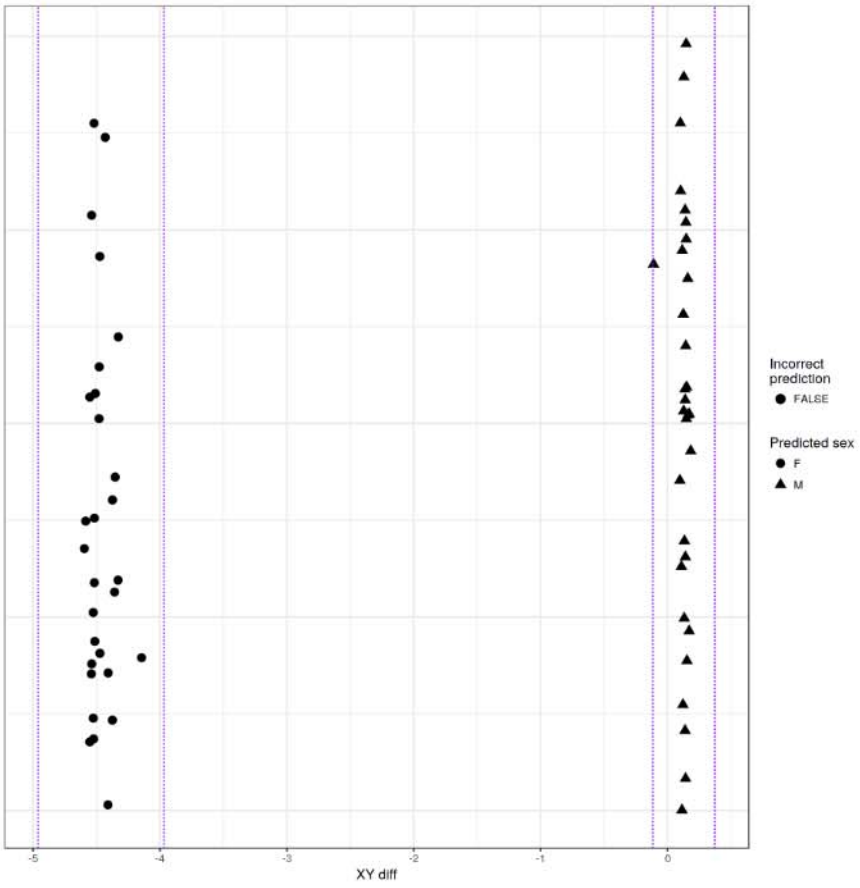

Figure 3.8: Predicted sex of each sample based on the sex chromosome copy numbers inferred from probe intensities for the 450k array data. No predicted sex values differ from their annotations. Plot generated by `meffil` QC report.

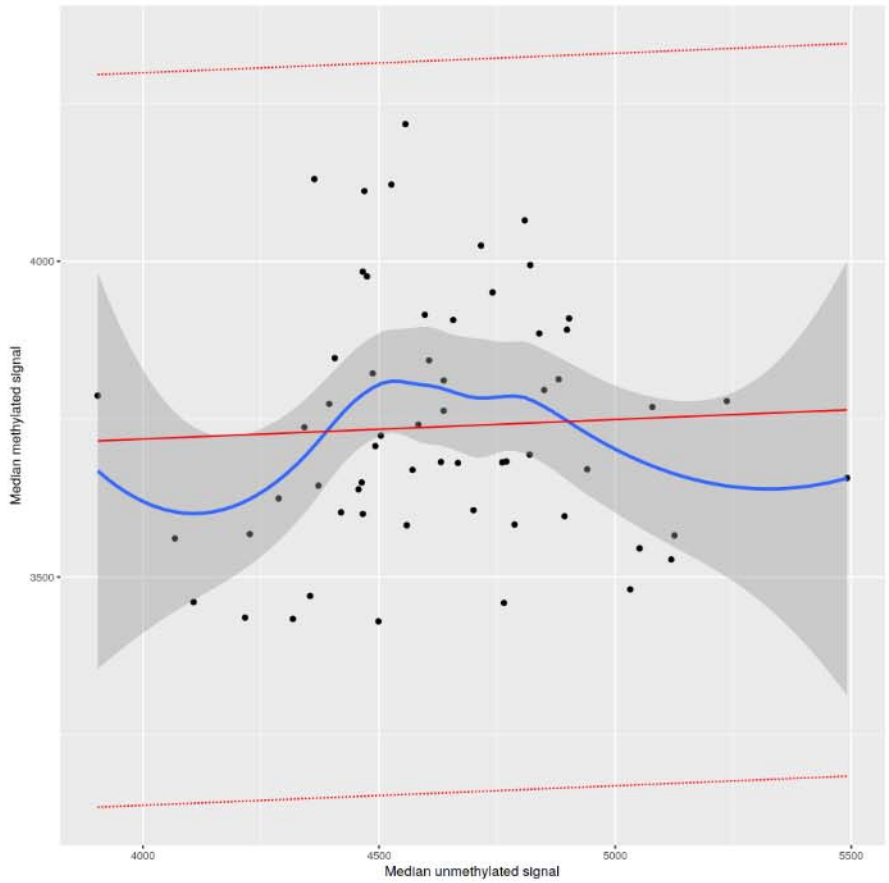

Figure 3.9: Median methylated signal vs unmethylated signal per sample for the 450k array data, solid red line indicates linear regression of median methylated signal vs median unmethylated signal with dotted red lines representing 3σ from the expected mean. Samples outside the expected range would be indicated in the legend. Plot generated by `meffil` QC report.

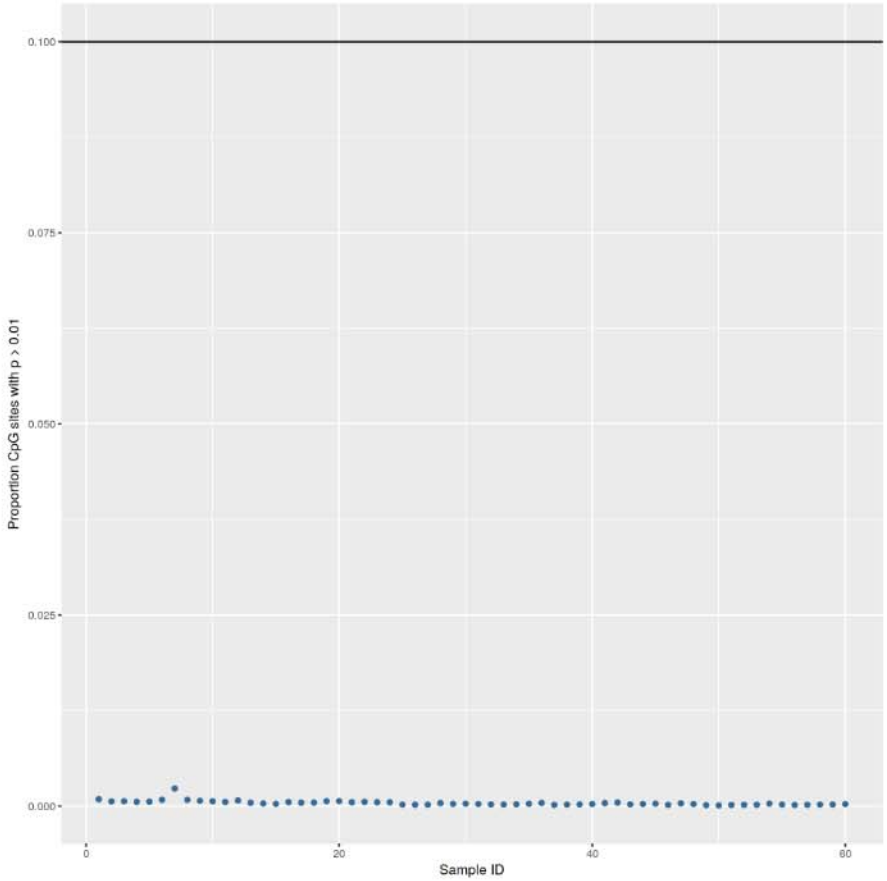

Figure 3.10: Proportion of probes with detection p-values >0.01 by sample for the 450k array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

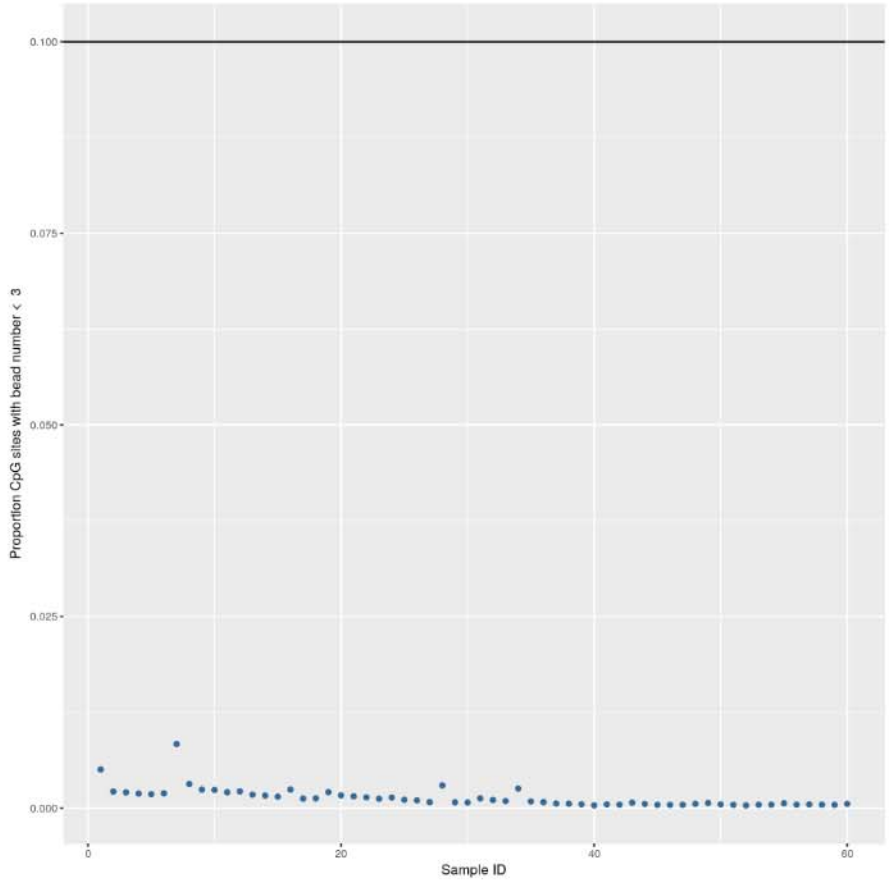

Figure 3.11: Proportion of probes with a bead count of < 3 by sample for the 450k array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

3.5.2.2 Probe QC

3.5.2.2.1 Probe QC - EPIC Arrays There were no outliers within the control probes Figure 3.12. 1,626 probes were excluded for having high background signal in a high proportion of samples (proportion of samples with detection p-value > 0.01 is > 0.1), (Figure 3.13). 162 probes were excluded for having low bead count in a high proportion of samples (proportion of samples with bead number < 3 is > 0.1), (Figure 3.14). Probes with poor technical quality were excluded from the analysis prior to functional normalisation.

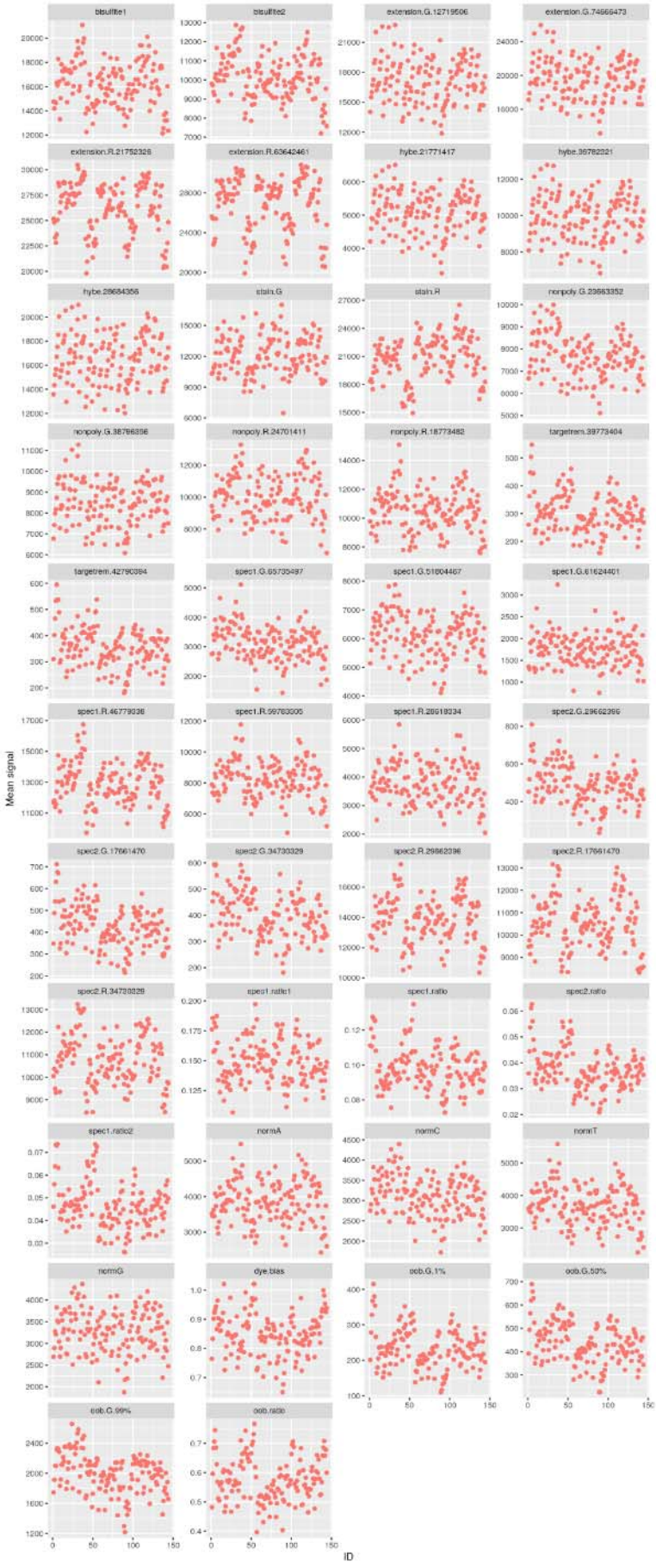

Figure 3.12: Control probe signal by sample for each summary group for the EPIC data. Outliers would be circled in black. Plot generated by `meffil` QC report.

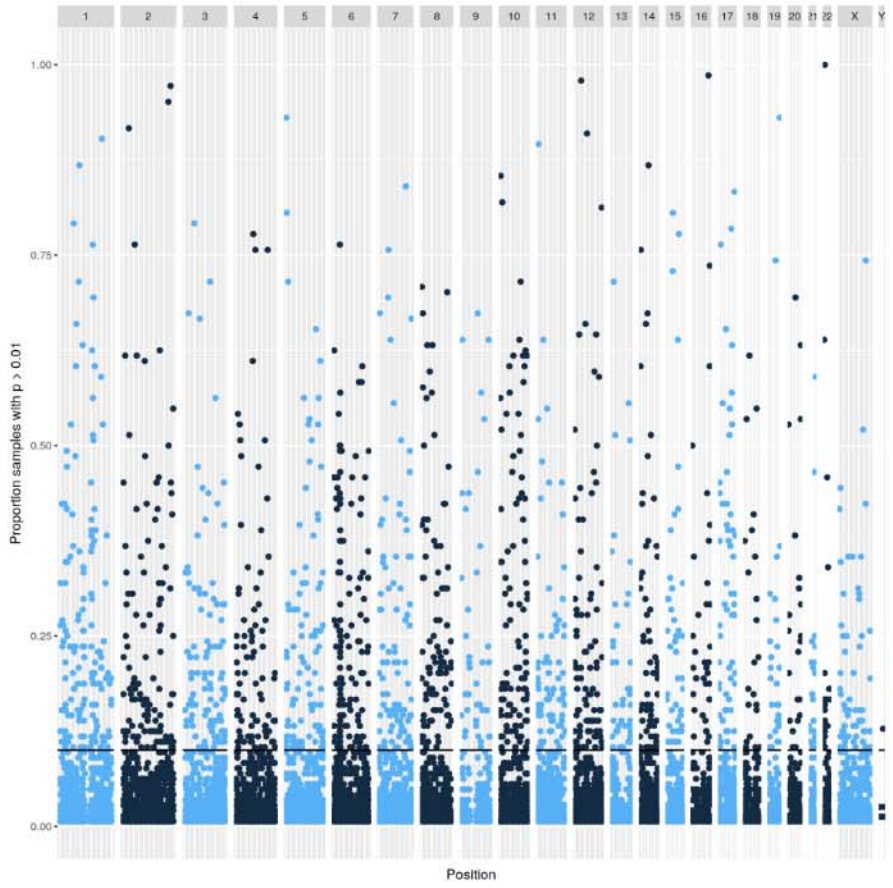

Figure 3.13: Undetectable probes across samples for EPIC data. Manhattan plot showing proportion of samples (y) in which a given probe (x) is not distinguishable from background noise, i.e. a detection p-value of > 0.01. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

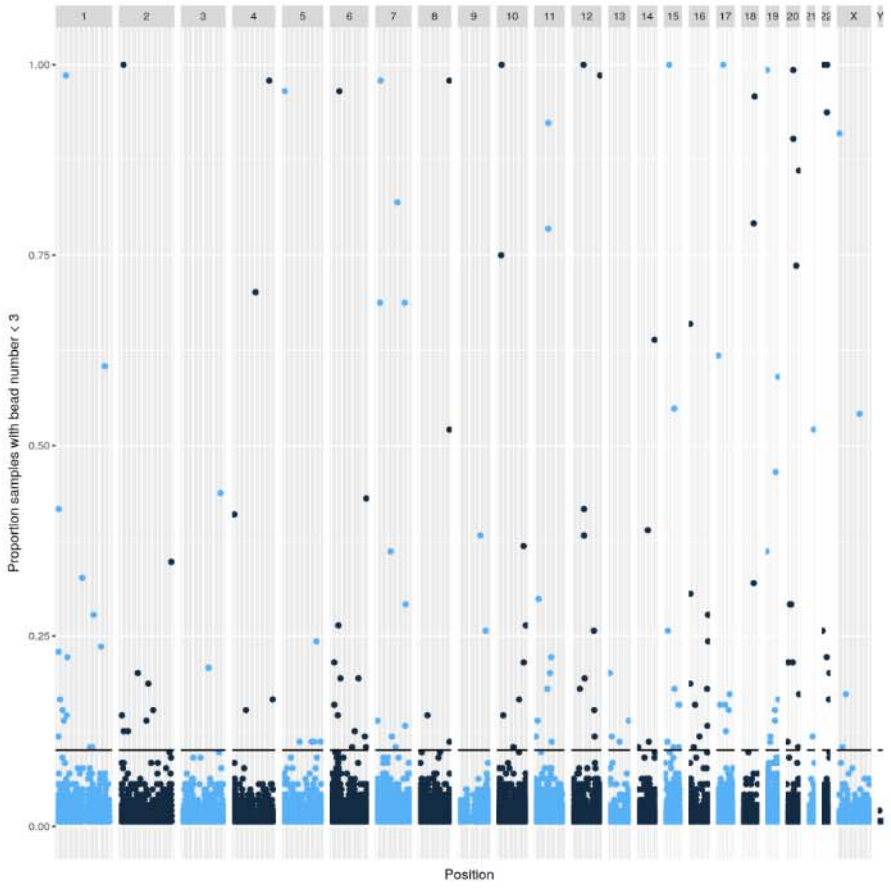

Figure 3.14: Low bead count probes across samples for EPIC data. Manhattan plot showing the proportion of samples (y) in which a given probe (x) has a bead count of <3. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

Problematic probes identified by Pidsley et al. [222] and those overlapping the regions identified by Bell et al. [107] were excluded from subsequent analysis after functional normalisation. This, including the poor quality probes, is a total of 57,396 unique probes excluded from the analysis (~6.62% of the total number of probes). Gap hunter identified a further 77,398 probes (8.9% of all probes) which might be subject to genetic confounding (Figure 3.15).

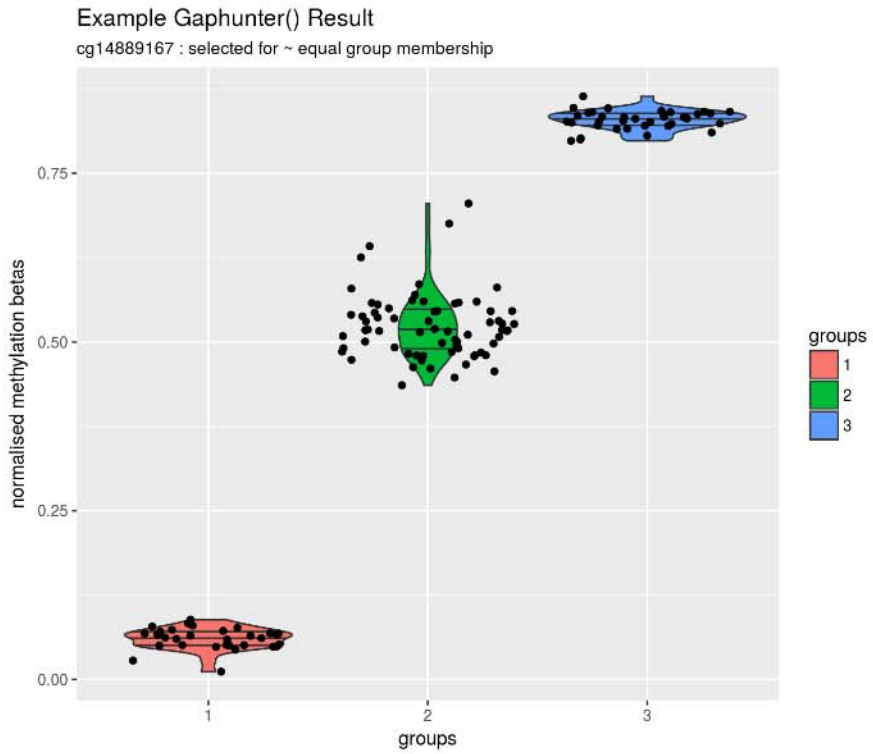

Figure 3.15: An example of the DNAm distribution for a result from the `gaphunter()` function. This is an example chosen to best exemplify the sort of result which is strongly suggestive of a genetic variant with an impact on methylation status acting on this site. It is unrepresentative of typical results from `gaphunter()` in that the groups have a relatively even membership, many results have a small number of individuals in one or more groups making it hard to distinguish methylation outliers caused by rarer genetic variants from those with other causes.

3.5.2.2.2 Probe QC - 450k Arrays 509 probes were excluded for having high background signal in a high proportion of samples (proportion of samples with detection p-value > 0.01 is > 0.1), (Figure 3.17). 1037 probes were excluded for having low bead count in a high proportion of samples (proportion of samples with bead number < 3 is > 0.1) (Figure 3.18). There was one sample (MAVIDOS ID 2183) with an outlier within the control probes, a dinitrophenyl labelled staining control probe, thus it was not excluded as only outliers in dye bias and bisulfite conversion control probes were deemed sufficient grounds for excluding a sample (Figure 3.16, Detailed contol probe desciption [221] p222).

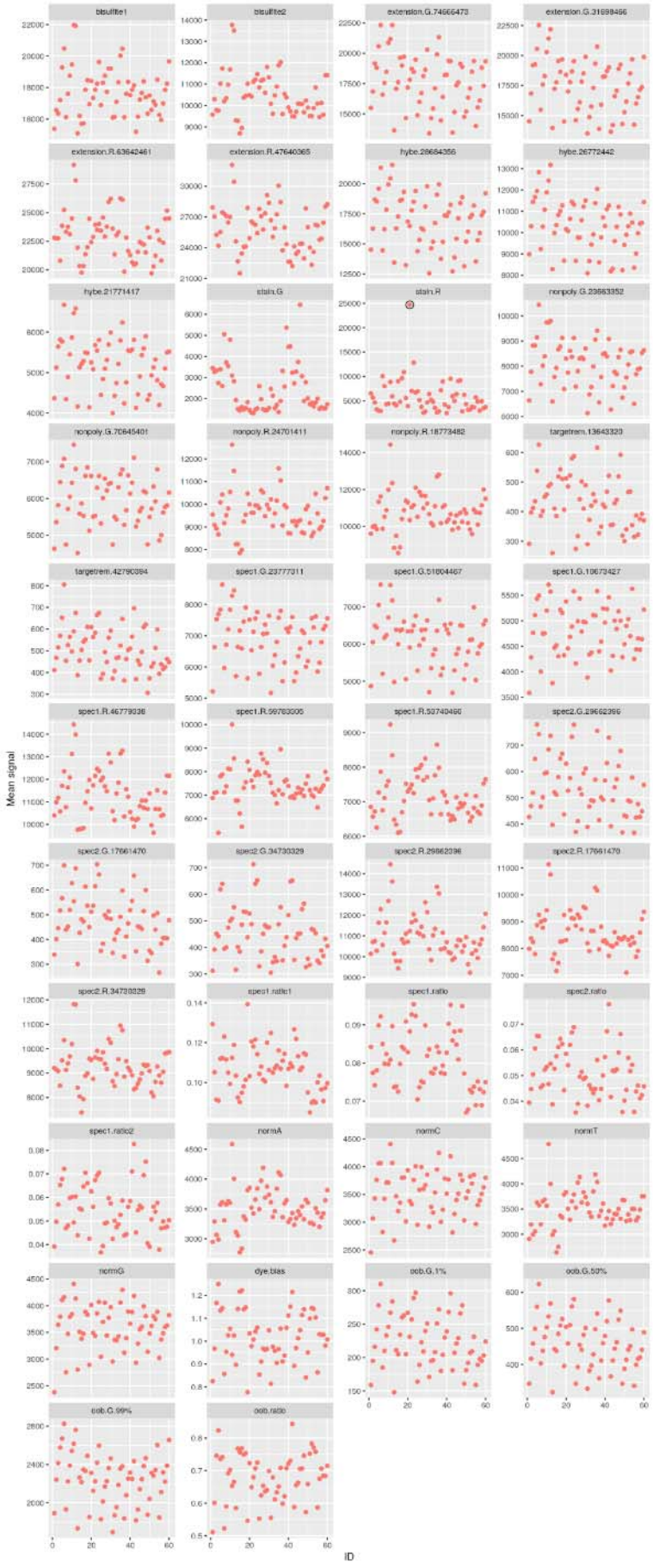

Figure 3.16: Control probe signal by sample for each summary group for the 450k data. Outliers would be circled in black. Plot generated by `meffil` QC report.

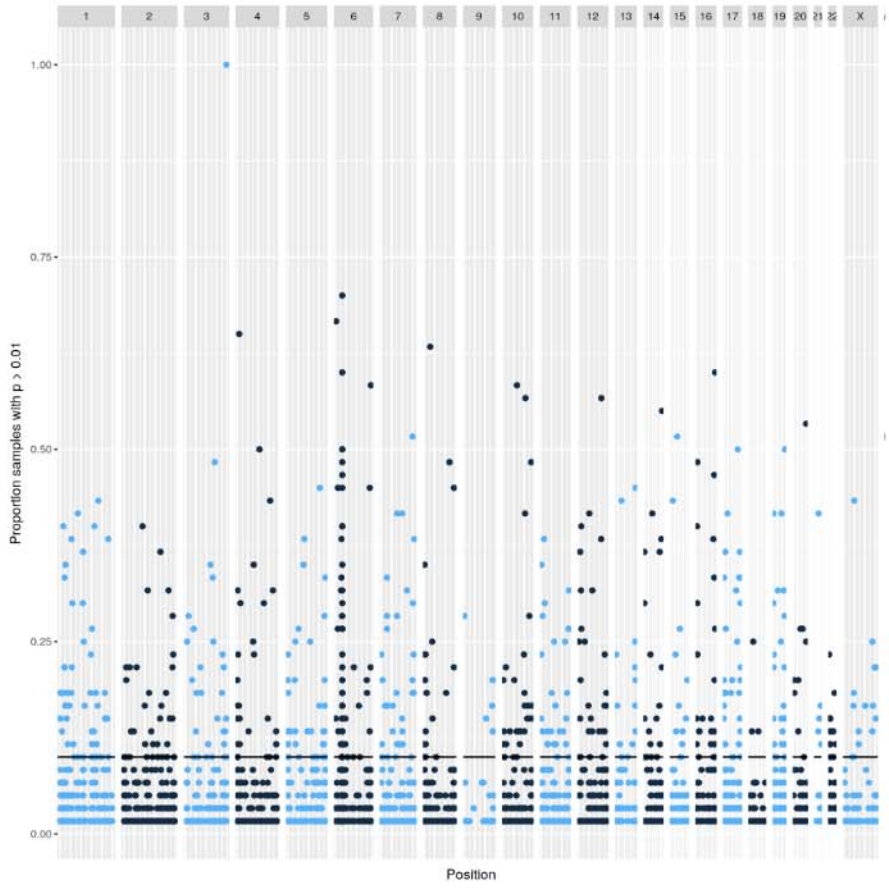

Figure 3.17: Undetectable probes across samples for 450k data. Manhattan plot showing proportion of samples (y) in which a given probe (x) is not distinguishable from background noise, i.e. a detection p-value of > 0.01. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

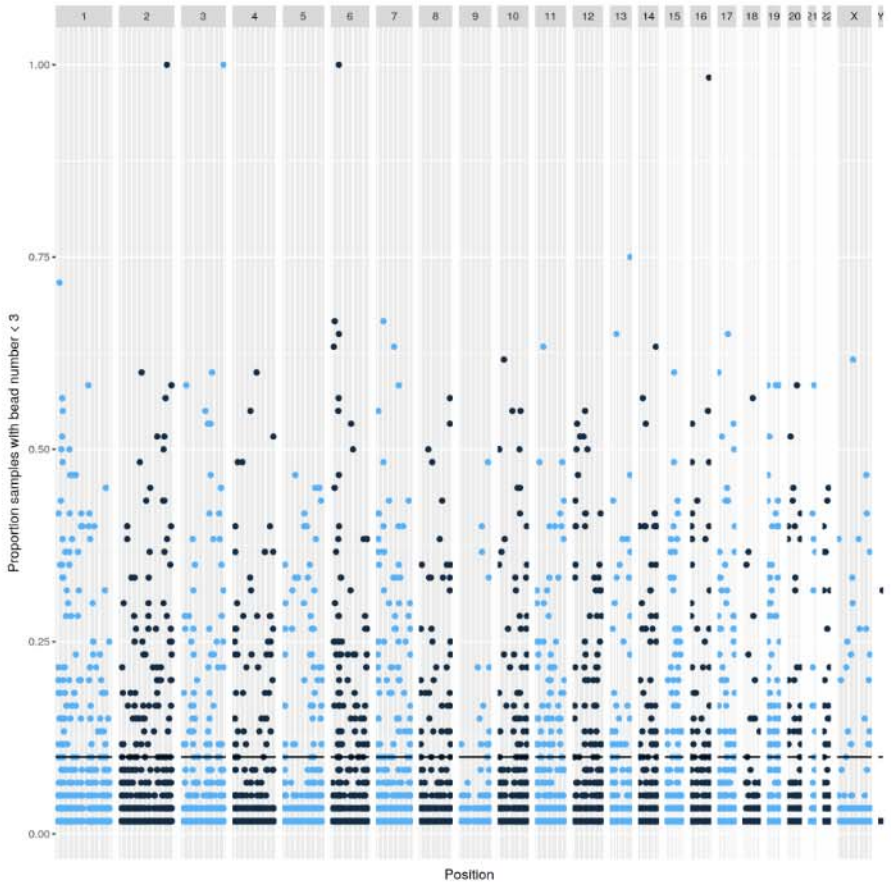

Figure 3.18: Low bead count probes across samples for 450k data. Manhattan plot showing the proportion of samples (y) in which a given probe (x) has a bead count of <3. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

All probes on the 'general mask' list from Zhou et al. [223] were excluded from the analysis following functional normalisation, leaving a total of 418,632 probes for subsequent analysis.

3.5.2.3 Functional Normalisation

An m of 6 was chosen for the EPIC arrays as this value produced the last steep drop in residual variation, see Figure 3.19. An m of 6 was chosen for the 450k arrays as this value produced the last steep drop in residual variation, see Figure 3.20.

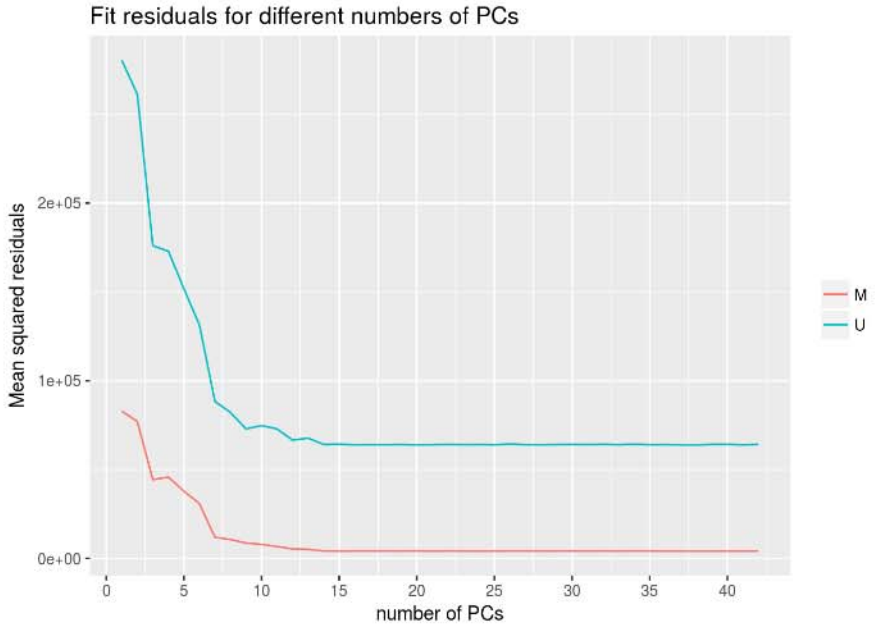

Figure 3.19: Residual variation remaining after functional normalisation of the top 20,000 most variable probes with m PCs from the control probe summary matrices for the EPIC array samples (n=137), for M = methylated and U = unmethylated probes.

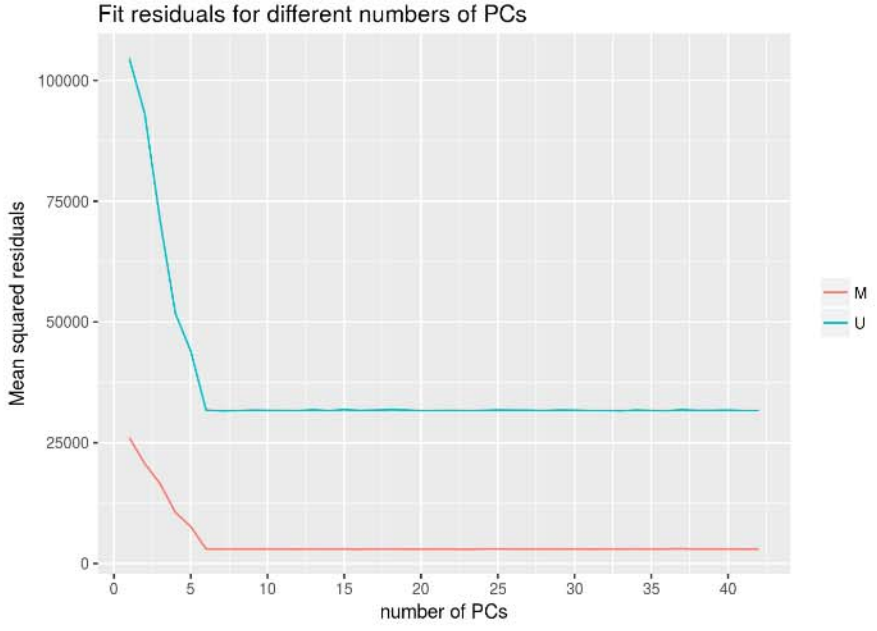

Figure 3.20: Residual variation remaining after functional normalisation of the top 20,000 most variable probes with m PCs from the control probe summary matrices for the 450k array samples (n=60), for M = methylated and U = unmethylated probes.

3.5.2.4 EWASs

All EWAS performed below were also performed in exactly the same fashion for the 60 450k arrays, none of these results were significant and they are not included here.

3.5.2.4.1 Neonatal Bone Mineral Content Figure 3.21 illustrates the distribution of Neonatal Bone Mineral Content, the outcome on which this EWAS was performed.

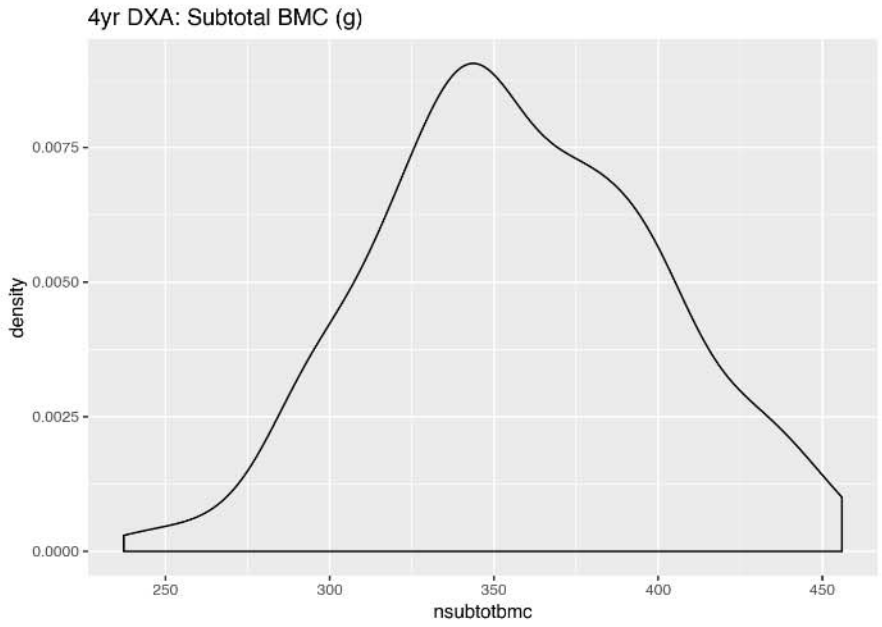

Figure 3.21: Distribution of Neonatal Bone Mineral Content (g) for individuals in the EWAS.

No CpGs fell below the Bonferroni corrected significance threshold for an association between DNA methylation at that locus and neonatal bone mineral content, Figure 3.22. Sex and sample age at DXA were included as covariates in the 'all' model. SVA generated 5 significant surrogate variables which were additionally used in the SVA model.

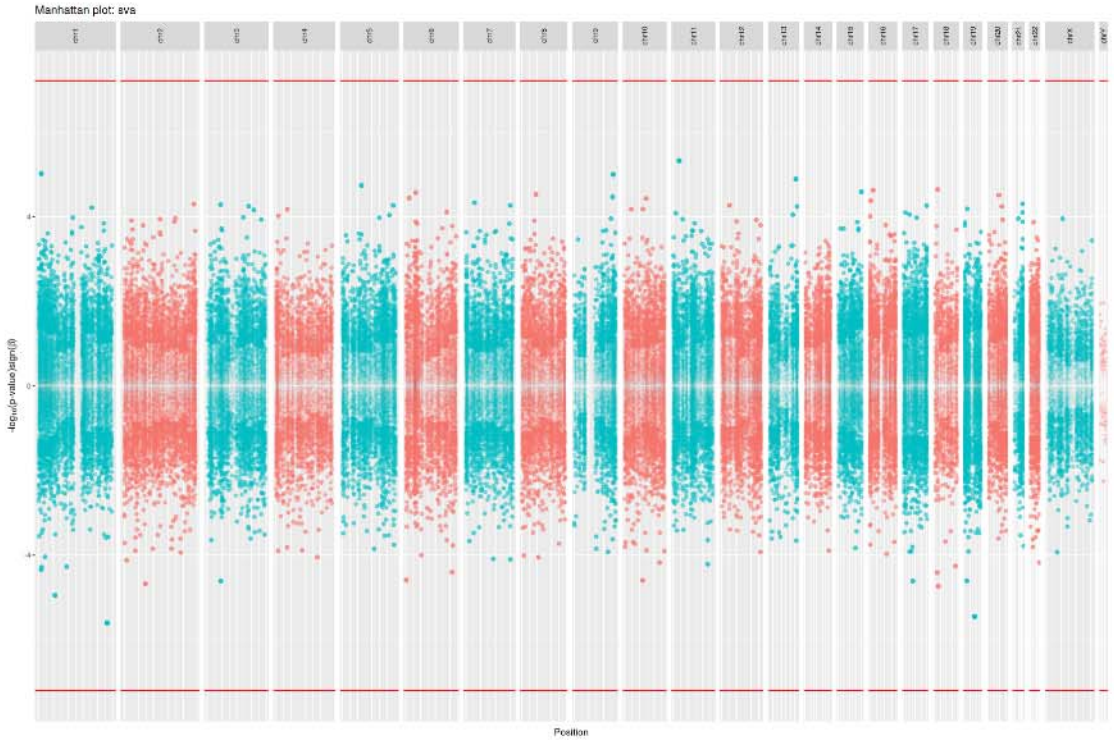

Figure 3.22: Results of EWAS for neonatal bone mineral content with SVA model. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. Size and transparency of points increases with $-log_{10}(p-value)$ such that the most significant CpGs are represented by the largest and least translucent points. x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 6.18×10^{-8} $(0.05 \div 808,585)$.

3.5.2.4.2 Intervention / Placebo No CpGs fell below the Bonferroni corrected significance threshold for an association between DNA methylation at that locus and Intervention/placebo group status, Figure 3.23. Sex and sample age at DXA were included as covariates in the 'all' model. SVA generated 5 significant surrogate variables which were additionally used in the SVA model.

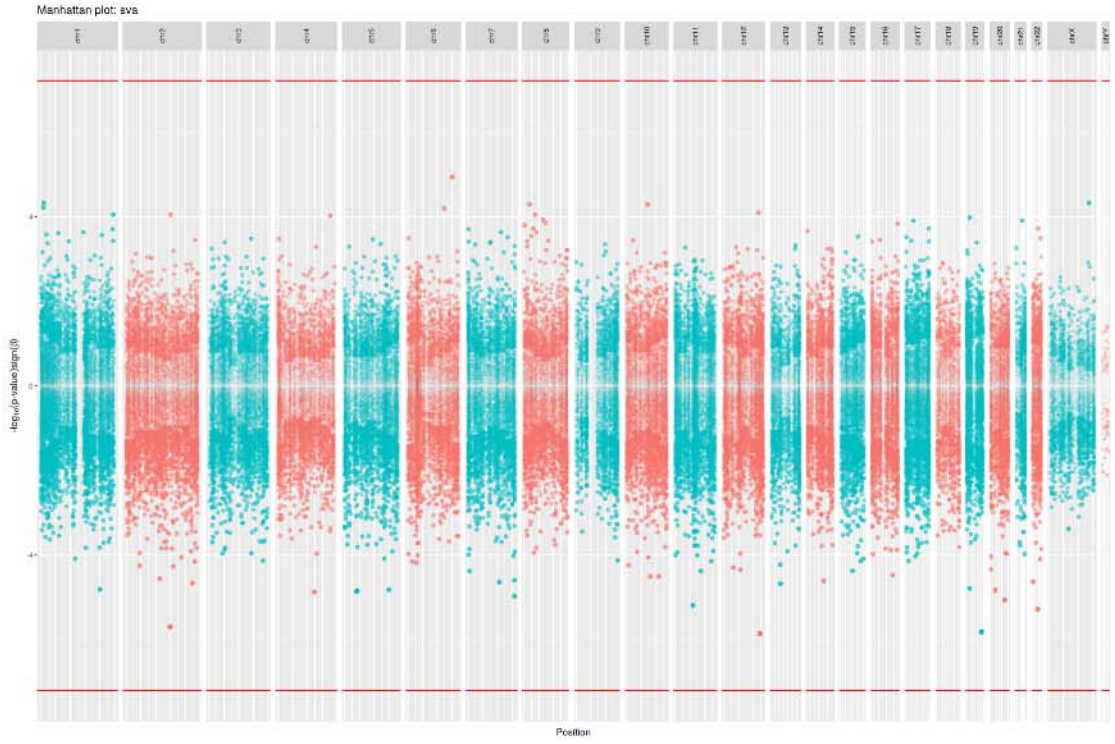

Figure 3.23: Results of EWAS for intervention/placebo group status with SVA model. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. Size and transparency of points increases with $-log_{10}(p-value)$ such that the most significant CpGs are represented by the largest and least translucent points. x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 6.18×10^{-8} $(0.05 \div 808,585)$.

3.5.2.4.3 Maternal Vitamin D (34wks) Maternal Vitamin D levels remain substantially overlapping between intervention and placebo groups at 34wks, see Figure 3.24. Thus maternal vitamin D at 34wks may prove a more useful variable to model than intervention/placebo status.

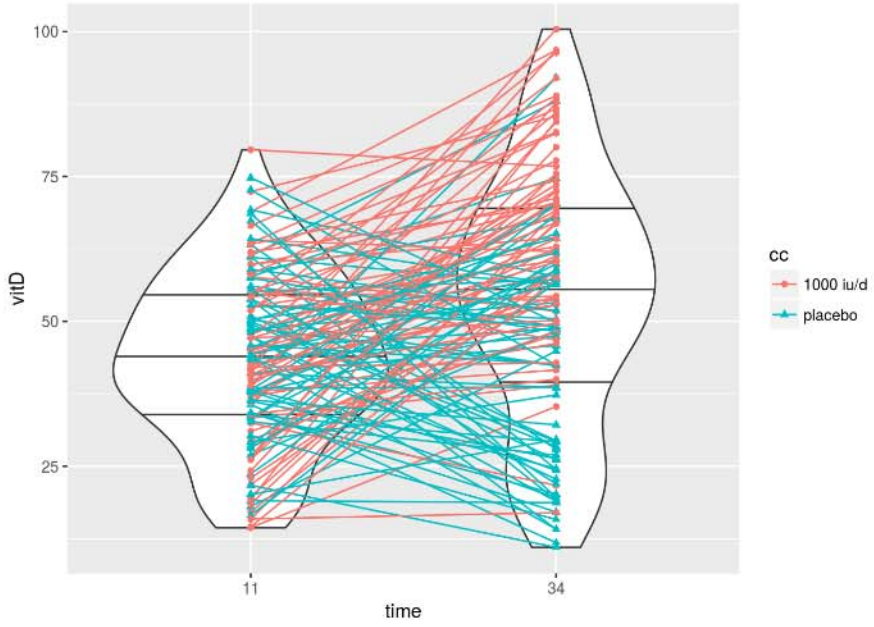

Figure 3.24: Maternal circulating 25(OH)-vitamin D levels (nmol/l) at 11 and 34wks gestation, supplementation with 1000 IU/day cholecalciferol began at week 14. Each participant is shown at both time points linked by a line to indicate the direction of change. The violin plots indicate the density of the distribution of vitamin D values at each time point with the 25^{th}, 50^{th}, 75^{th} quantiles indicated with horizontal black lines. The colour indicates Intervention (Red) / Placebo (Blue) group

No CpGs fell below the Bonferroni corrected significance threshold for an association between DNAm at that locus and maternal vitamin D levels at 34wks gestation, (Figure 3.25). Sex and sample age at DXA were included as covariates in the 'all' model. SVA generated 5 significant surrogate variables which were additionally used in the SVA model.

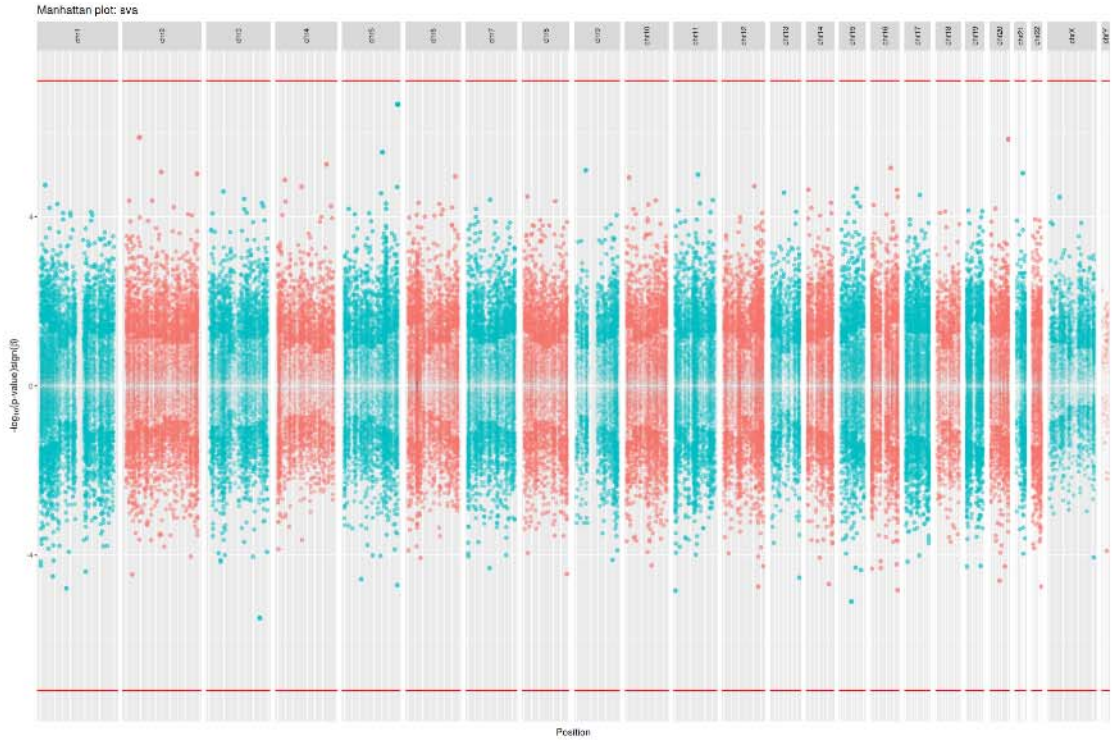

Figure 3.25: Results of EWAS for maternal circulating 25(OH)-vitamin D levels (nmol/l) levels at 34wks gestation with SVA model. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. Size and transparency of points increases with $-log_{10}(p-value)$ such that the most significant CpGss are represented by the largest and least transleucent points. x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 6.18×10^{-8} $(0.05 \div 808,585)$.

3.5.2.4.4 Change in Maternal Vitamin D Figure 3.26 illustrates the change in maternal vitamin D from 11 to 34 weeks gestation. No CpGs fell below the Bonferroni corrected significance threshold for an association between DNAm at that locus and change in maternal vitamin D from 11 to 34wks gestation, Figure 3.27. SVA generated 5 significant surrogate variables which were additionally used in the SVA model.

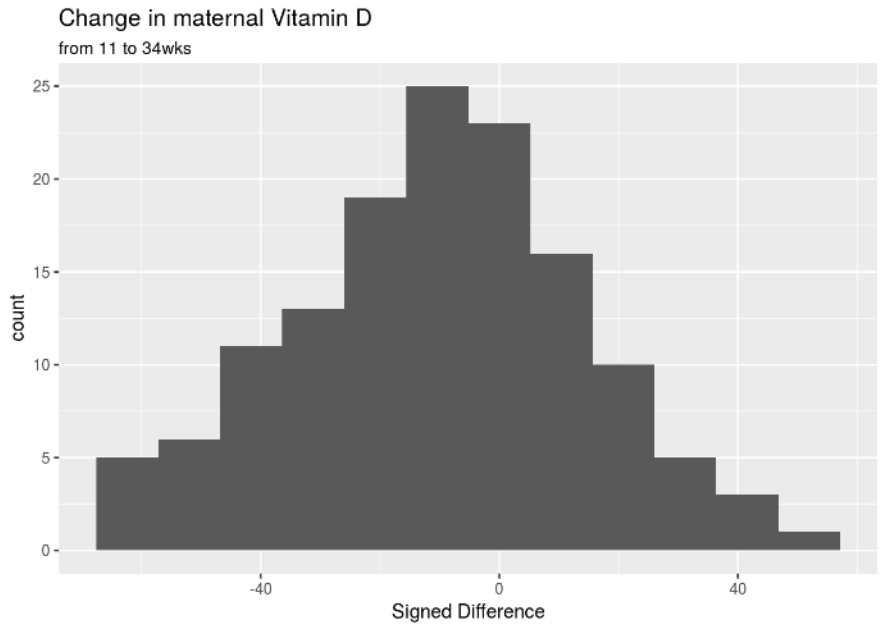

Figure 3.26: Distribution of the changes in maternal circulating 25(OH)-vitamin D levels (nmol/l) levels from 11 to 34 weeks gestation.

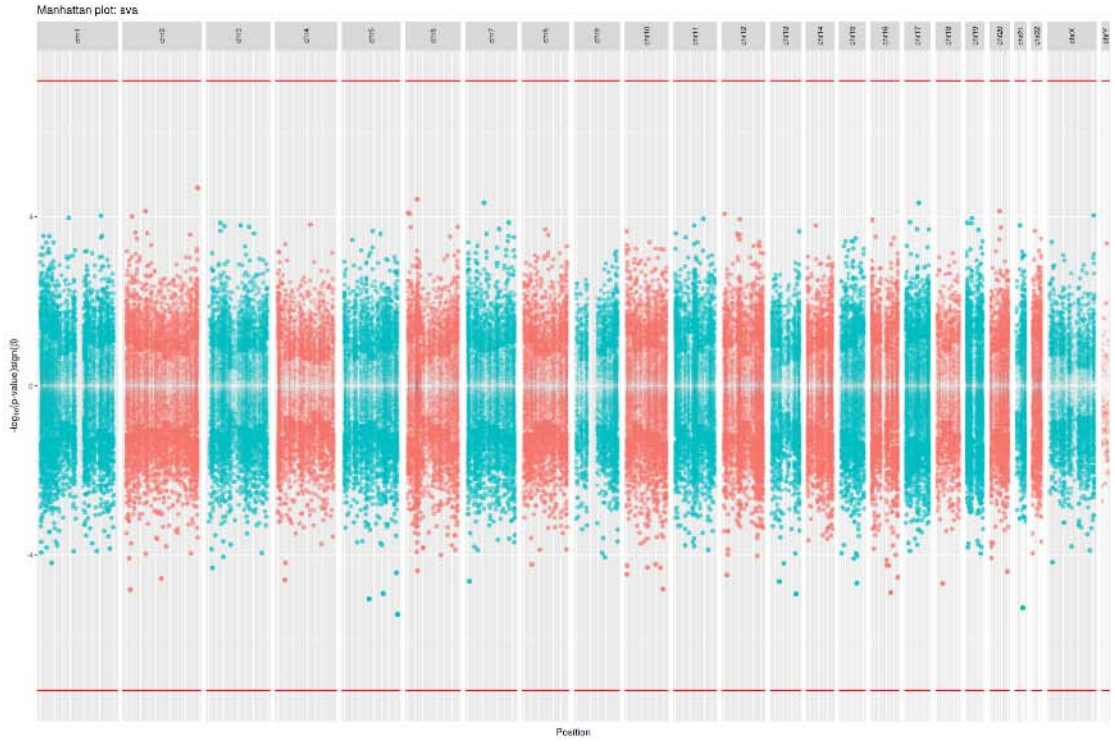

Figure 3.27: Results of EWAS for change in maternal circulating 25(OH)-vitamin D levels (nmol/l) levels from 11 to 34wks gestation with SVA model. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. Size and transparency of points increases with $-log_{10}(p-value)$ such that the most significant CpGs are represented by the largest and least translucent points. x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 6.18×10^{-8} $(0.05 \div 808,585)$.

3.5.2.5 Concordance of EPIC and 450k EWAS results

Concordance between the rankings of CpGss would suggest that the EWAS may be capturing a 'real' signal that is simply below the significance threshold with the statistical sensitivity/power that is available in this dataset. The concordance between the EPIC and 450k datasets (Figures 3.28 & 3.29) appears to be at roughly the level expected by chance. This does not lend support to the possibility that there are associations between the variables of interest and DNAm that are beneath the current sensitivity of the study, it does not, however, rule this out. In the absence of any CpGs above the significance threshold and with poor concordance between the 450k and EPIC array p-value rankings no further analyses such as gene set enrichment and differentially methylated region (DMR) calling have been carried out at this time. Use of the two similar cell-free cell-type correction methods SVA and iSVA for the same dataset produce highly concordant results so are used as an example of how a highly concordant result would appear in the plots of these results.

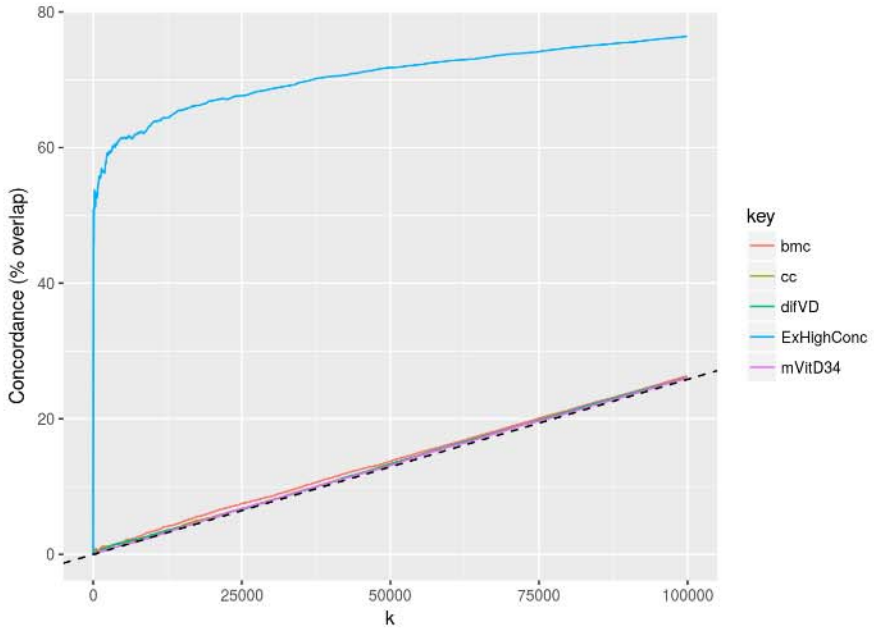

Figure 3.28: Concordance between the top 100,000 probes in common between the EWASs run on the EPIC (n=137) and 450k (n=60) data sets. bmc = bone mineral content, cc = Intervention / Placebo, difVD = Change in Vitamin D from 11 to 34wks, ExHighConc = Example of High Concordance generated using SVA vs iSVA results for the 450k intervention/placebo EWAS. Dotted line denotes concordance expected by chance (intersects 50% at 387,511, the number of shared probes).

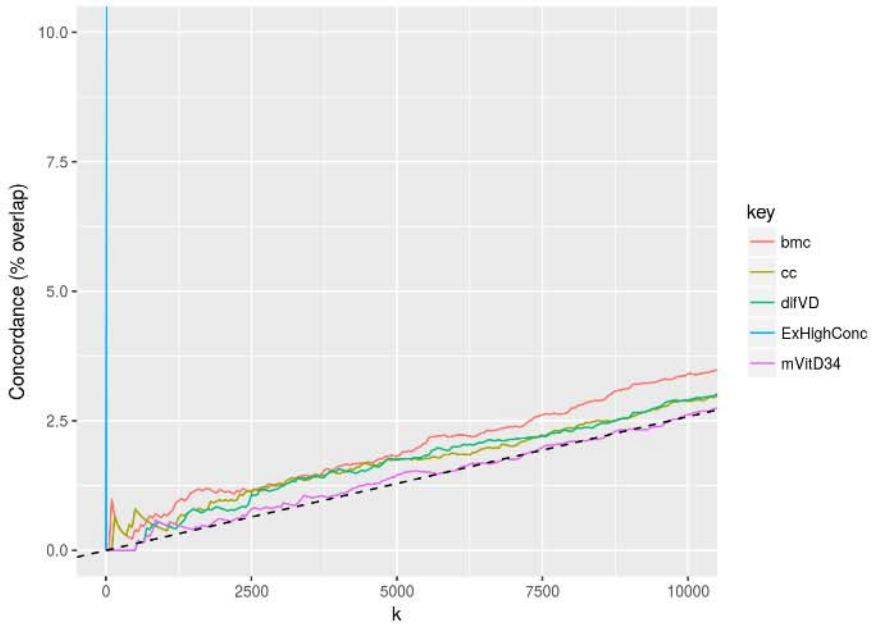

Figure 3.29: Concordance between the top 10,000 probes in common between the EWASs run on the EPIC (n=137) and 405k (n=60) data sets. bmc = bone mineral content, cc = Intervention / Placebo, difVD = Chance in Vitamin D from 11 to 34wks, ExHighConc = Example of High Concordance generated using SVA vs iSVA results for the 450k intervention/placebo EWAS. Dotted line denotes concordance expected by chance (intersects 50% at 387,511, the number of shared probes).

3.5.3 MAVIDOS phase II

DNA methylation at none of the CpGs was significantly associated with bone mineral content at 4 years.

3.5.3.1 Whole Array QC

The predicted sex of the samples generated using sex chromosome probe intensities was checked against that in the sample annotation and 5 mismatches were found (Figure 3.30). Two samples did not have sex information, 2 had predicted sex discordant with their annotated sex and 1 was ambiguous these were excluded from further analysis.

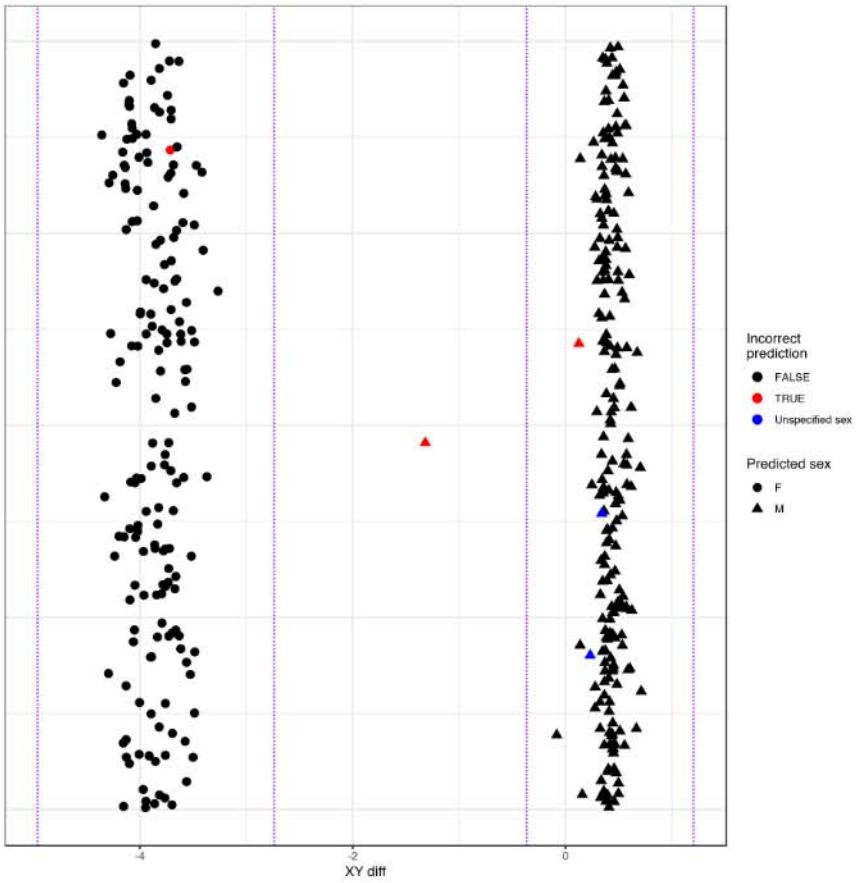

Figure 3.30: Predicted sex of each sample based on the sex chromosome copy numbers inferred from probe intensities for the EPIC array data. Mismatches between the predicted sex and that asserted in the sample annotation metadata are shown in red. Two predicted sex values differ from their annotations. Plot generated by `meffil` QC report.

Arrays: 201516310023 (mavid: 1490), 201516320022 (mavid: 1672), 201530430013 (mavvid: 4090), 201530430015 (mavid: 1903), & 202410280028 (mavid: 2078) were excluded as their median methylated signal was more than 3σ from the expected value, (Figure 3.31). No samples were excluded for having a higher than expected proportion of undetected probes (proportion of probes with detection p-value > 0.01 is > 0.1) (Figure 3.32). No samples were excluded for having a high proportion of probes with low bead counts (proportion of probes with bead number < 3 is > 0.1), (Figure 3.7).

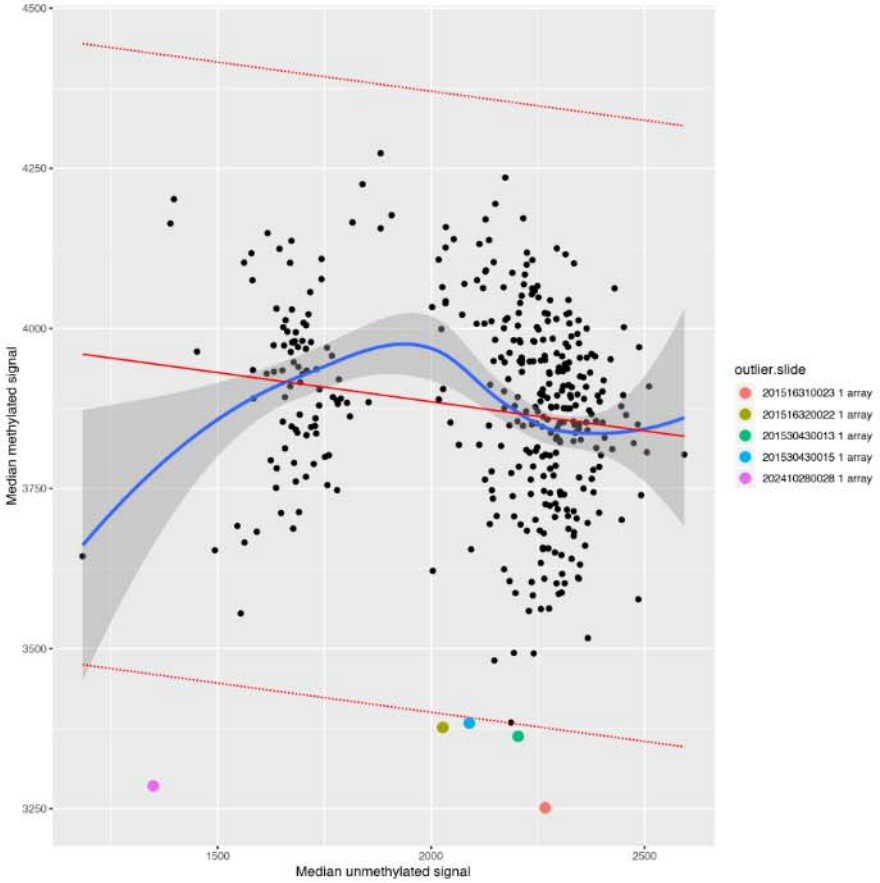

Figure 3.31: Median methylated signal vs unmethylated signal per sample for the EPIC array data, solid red line indicates linear regression of median methylated signal vs median unmethylated signal with dotted red lines representing 3σ from the expected mean. Samples outside the expected range are indicated in the legend. Plot generated by `meffil` QC report.

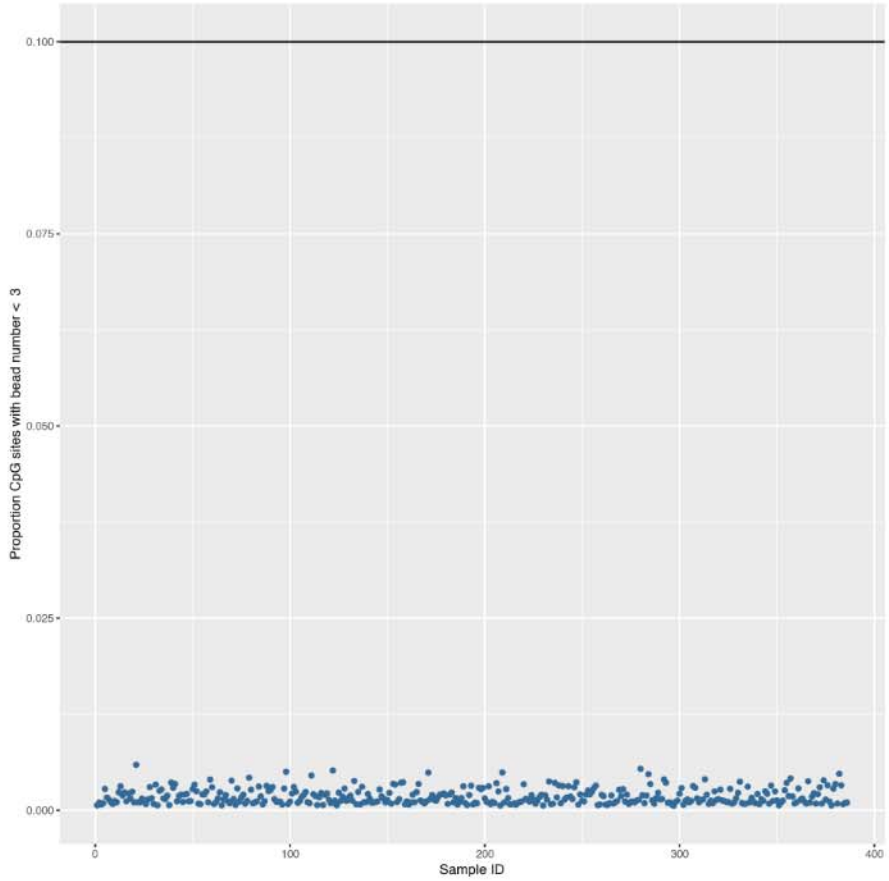

Figure 3.32: Proportion of probes with detection p-values >0.01 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

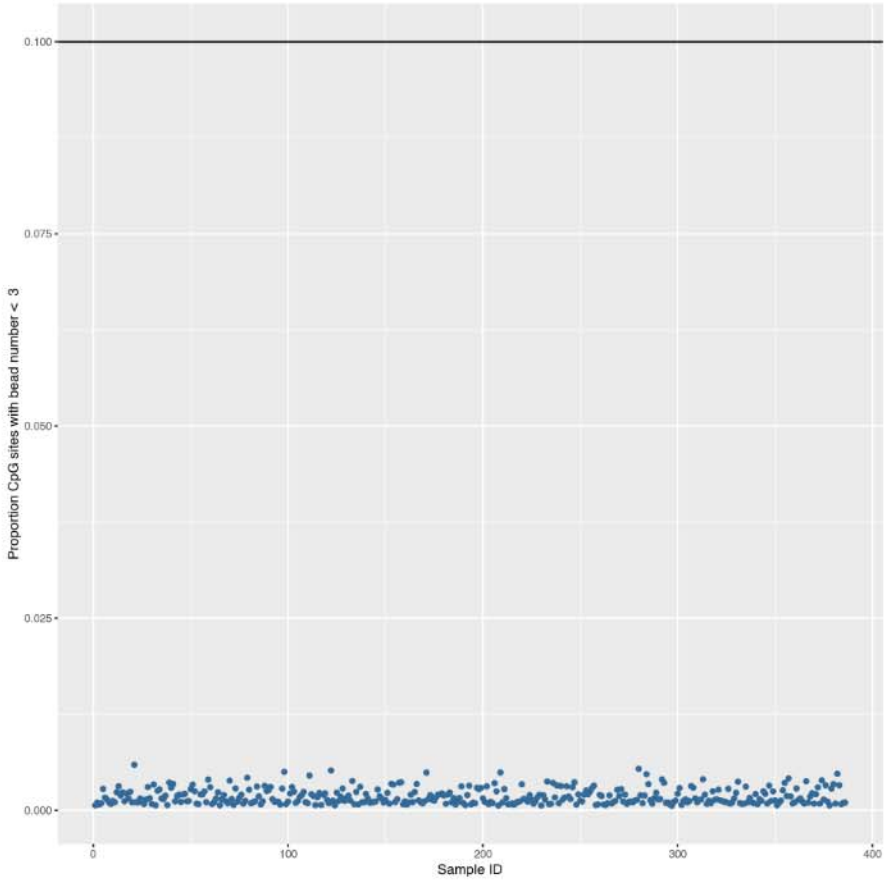

Figure 3.33: Proportion of probes with a bead count of < 3 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

3.5.3.2 Probe QC

There was one outlier within the control probes, in a non-critical specificity control probe for detecting non-specific methylation detection over an unmethylated background (Figure 3.12). 1,317 probes were excluded for having high background signal in a high proportion of samples (proportion of samples with detection p-value > 0.01 is > 0.1), (Figure 3.13). 220 probes were excluded for having low bead count in a high proportion of samples (proportion of samples with bead number < 3 is > 0.1), (Figure 3.14). Probes with poor technical quality were excluded from the analysis prior to functional normalisation.

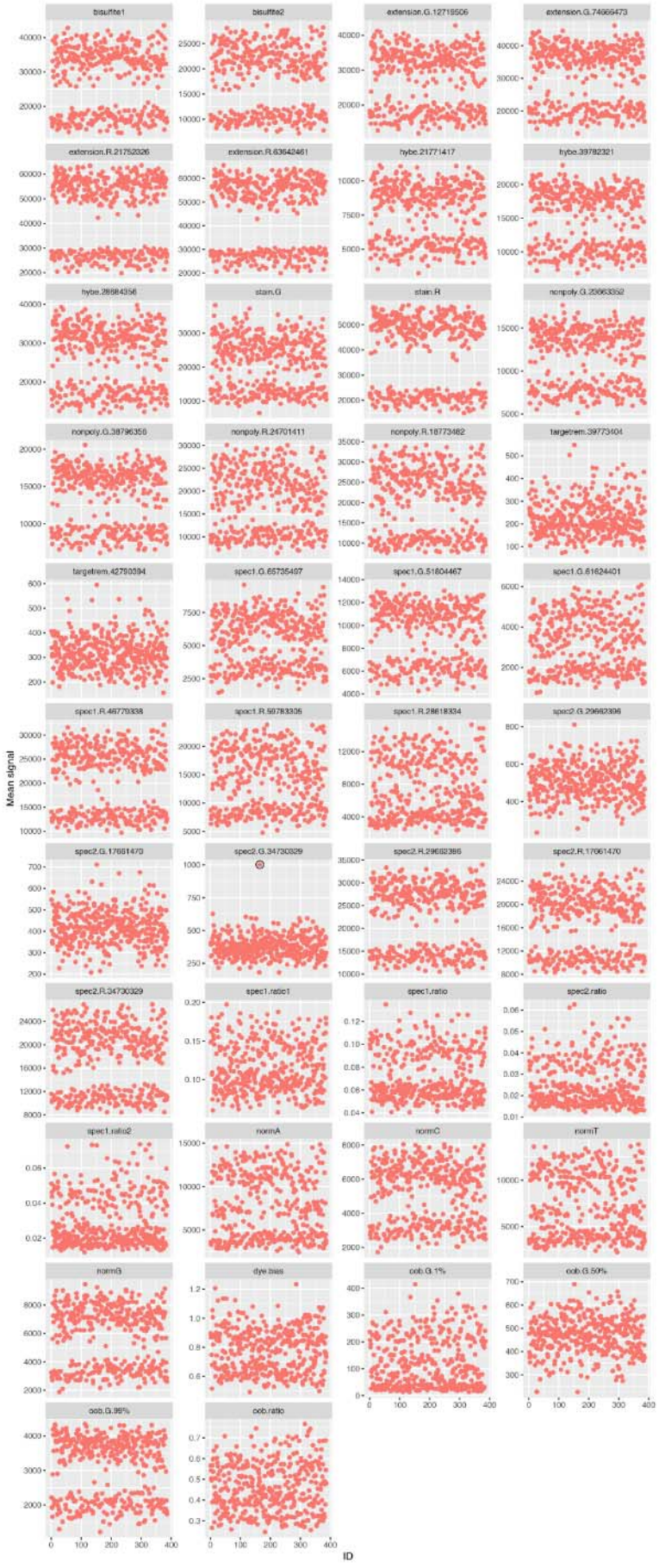

Figure 3.34: Control probe signal by sample for each summary group for the EPIC data. Outliers would be circled in black. Plot generated by `meffil` QC report.

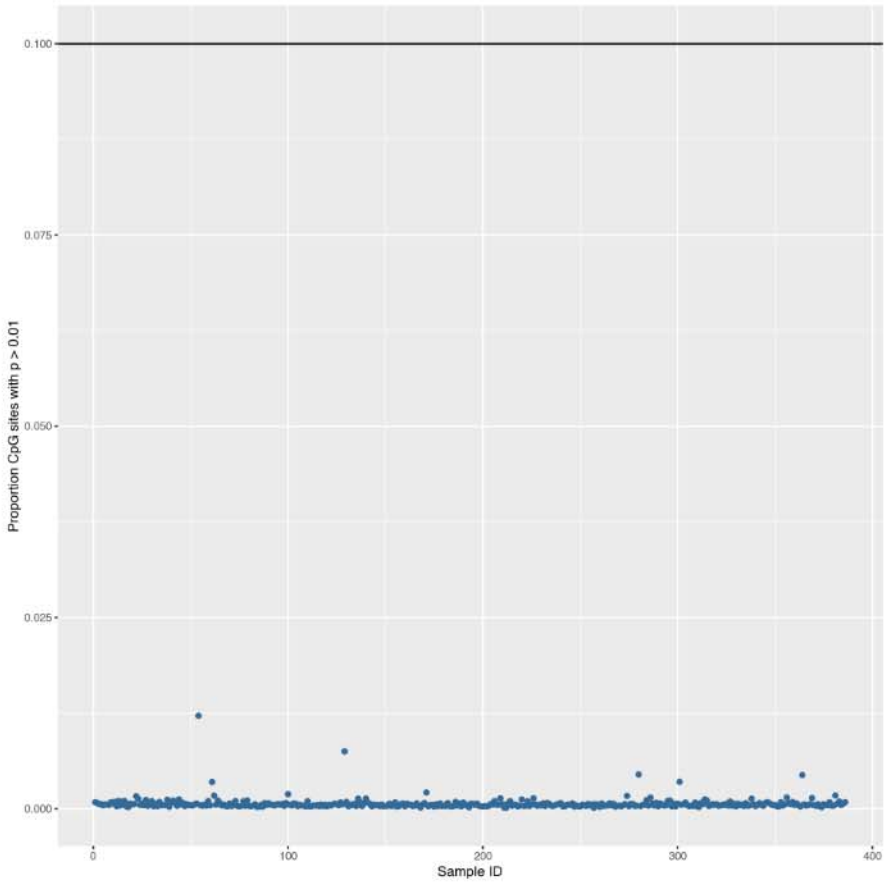

Figure 3.35: Undetectable probes across samples for EPIC data. Manhattan plot showing proportion of samples (y) in which a given probe (x) is not distinguishable from background noise, i.e. a detection p-value of > 0.01. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

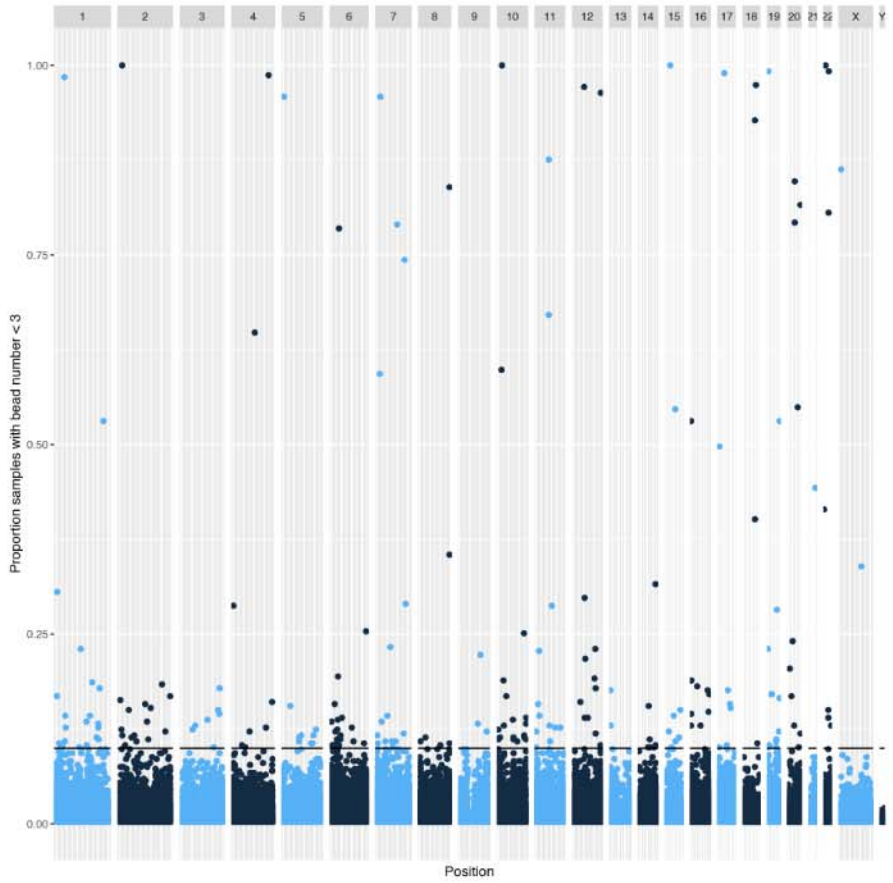

Figure 3.36: Low bead count probes across samples for EPIC data. Manhattan plot showing the proportion of samples (y) in which a given probe (x) has a bead count of <3. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

3.5.3.3 Functional Normalisation

An m of 6 was chosen as this value produced the last steep drop in residual variation, see Figure 3.37.

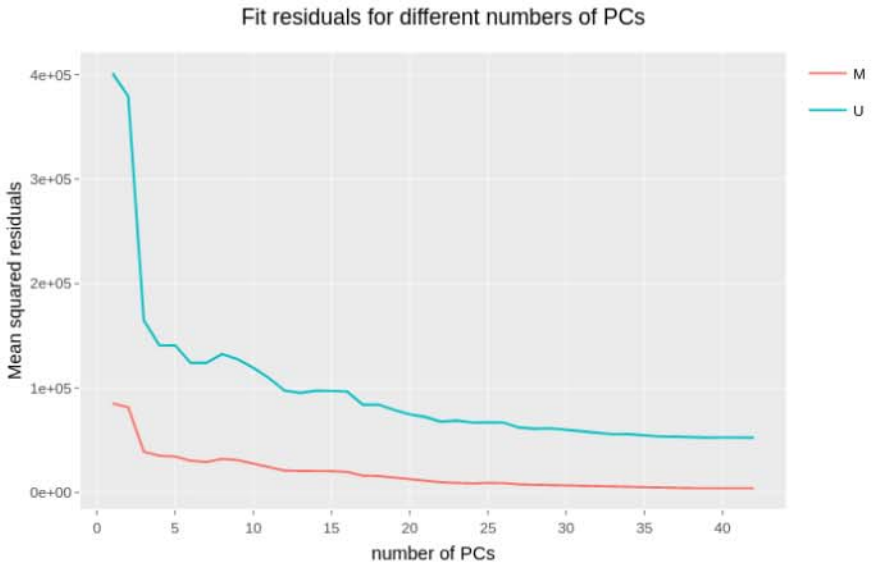

Figure 3.37: Residual variation remaining after functional normalisation of the top 20,000 most variable probes with m PCs from the control probe summary matrices for the EPIC array samples (n=237), for M = methylated and U = unmethylated probes.

3.5.3.4 EWAS

3.5.3.4.1 Bone Mineral Content at 4 years The model for this EWAS attempted to predict whole body (minus head) bone mineral content (g) (Figure 3.38) correcting for: Cell type composition (B-cells, CD4+T, CD8+T, Granulocytes, Monocytes, Natural Killer cells, Endothelial cells, Epithelial cells, Stromal cells), Mother's age at birth, Sex, Mother's BMI at 11 weeks gestation, whether the mother smoked during pregnancy, & Gestational Age. The EWAS was performed on 237 individuals. surrogate variable analysis identified 19 significant surrogate variables which were included in the model. No CpGs fell below the Bonferroni corrected significance threshold for an association between DNAm at that locus and bone mineral content, in any of the models including SVA (Figure 3.39).

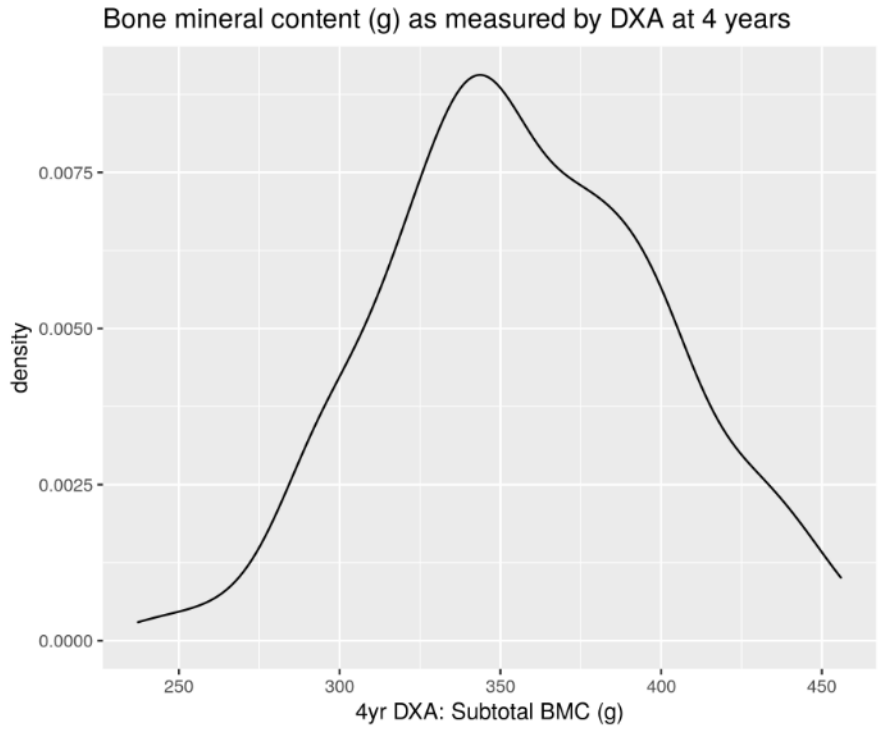

Figure 3.38: Distribution of whole body (minus head) bone mineral content in grams at 4 years of age.

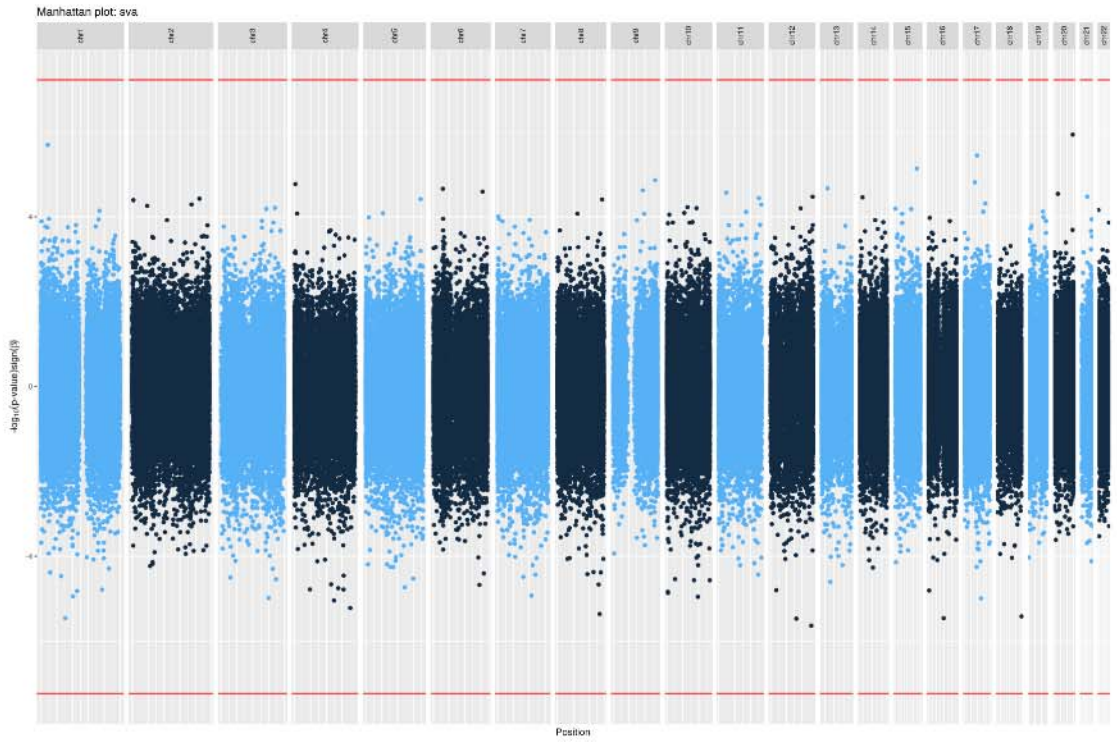

Figure 3.39: Results of EWAS for whole body minus head bone mineral content (g) with SVA model (n = 237). Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. Size and transparency of points increases with $-log_{10}(p-value)$ such that the most significant CpGs are represented by the largest and least translucent points. x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 6.18×10^{-8} $(0.05 \div 808,585)$.

3.5.3.5 RXRA & CDKN2A

None of the probes associated with the *RXRA* and *CDKN2A* loci exhibited genome wide significant changes in DNA methylation however nominally significant ($p < 0.05$) changes are observed at some of the probes within these genes in the MAVIDOS phase II analysis dataset. Table 3.3 lists these model results for maternal vitamin D levels and table 3.4 for bone mineral content and density. Appendix 6.13 provides the complete list of these model results not just the nominally significant results listed here.

The result that 25 of 27 probes with nominally significant changes in DNA methylation are for decreasing methylation with increased vitamin D is consistent with the lower methylation in the offspring of the supplemented mothers in the original publication on the *RXRA* changes [263], Figure 3.40. However at 34 weeks only 4 probes showed nominally significant changes and 3 of the 4 were increases in methylation with vitamin D, Figure 3.41.

Table 3.3: Results for modelling Maternal Vitamin D levels with umbilical cord tissue DNA methylation levels and covariates at probes associated with the *RXRA* and *CDKN2A* genes. vit D: maternal circulating vitamin D nmol/l

Timepoint	Model type	Gene	position	Probe	Slope	p-value
11 weeks	Corrected	CDKN2A	Chr9: 21,968,233- 21,968,233	cg12840719	230.2170	0.0066
			Chr9: 21,986,062- 21,986,062	cg00550721	-63.8120	0.0291
			Chr9: 21,991,666- 21,991,666	cg13109095	-122.1244	0.0358
			Chr9: 21,975,898- 21,975,898	cg10458809	27.5771	0.0365
			Chr9: 21,971,256- 21,971,256	cg08686553	96.1148	0.0378
	Uncorrected		Chr9: 21,968,233- 21,968,233	cg12840719	180.7349	0.0147
			Chr9: 21,986,062- 21,986,062	cg00550721	-50.0384	0.0411
	Corrected	RXRA	Chr9:137,329,237-137,329,237	cg14344989	-92.4988	0.0005
			Chr9:137,250,935-137,250,935	cg02059519	-106.0882	0.0007
			Chr9:137,271,457-137,271,457	cg13312109	-74.2411	0.0007
			Chr9:137,263,644-137,263,644	cg05829961	-106.9248	0.0019
			Chr9:137,283,224-137,283,224	cg00622905	-108.5121	0.0021
			Chr9:137,260,147-137,260,147	cg04900958	-99.1450	0.0033
			Chr9:137,268,524-137,268,524	cg10122776	-141.6744	0.0035
			Chr9:137,266,722-137,266,722	cg06831947	-109.8622	0.0036
			Chr9:137,279,180-137,279,180	cg14051662	-82.8152	0.0040
			Chr9:137,262,485-137,262,485	cg25619159	-90.9658	0.0043
			Chr9:137,329,241-137,329,341	cg24164254	-88.6921	0.0068
			Chr9:137,282,509-137,282,509	cg00735574	-95.1502	0.0105
			Chr9:137,301,743-137,301,743	cg08868540	-130.8756	0.0105
			Chr9:137,302,231-137,302,231	cg13413384	-92.9525	0.0107
			Chr9:137,252,129-137,252,129	cg14236758	-89.7349	0.0114
			Chr9:137,326,282-137,326,382	cg25300535	27.8782	0.0142
			Chr9:137,271,104-137,271,104	cg01894436	-96.1754	0.0225
			Chr9:137,272,766-137,272,766	cg13847322	-73.9350	0.0230
			Chr9:137,299,685-137,299,685	cg00545196	-63.1364	0.0282
			Chr9:137,268,074-137,268,074	cg14121282	-65.2194	0.0323
			Chr9:137,257,228-137,257,228	cg10271868	-62.4176	0.0361
			Chr9:137,298,350-137,298,350	cg01063003	-71.8198	0.0376
			Chr9:137,258,920-137,258,920	cg13786567	-63.6464	0.0397
			Chr9:137,265,865-137,265,865	cg14484045	-77.6335	0.0440
			Chr9:137,270,186-137,270,186	cg13941235	-44.6552	0.0466
			Chr9:137,225,311-137,225,311	cg13689099	-49.7538	0.0487
	Uncorrected		Chr9:137,266,722-137,266,722	cg06831947	-93.4263	0.0020
			Chr9:137,250,935-137,250,935	cg02059519	-79.0829	0.0022
			Chr9:137,271,457-137,271,457	cg13312109	-50.1030	0.0066
			Chr9:137,268,524-137,268,524	cg10122776	-108.6295	0.0122
			Chr9:137,262,485-137,262,485	cg25619159	-71.3247	0.0135
			Chr9:137,279,180-137,279,180	cg14051662	-62.6430	0.0146
			Chr9:137,260,147-137,260,147	cg04900958	-63.4042	0.0187
			Chr9:137,326,282-137,326,382	cg25300535	24.7823	0.0230
			Chr9:137,282,509-137,282,509	cg00735574	-56.2782	0.0273
			Chr9:137,268,074-137,268,074	cg14121282	-43.7701	0.0299
			Chr9:137,221,918-137,221,918	cg00639224	32.4576	0.0406
			Chr9:137,272,766-137,272,766	cg13847322	-53.5243	0.0433
34 weeks	Corrected	CDKN2A	Chr9: 21,975,622- 21,975,622	cg25339269	-472.4791	0.0081
		RXRA	Chr9:137,217,473-137,217,473	cg14204281	-258.5922	0.0259
			Chr9:137,301,635-137,301,635	cg18566083	114.9198	0.0483
			Chr9:137,215,264-137,215,364	cg04329455	99.9363	0.0489
	Uncorrected		Chr9:137,299,285-137,299,285	cg13865488	59.8832	0.0427

[a] Corrected: vit D ~ beta + wageb + Sex + awbmi + smokpreg + egest + Bcell + CD4T + CD8T + Gran + Mono + NK + Endothelial + Epithelial + Stromal, Uncorrected: vit D ~ beta, beta: methylation value

Table 3.4: Results for modelling bone mineral content and density at 4 years of age with umbilical cord tissue DNA methylation levels and covariates at probes associated with the *RXRA* and *CDKN2A* genes. Corrected: BMC/D ~ beta + wageb + Sex + awbmi + smokpreg + egest + Bcell + CD4T + CD8T + Gran + Mono + NK + Endothelial + Epithelial + Stromal, Uncorrected: BMC/D ~ beta, beta: methylation value, BMC/D Bone Mineral Content (g)/Density(gcm-2) at 4 years (DXA)'

Metric	Model type	Gene	position	Probe	Slope	p-value
BMC	Corrected	CDKN2A	Chr9: 21,991,666- 21,991,666	cg13109095	-353.4675	0.0270
			Chr9: 21,975,243- 21,975,243	cg16468520	1968.4218	0.0347
			Chr9: 21,975,023- 21,975,023	cg26269171	906.1528	0.0468
	Uncorrected		Chr9: 21,975,243- 21,975,243	cg16468520	2174.9837	0.0117
			Chr9: 21,975,495- 21,975,495	cg06955353	1041.1349	0.0239
			Chr9: 21,975,023- 21,975,023	cg26269171	979.8917	0.0268
			Chr9: 21,975,898- 21,975,898	cg10458809	72.1257	0.0339
			Chr9: 21,975,305- 21,975,305	cg02008397	939.3506	0.0416
			Chr9: 21,974,704- 21,974,704	cg13601799	388.9265	0.0462
	Corrected	RXRA	Chr9:137,232,372-137,232,372	cg14462321	-232.0554	0.0130
			Chr9:137,210,392-137,210,392	cg03416552	202.2096	0.0197
			Chr9:137,299,436-137,299,436	cg14654324	136.6815	0.0383
	Uncorrected		Chr9:137,232,372-137,232,372	cg14462321	-206.8839	0.0050
			Chr9:137,299,436-137,299,436	cg14654324	137.8739	0.0319
			Chr9:137,329,237-137,329,237	cg14344989	-88.5216	0.0379
			Chr9:137,291,076-137,291,076	cg02766323	-360.6465	0.0439
BMD	Corrected	CDKN2A	Chr9: 21,975,243- 21,975,243	cg16468520	1.5214	0.0355
	Uncorrected		Chr9: 21,975,243- 21,975,243	cg16468520	1.5416	0.0259
			Chr9: 21,974,704- 21,974,704	cg13601799	0.3137	0.0447
	Corrected	RXRA	Chr9:137,232,372-137,232,372	cg14462321	-0.1705	0.0188
			Chr9:137,268,524-137,268,524	cg10122776	-0.2328	0.0252
			Chr9:137,299,436-137,299,436	cg14654324	0.1098	0.0319
			Chr9:137,302,231-137,302,231	cg13413384	-0.1657	0.0335
			Chr9:137,264,876-137,264,876	cg18483374	0.1447	0.0351
			Chr9:137,210,289-137,210,289	cg01557614	0.1434	0.0457
	Uncorrected		Chr9:137,232,372-137,232,372	cg14462321	-0.1758	0.0029
			Chr9:137,264,876-137,264,876	cg18483374	0.1344	0.0085
			Chr9:137,268,524-137,268,524	cg10122776	-0.2471	0.0097
			Chr9:137,299,436-137,299,436	cg14654324	0.1126	0.0287
			Chr9:137,326,227-137,326,227	cg14265066	-0.2243	0.0327
			Chr9:137,228,542-137,228,542	cg04875697	-0.2058	0.0438

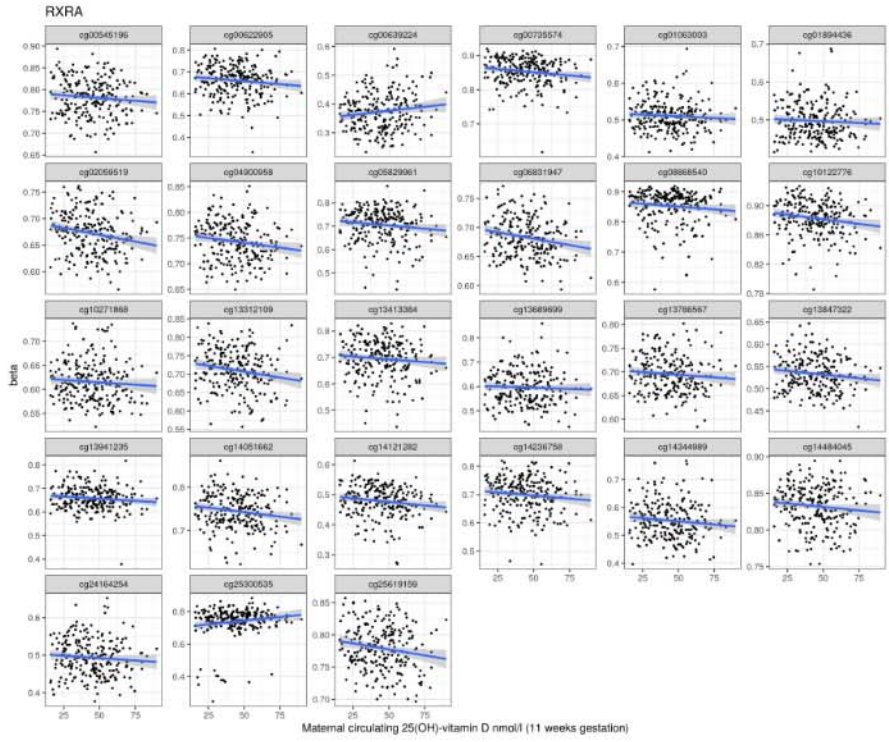

Figure 3.40: *RXRA* gene associated probes showing nominally significant (p<0.05) changes in DNA methylation with maternal vitamin D at 11 weeks gestation

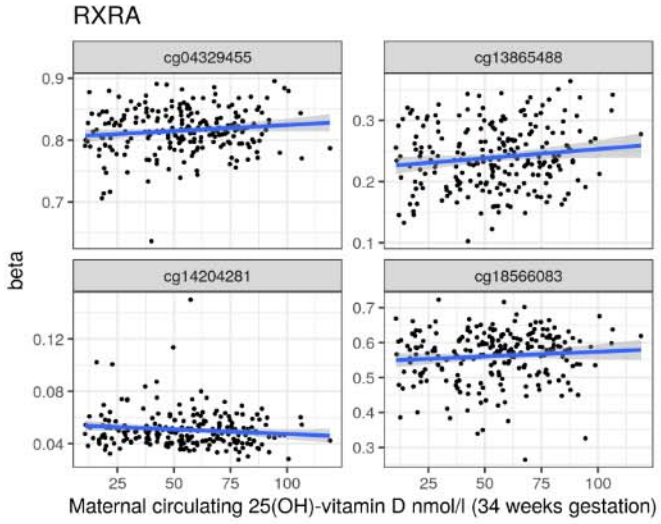

Figure 3.41: *RXRA* gene associated probes showing nominally significant (p<0.05) changes in DNA methylation with maternal vitamin D at 34 weeks gestation

The inverse relationship between DNA methylation and bone mineral content / density in the *ANRIL* promoter region of *CDKN2A* is not seen in the majority of *CDKN2A* probes which show nominally significant changes with bone mineral content / Density in this analysis. Of the 7 probes associated with BMC 1 shows a decrease (Figure 3.42) and both probes associated with

BMD show an increase (Figure 3.43).

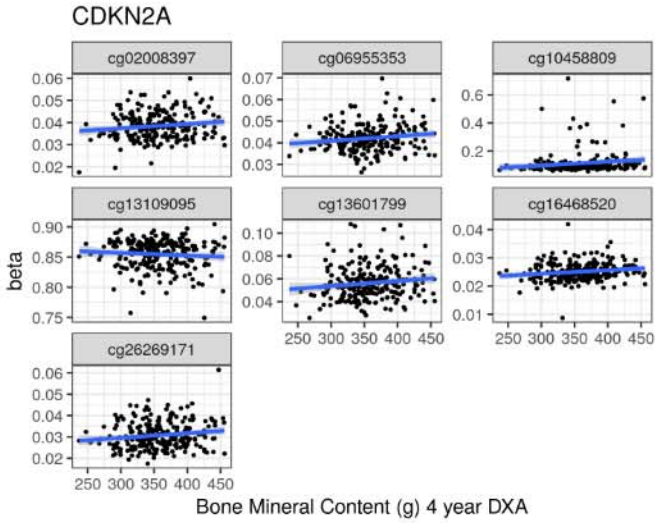

Figure 3.42: *CDKN2A* gene associated probes showing nominally significant (p<0.05) changes in DNA methylation with bone mineral density at 4 years

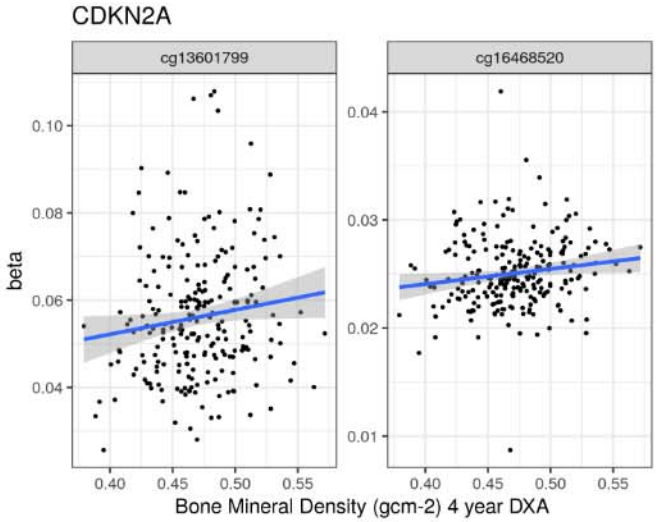

Figure 3.43: *CDKN2A* gene associated probes showing nominally significant (p<0.05) changes in DNA methylation with bone mineral density at 4 years

These results do not correspond to the same CpGs examined in the previous studies, either in the case of *RXRA* [263] or *CDKN2A* [264]. Thus these results cannot directly support or contradict these findings for the same specific CpG sites, making their interpretation somewhat challenging. In the case of *RXRA* these results appear somewhat supportive of the previous observation that there is lower methylation at higher vitamin D levels at least when considering only the 11 weeks gestation data. Whereas the *CDKN2A* results do not support the previous finding of an inverse

Table 3.5: Summary statistics for variables used in the SWS EWAS. Blood cell count information is estimated based on the methylation data as described in the methods. (Ob.) obstetric exam, (init.) initial survey.

Description	mean	median	min	max
Ob: Woman's age at child's birth	31.67	32.08	23.99	38.12
Init: Woman's body mass index	24.82	24.29	17.40	40.17
Smoking in pregnancy	0.14	0.00	0.00	1.00
Ob: Gestational age - from LMP data, or U/S scan or Elaine	39.84	40.19	33.14	42.73
B cells	0.09	0.09	0.01	0.21
CD4+ T-cells	0.17	0.16	0.05	0.34
CD8+ T-cells	0.13	0.13	0.05	0.31
Gran	0.45	0.43	0.19	0.66
Monocytes	0.04	0.04	0.00	0.11
Natural Killer Cells	0.00	0.00	0.00	0.07
Red Blood Cells	0.14	0.11	0.04	0.39
8 yr DXA: Total BMC (kg), without heads,adjusted for sex and age	0.72	0.70	0.46	1.08
6 yr DXA: Total BMC (kg), without heads,adjusted for sex and age	0.55	0.55	0.39	0.73
6yr PQCT: 38% - Periosteal circumference (circular ring model) (mm)	52.35	51.86	44.10	64.35
6yr PQCT: 38% - Cortical density (mg/cm3)	1036.39	1040.60	949.90	1111.10

relationship between DNA methylation and bone mineral content / density at the loci which they target.

3.5.4 Southampton Women's Survey (SWS)

Summary statistics for the variables used in the EWAS models are provided in tables 3.5 & 3.6.

DNA methylation at two CpGs were significantly associated with total bone mineral content at 6 years and periosteal circumference at 6 years respectively. CpG cg26559250 located at Chr6:157,653,445-157,653,447 at the ZDHHC14 (zinc finger DHHC-type palmitoyltransferase 14) gene showed an increase of DNA methylation with increasing bone mineral content with a significance of 2.52×10^{-8} CpG cg22570676 located at Chr19:2,527,492-2,527,494 at the GNG7 (G protein subunit gamma 7) gene showed an increase of DNA methylation with increasing periosteal circumference with a significance of 4.24×10^{-8}. Neither probe is flagged for known technical issues or genetic confounding [223].

3.5.4.1 Whole Array QC

The predicted sex of the samples generated using sex chromosome probe intensities was checked against that in the sample annotation and no mismatches were found (Figure 3.44). No samples were excluded for having mismatched Sex.

Table 3.6: Summary statistics for variables used in the SWS EWAS divided by the sex of the child. Blood cell count information is estimated based on the methylation data as described in the methods. n = 464, Female n=226, Male n=238. (Ob.) obstetric exam, (init.) initial survey.

Sex	Description	mean	median	min	max
Female	Ob: Woman's age at child's birth	31.97	32.23	23.99	38.12
	Init: Woman's body mass index	25.47	24.67	18.60	40.17
	Smoking in pregnancy	0.14	0.00	0.00	1.00
	Ob: Gestational age - from LMP data, or U/S scan or Elaine	39.87	40.21	33.14	42.73
	B cells	0.10	0.09	0.01	0.21
	CD4+ T-cells	0.17	0.17	0.05	0.34
	CD8+ T-cells	0.12	0.12	0.05	0.31
	Granulocytes	0.47	0.48	0.19	0.66
	Monocytes	0.04	0.04	0.00	0.11
	Natural Killer Cells	0.00	0.00	0.00	0.04
	Red Blood Cells	0.12	0.10	0.04	0.32
	8 yr DXA: Total BMC (kg), without heads,adjusted for sex and age	0.72	0.72	0.46	1.08
	6 yr DXA: Total BMC (kg), without heads,adjusted for sex and age	0.56	0.56	0.39	0.73
	6yr PQCT: 38% - Periosteal circumference (circular ring model) (mm)	52.69	51.99	44.74	64.35
	6yr PQCT: 38% - Cortical density (mg/cm3)	1039.99	1041.40	949.90	1110.90
Male	Ob: Woman's age at child's birth	31.27	31.80	24.17	37.73
	Init: Woman's body mass index	23.98	23.12	17.40	32.42
	Smoking in pregnancy	0.14	0.00	0.00	1.00
	Ob: Gestational age - from LMP data, or U/S scan or Elaine	39.80	39.79	35.77	42.14
	B cells	0.09	0.09	0.05	0.18
	CD4+ T-cells	0.17	0.15	0.06	0.33
	CD8+ T-cells	0.14	0.13	0.07	0.26
	Granulocytes	0.42	0.40	0.24	0.64
	Monocytes	0.04	0.04	0.00	0.08
	Natural Killer Cells	0.01	0.00	0.00	0.07
	Red Blood Cells	0.16	0.13	0.04	0.39
	8 yr DXA: Total BMC (kg), without heads,adjusted for sex and age	0.70	0.69	0.50	1.00
	6 yr DXA: Total BMC (kg), without heads,adjusted for sex and age	0.53	0.52	0.42	0.67
	6yr PQCT: 38% - Periosteal circumference (circular ring model) (mm)	51.89	51.83	44.10	62.47
	6yr PQCT: 38% - Cortical density (mg/cm3)	1031.69	1030.10	961.00	1111.10

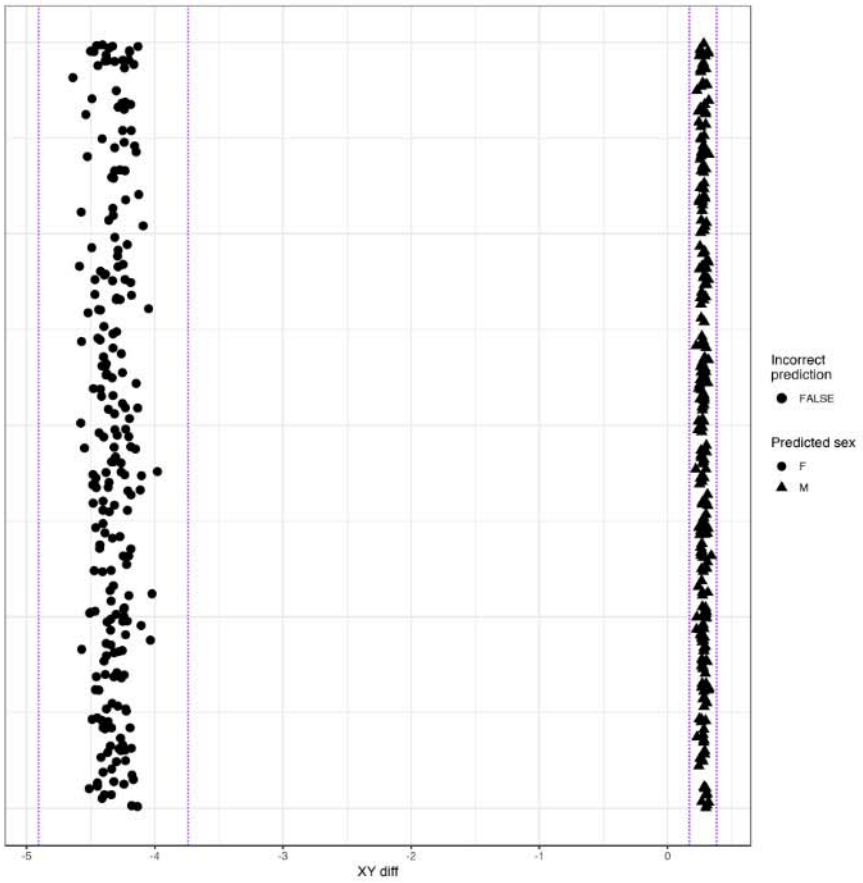

Figure 3.44: Predicted sex of each sample based on the sex chromosome copy numbers inferred from probe intensities for the EPIC array data. Mismatches between the predicted sex and that asserted in the sample annotation metadata are shown in red. Plot generated by `meffil` QC report.

No samples were excluded for having a median methylated signal that was more than 3σ from the expected value, (Figure 3.45). No samples were excluded for having a higher than expected proportion of undetected probes (proportion of probes with detection p-value > 0.01 is > 0.1) (Figure 3.46). No samples were excluded for having a high proportion of probes with low bead counts (proportion of probes with bead number < 3 is > 0.1), (Figure 3.47).

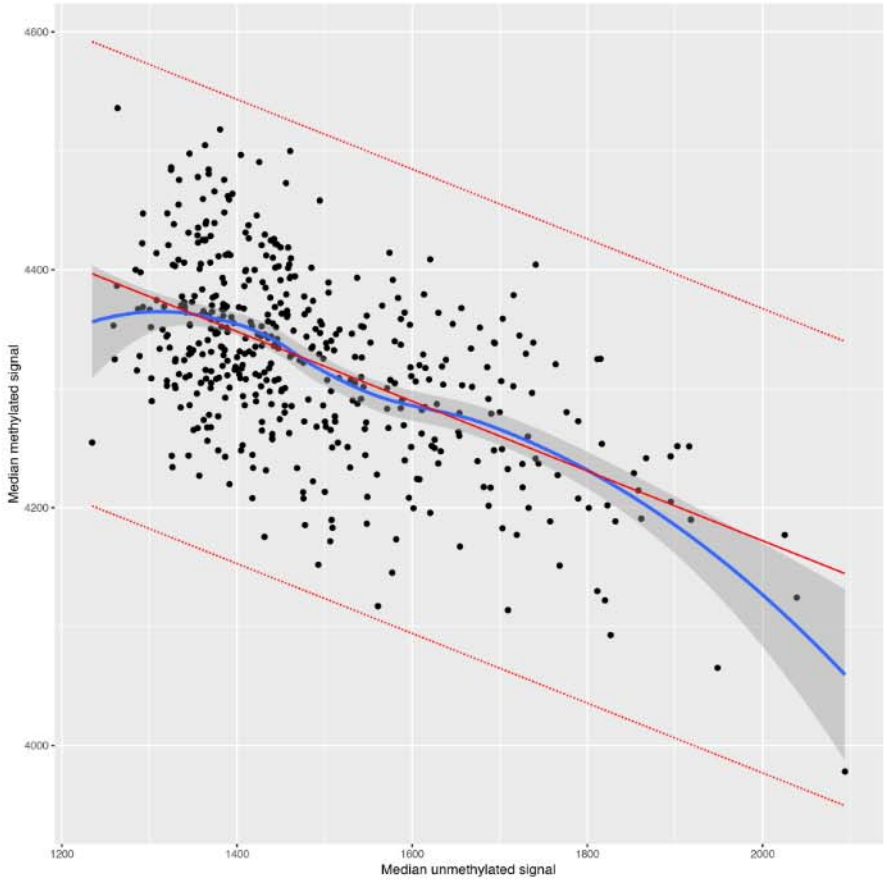

Figure 3.45: Median methylated signal vs unmethylated signal per sample for the EPIC array data, solid red line indicates linear regression of median methylated signal vs median unmethylated signal with dotted red lines representing 3σ from the expected mean. Samples outside the expected range are indicated in the legend. Plot generated by `meffil` QC report.

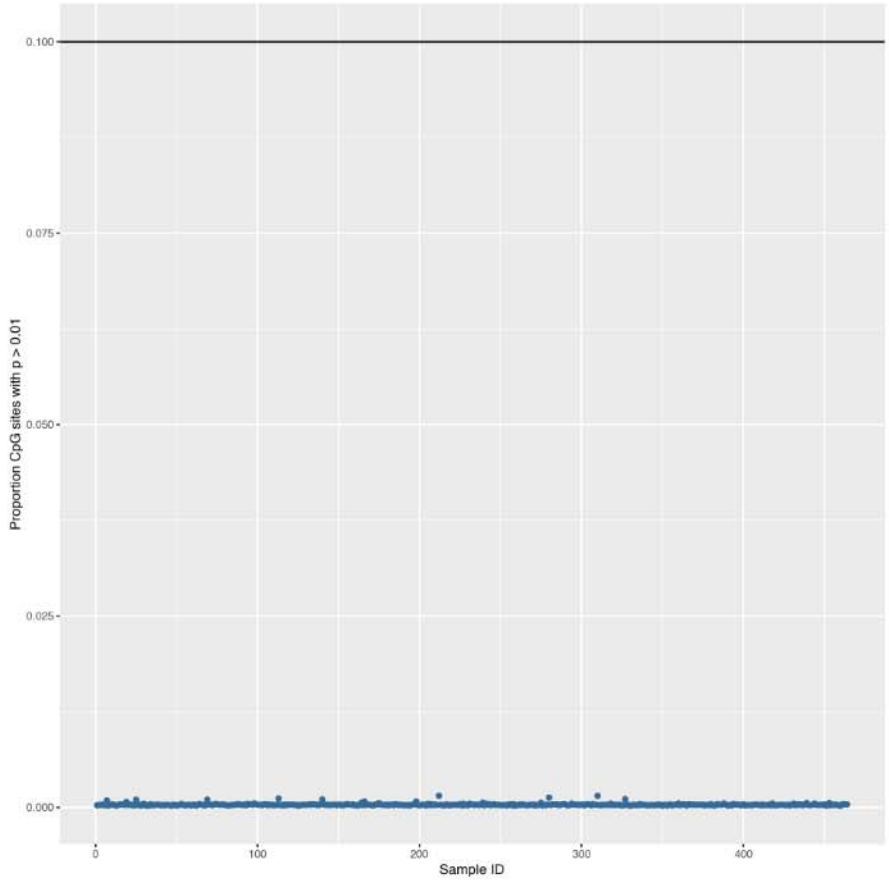

Figure 3.46: Proportion of probes with detection p-values >0.01 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

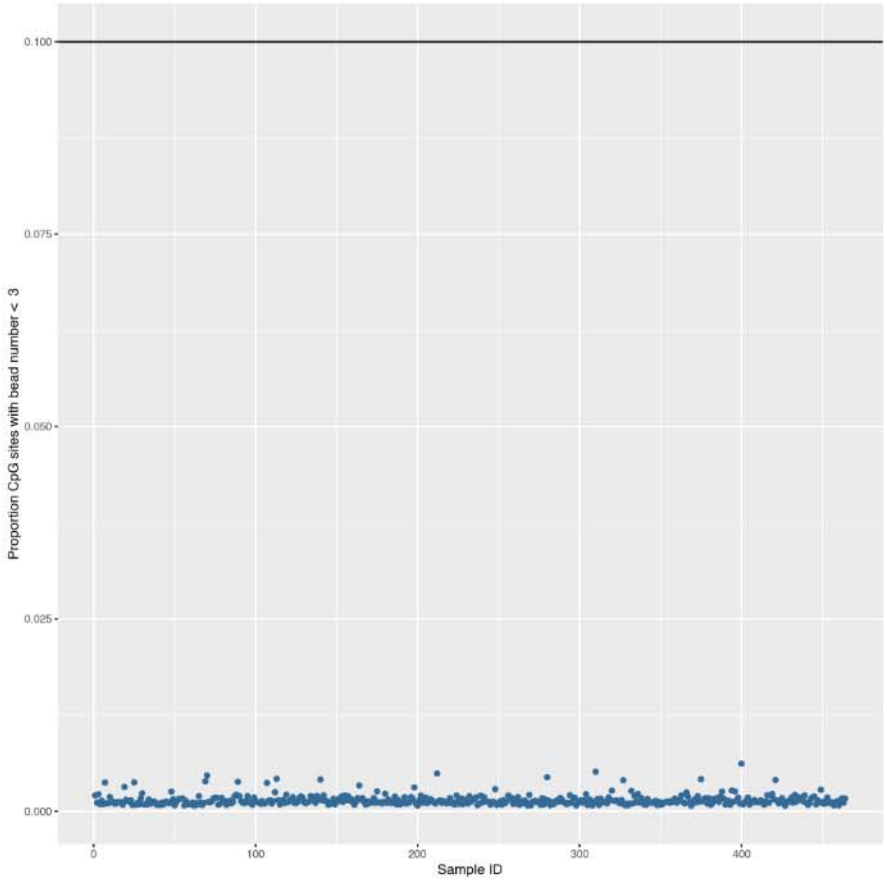

Figure 3.47: Proportion of probes with a bead count of < 3 by sample for the EPIC array data. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

3.5.4.2 Probe QC

There were no outliers within the control probes (Figure 3.48). 833 probes were excluded for having high background signal in a high proportion of samples (proportion of samples with detection p-value > 0.01 is > 0.1), (Figure 3.49). 127 probes were excluded for having low bead count in a high proportion of samples (proportion of samples with bead number < 3 is > 0.1), (Figure 3.50). Probes with poor technical quality were excluded from the analysis prior to functional normalisation.

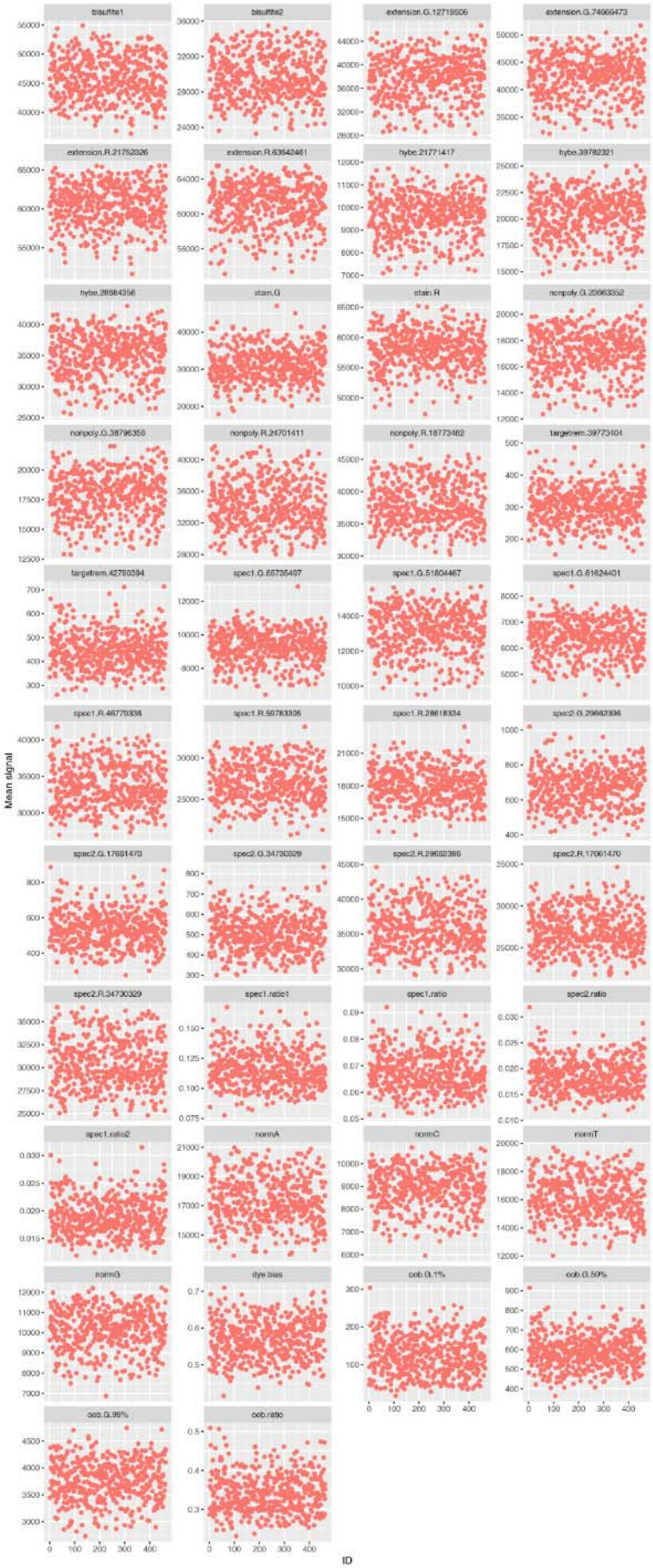

Figure 3.48: Control probe signal by sample for each summary group for the EPIC data. Outliers would be circled in black. Plot generated by `meffil` QC report.

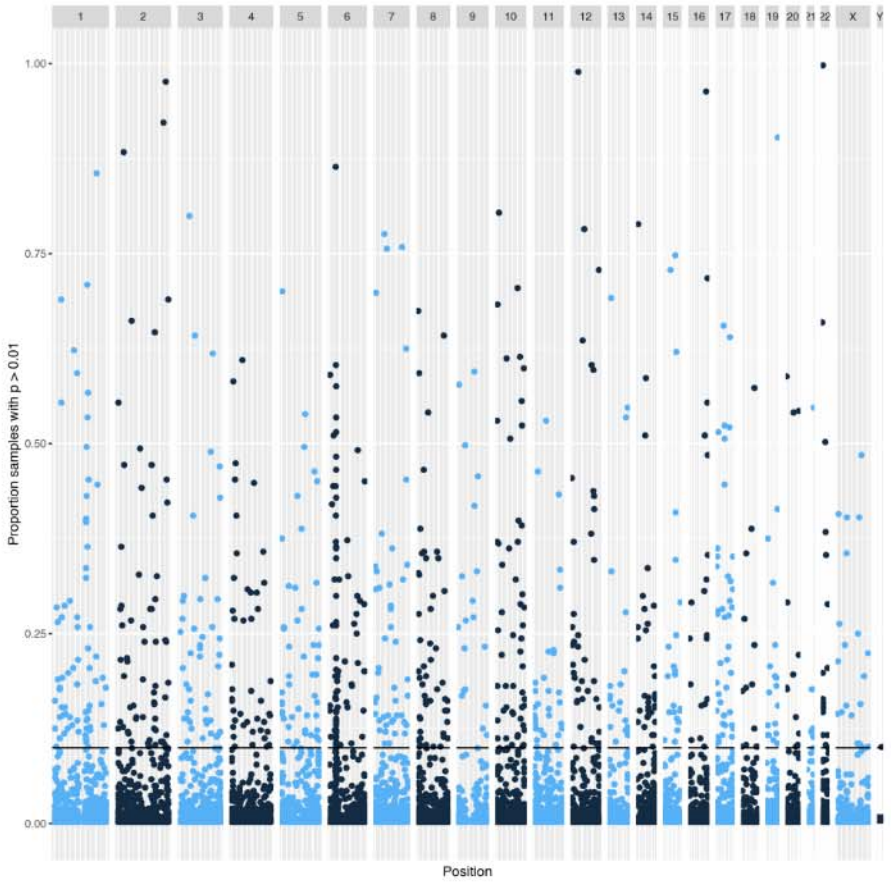

Figure 3.49: Undetectable probes across samples for EPIC data. Manhattan plot showing proportion of samples (y) in which a given probe (x) is not distinguishable from background noise, i.e. a detection p-value of > 0.01. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

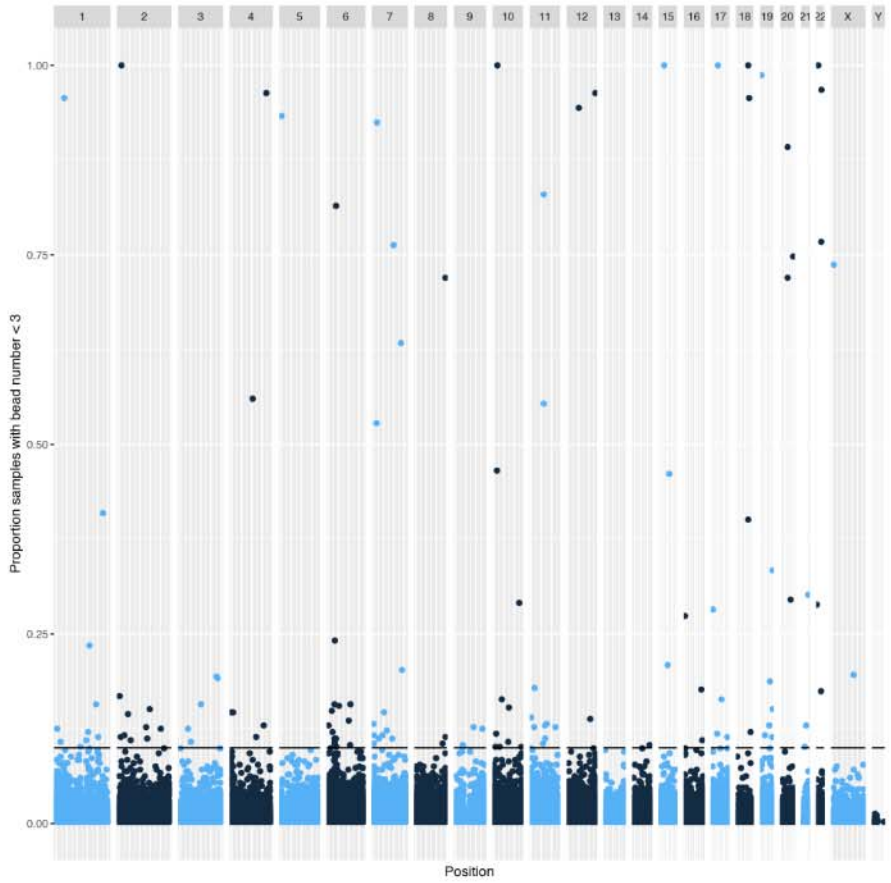

Figure 3.50: Low bead count probes across samples for EPIC data. Manhattan plot showing the proportion of samples (y) in which a given probe (x) has a bead count of <3. Black line indicates the exclusion threshold of 0.1. Plot generated by `meffil` QC report.

3.5.4.3 Functional Normalisation

An m of 12 was chosen as this value produced the last steep drop in residual variation, see Figure 3.51.

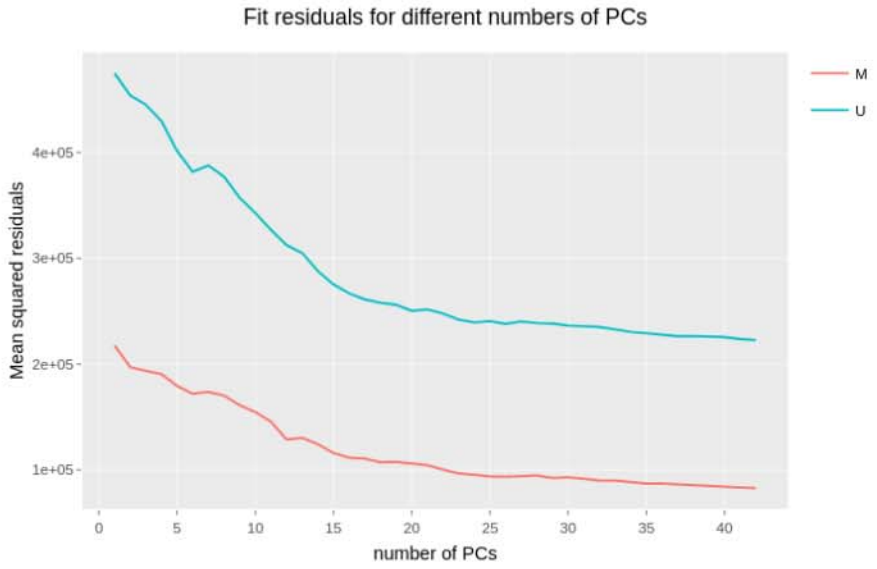

Figure 3.51: Residual variation remaining after functional normalisation of the top 20,000 most variable probes with m PCs from the control probe summary matrices for the EPIC array samples (n=464), for M = methylated and U = unmethylated probes.

3.5.4.4 EWAS

For all the EWAS, blood cell-type counts were estimated using the Houseman method [272] and the cord blood cell-type reference panel from Bakulski et al. [275]. The Cell-types estimated were: B cells, CD4+ T cells, CD8+ T Cells, Granulocytes, Monocytes, Natural Killer cells, & Erythrocytes. In addition to the estimated blood cell counts all models included as covariates: Maternal Age at time of birth (years), Sex, maternal BMI at 11 weeks gestation, parity, whether or not the mother smoked during pregnancy, and gestational age.

EWAS for bone mineral content at 6 and 8 years would be expected to produce similar results given that BMC at these ages are highly correlated (Figure 3.52).

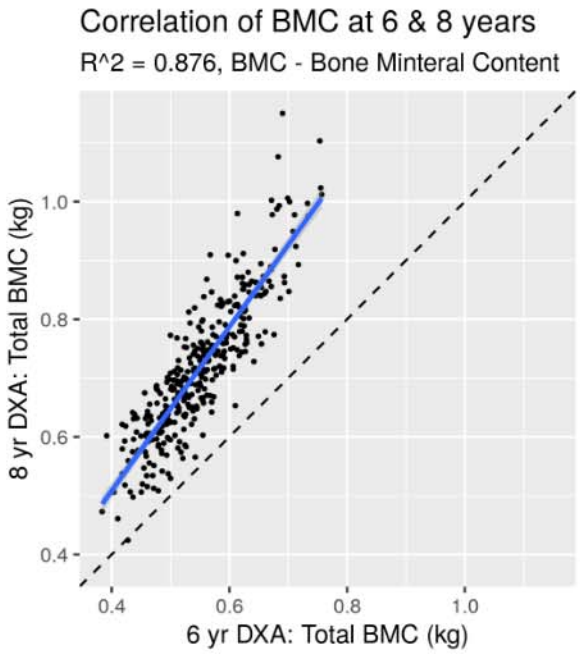

Figure 3.52: Bone Mineral content at 6 and 8 years of are correlated with an $R^2 = 0.88$.

3.5.4.4.1 Total Bone Mineral Conent at 8 years Figure 3.53, illustrates the distribution of bone mineral content at 8 years in the 408 individuals in this EWAS. Surrogate variable analysis identified 95 significant surrogate variables, this is likely an overestimate stemming from small amounts of variation remaining unaccounted for by the manual model thus the Manhattan plots based on the manual model were included. No CpGs fell below the Bonferroni corrected significance threshold (5.92×10^{-8}) for an association between DNA methylation at that locus and total bone mineral content minus head at 8 years adjusted for age and sex, Figure 3.54.

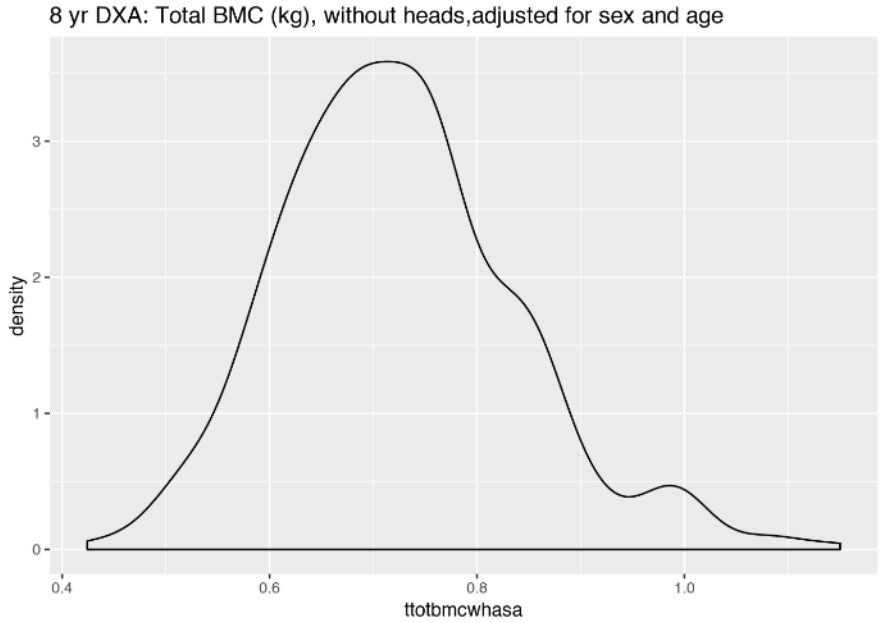

Figure 3.53: Distribution of whole body (minus head) bone mineral content in kg (ttotbmcwhasa) at 8 years of age (n = 408), adjusted for sex and age, as measured by DXA.

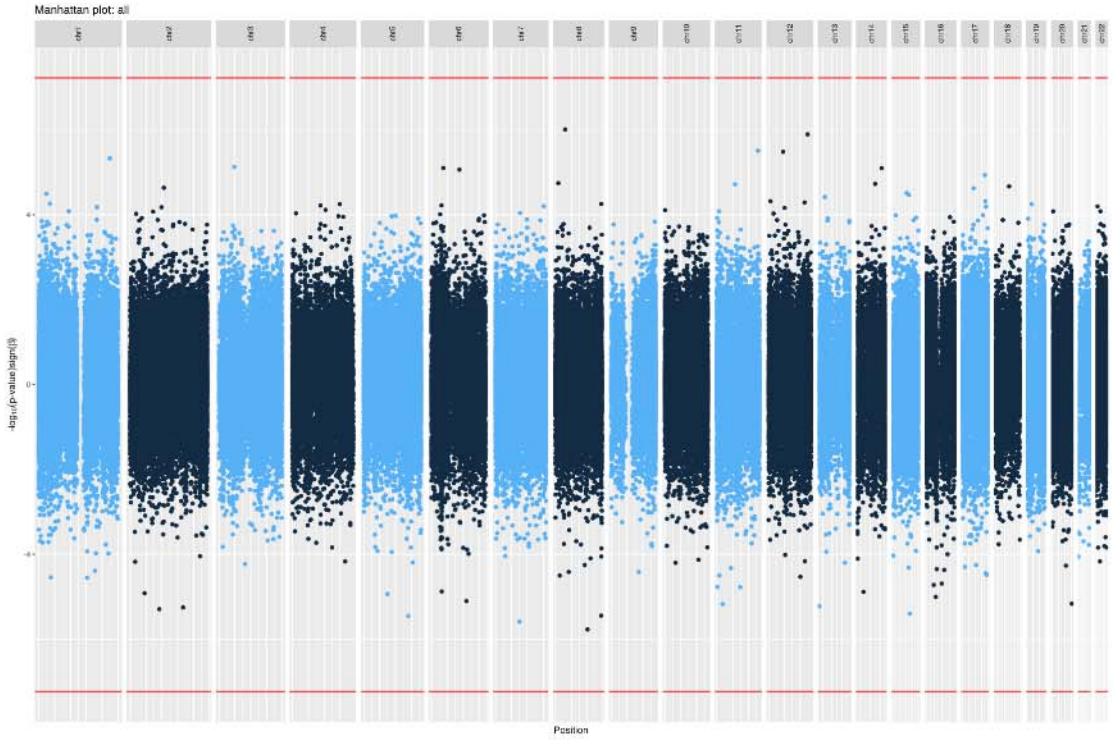

Figure 3.54: Results of EWAS for whole body (minus head) bone mineral content in kg at 8 years of age (n = 408), adjusted for sex and age. The 'all' model results are shown here. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. The x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 5.92×10^{-8}.

3.5.4.4.2 Total Bone Mineral Conent at 6 years Figure 3.55, illustrates the distribution of bone mineral content at 6 years in the 402 individuals in this EWAS. Surrogate variable analysis identified 97 significant surrogate variables, this is likely an overestimate stemming from small amounts of variation remaining unaccounted for by the manual model thus Manhattan plots based on the manual model have been included.

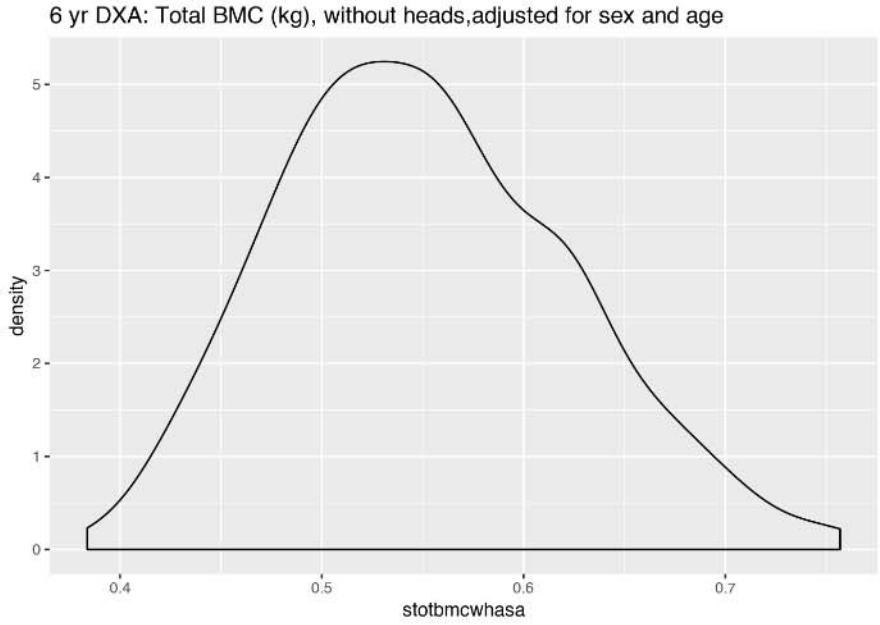

Figure 3.55: Distribution of whole body (minus head) bone mineral content in kg (stotbmcwhasa) at 6 years of age (n = 402), adjusted for sex and age, as measured by DXA.

One CpG fell below the Bonferroni corrected significance threshold (5.92×10^{-8}) for an association between DNA methylation at that locus and total bone mineral content minus head at 6 years adjusted for age and sex, Figure 3.56. This CpG was cg26559250 which is located at Chr6:157,653,445-157,653,447 adjacent to the ZDHHC14 (zinc finger DHHC-type palmitoyltransferase 14) gene. cg26559250 was significant ($p = 2.52 \times 10^{-8}$, increase of 1.46% per kg) in the 'all' model and was also Bonferroni significant in the uncorrected model. However, cg26559250 was not significant in the SVA or iSVA models suggesting that it may be attributable to batch or cell-type effects.

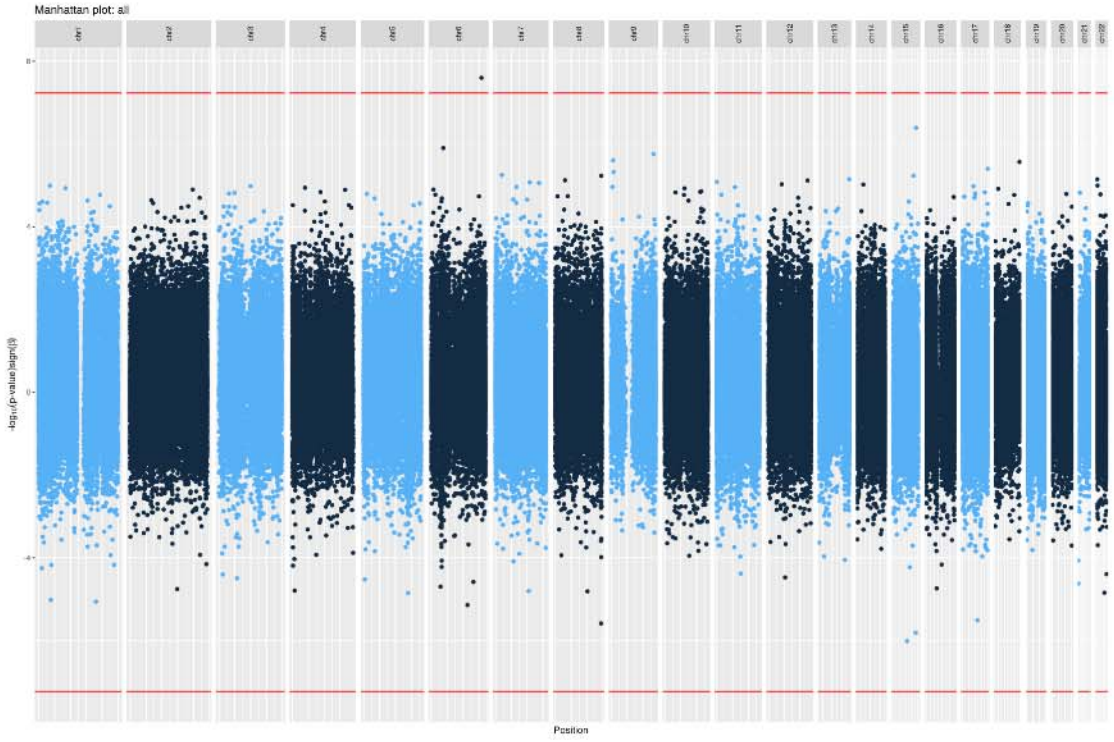

Figure 3.56: Results of EWAS for whole body (minus head) bone mineral content in kg at 6 years of age (n = 402), adjusted for sex and age. The 'all' model results are shown here. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. The x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 5.92×10^{-8}.

3.5.4.4.3 Periosteal Circumference at 6 years Figure 3.57, illustrates the distribution of periosteal circumference at 38% from the distal end of the tibia at 6 years (mm) in the 141 individuals in this EWAS. Surrogate variable analysis identified 37 significant surrogate variables, this is likely an overestimate stemming from small amounts of variation remaining unaccounted for by the manual model thus Manhattan plots based on the manual model have been included.

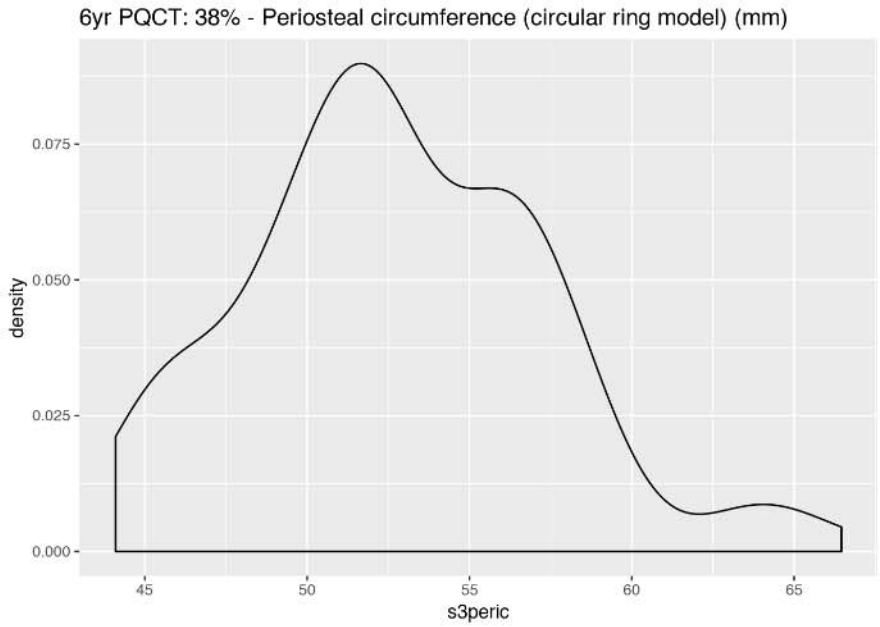

Figure 3.57: Distribution of periosteal circumference at 38% from the distal end of the tibia (mm) at 6 years of age (s3peric), adjusted for sex and age, as measured by PQCT (n = 141).

One CpG fell below the Bonferroni corrected significance threshold for an association between DNA methylation at that locus and periosteal circumference at 38% from the distal end of the tibia at 6 years (mm) adjusted for age and sex, Figure 3.58. This CpG was cg22570676 which is located at Chr19:2,527,492-2,527,494 at the GNG7 (G protein subunit gamma 7) gene. cg22570676 was significant ($p = 4.24 \times 10^{-8}$, increase of 0.370% per mm) in the 'all' model and was also Bonferroni significant in the uncorrected model. However, cg22570676 was not significant in the SVA or iSVA models suggesting that it may be attributable to batch or cell-type effects.

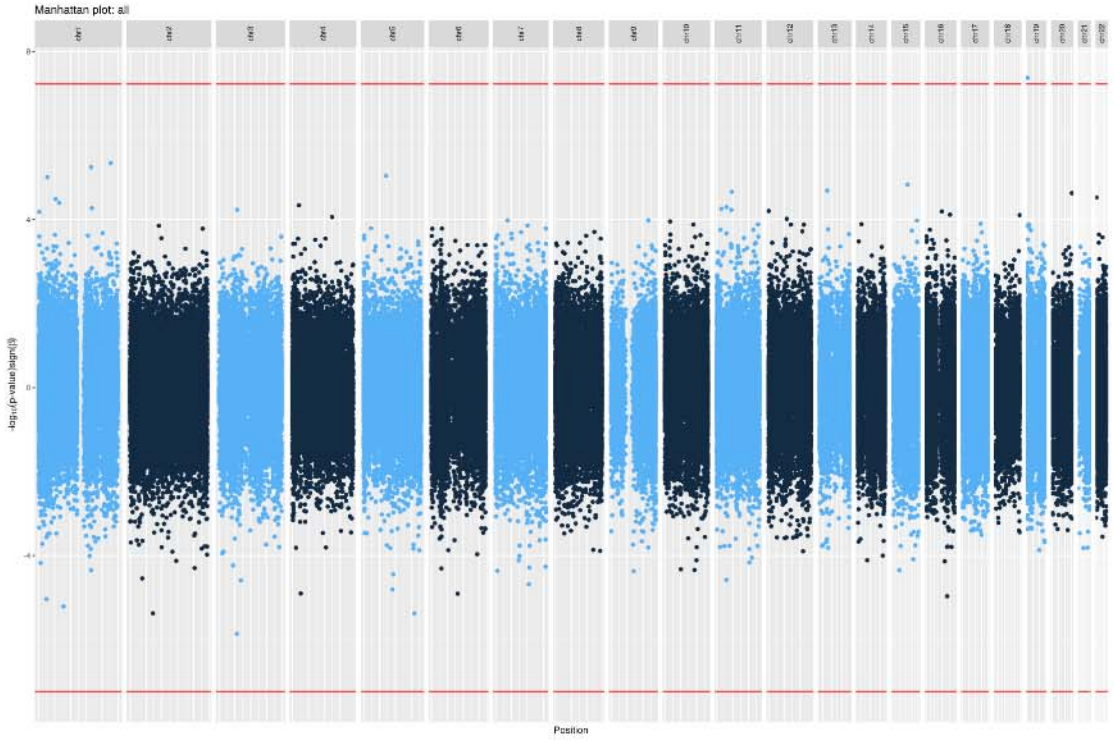

Figure 3.58: Results of EWAS for periosteal circumference at 38% from the distal end of the tibia (mm) at 6 years of age (n = 141), adjusted for sex and age. The 'all' model results are shown here. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. The x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 5.92×10^{-8}

3.5.4.4.4 Cortical Denisty at 6 years Figure 3.59, illustrates the distribution of cortical density at 38% from the distal end of the tibia at 6 years ($mg\ cm^{-3}$) in the 141 individuals in this EWAS. No CpGs fell below the Bonferroni corrected significance threshold (5.92×10^{-8}) for an association between DNA methylation at that locus and cortical density at 6 years, Figure 3.60. Surrogate variable analysis identified 37 significant surrogate variables, this is likely an overestimate stemming from small amounts of variation remaining unaccounted for by the manual model thus Manhattan plots based on the manual model.

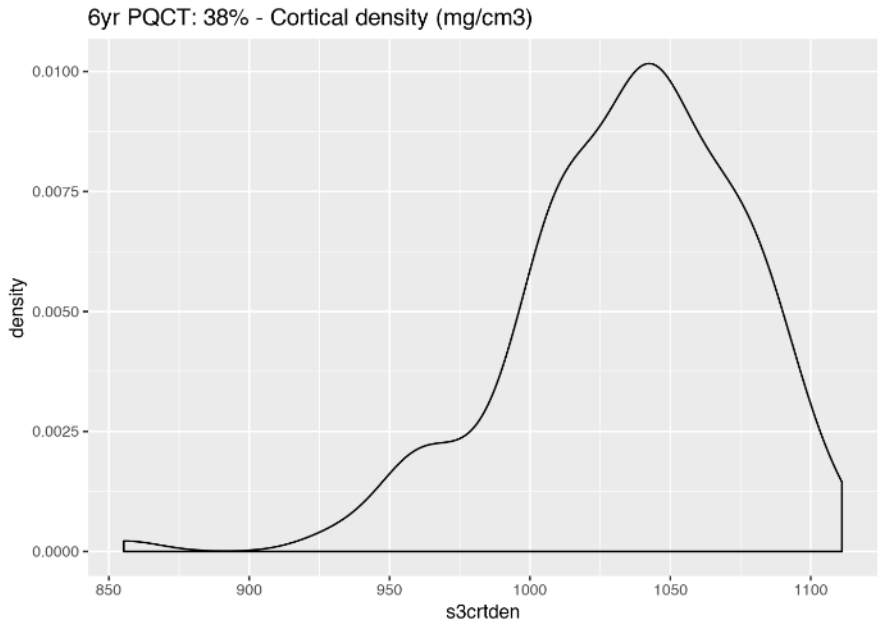

Figure 3.59: Distribution of cortical density at 38% from the distal end of the tibia ($mg\ cm^{-3}$) at 6 years of age (s3crtden), as measured by PQCT (n = 141).

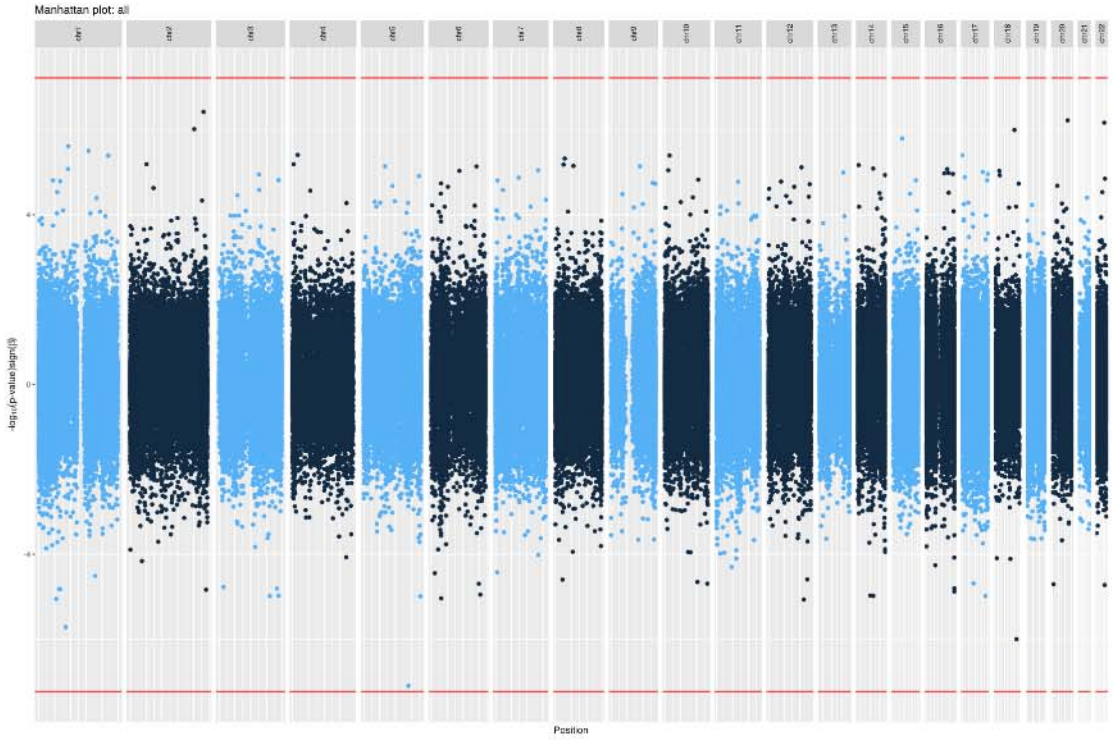

Figure 3.60: Results of EWAS for cortical density at 38% from the distal end of the tibia ($mg\ cm^{-3}$) at 6 years of age (n = 141). The 'all' model results are shown here. Bidirectional Manhattan plot on which $-log_{10}(p-value)$ is plotted on the y axis and the sign of this value represents the direction of change. The x axis represents chromosomes and position thereupon. Red line indicates the significance threshold of 5.92×10^{-8}

3.6 Discussion

EWAS for 9 outcomes were carried out across 3 sets of samples from MAVIDOS and SWS using the EPIC and 450K array platforms. No significant results were found in either the first or second phase of the MAVIDOS analysis. Two possible results for bone outcomes at 6 years were identified in the SWS data but did not remain in models including surrogate variables for possible confounding effects. The effect observed at CpG cg26559250, adjacent to the ZDHHC14 gene, in EWAS for bone mineral content at 6 years was not seen at the 8 year time point. Furthermore both this result and the finding for CpG cg22570676 and periosteal circumference were of small effect sizes 1.46% per kg, 0.370% per mm respectively. Consequently these results should be treated with considerable caution, they are in need of replication before they can be considered reliable. No Bonferroni significant DNA methylation changes at the *CDKN2A* and *RXRA* loci which have been previously associated with maternal vitamin D and bone phenotypes were identified, despite the presence of 95 and 75 probes annotated as being in the vicinity of these genes respectively.

Whilst the calculation of power for EWAS is a complex and rather understudied problem [282–284] it is possible to achieve some approximations using Cohen's methods [285]. To achieve the modest goal of 80% power for a small effect size ($r^2 = 0.02$) in a linear regression analysis (F-test) with 7 covariates at a significance level suitable for the EPIC array of $p = 5.92 \times 10^{-8}$ an n of 2607 is needed. Seven was the smallest number of covariates used in an SVA model in these analyses, models in the SWS analyses had 13 manually specified variables. When considering what is for EWAS a very large effect size ($r^2 = 0.15$) with the 13 covariates used in the SWS models it is possible to achieve 80% power for an n of 374. This is an effect size in line with the effects of smoking on DNA methylation at some loci [286,287]. This would make the two largest EWAS performed here for BMC at 8 and 6 years ($n = 408$, $n = 402$ respectively) powered only to find large effect sizes with just over 80% probability. The most generous set of parameters (80% power, $r^2 = 0.15$, $p = 5.92 \times 10^{-8}$, & 2 covariates) yield an n of 259, more realistic numbers (90% power, $r^2 = 0.02$, $p = 5.92 \times 10^{-8}$, & 13 covariates) yield an n of 3370. EWAS have identified biologically relevant changes associated with environmental exposures in DNA methylation with magnitudes of less than a single percentage point, and percentage changes in the low single digits are not uncommon in EWAS [287]. This makes all of the EWAS performed here underpowered to identify small DNA methylation changes which might be expected to occur. A collaboration with colleagues at the MRC-IEU, University of Bristol is underway to perform a meta-analysis to include these data with similar results from other cohorts in order to increase the power of these analyses. The covariates included in the models for the second phase of MAVIDOS analysis and the SWS analysis are matched to those being used by our collaborators to maximise the comparability of our results.

Given that the effect of maternal vitamin D on neonatal bone mass appears to be seasonal, with only babies born in the winter months showing statistically significant benefits of supplementation [208], it would be interesting to perform seasonally stratified EWAS were sufficient numbers available to do so with reasonable power. An initial step might be to ascertain in birth season has a significant interaction with DNA methylation state when predicting maternal vitamin D levels or bone outcomes in EWAS models.

Attempting to identify small changes in the overall DNA methylation state of complex populations of cells like blood and umbilical cord tissue which are associated with phenotypes such as circulating maternal vitamin D is a technically challenging undertaking. This work provides two candidates for further analysis for associations between DNA methylation and bone health outcomes at 6 years of age. Furthermore, these results have contributed to a larger meta analysis of other studies based on the DNA methylation array data in additional populations with greater power to detect associations with metrics of bone health.

Chapter 4

The Genomic Loci of Specific Human tRNA Genes Exhibit Ageing-Related DNA Hypermethylation

4.1 Abstract

Understanding how the epigenome deteriorates with age and subsequently impacts on biological function may bring unique insights to ageing-related disease mechanisms. As a central cellular apparatus, tRNAs are fundamental to the information flow from DNA to proteins. Whilst only being transcribed from ~46kb (<0.002%) of the human genome, their transcripts are the second most abundant in the cell. Furthermore, it is now increasingly recognised that tRNAs and their fragments also have complex regulatory functions. In both their core translational and additional regulatory roles, tRNAs are intimately involved in the control of metabolic processes known to affect ageing. Experimentally DNA methylation can alter tRNA expression, but little is known about the genomic DNA methylation state of tRNAs.

Here, we find that the human genomic tRNA loci are enriched for ageing-related DNA hypermethylation. We initially identified DNA hypermethylation of 44 and 21 specific tRNA genes, at study-wide ($p < 8.36 \times 10^{-5}$) and genome-wide ($p < 4.34 \times 10^{-9}$) significance, respectively, in 4,350 MeDIP-seq peripheral blood DNA methylomes (16 - 82 years). This starkly contrasted with 0 hypomethylated at both these significance levels. Further analysing the 21 genome-wide results, we found 3 of these tRNAs to be independent of major changes in cell-type composition

(tRNA-iMet-CAT-1-4, tRNA-Ser-AGA-2-6, tRNA-Ile-AAT-4-1). We also excluded the ageing-related changes being due to the inherent CpG density of the tRNAome by permutation analysis (1,000x, Empirical p-value $< 1 \times 10^{-3}$). We additionally explored 79 tRNA loci in an independent cohort using Fluidigm deep targeted bisulfite-sequencing of pooled DNA (n=190) across a range of 4 timepoints (aged ~4, ~28, ~63, ~78 years). This revealed these ageing changes to be specific to particular isodecoder copies of these tRNAs (tRNAs coding for the same amino acid but with sequence body differences) and included replication of 2 of the 3 genome-wide tRNAs (tRNA-iMet-CAT-1-4, tRNA-Ser-AGA-2-6). Additionally, this isodecoder-specificity may indicate the potential for regulatory fragment changes with age.

In this study we provide the first comprehensive evaluation at the genomic DNA methylation state of the human tRNAome, revealing a discreet and strongly directional hypermethylation with advancing age.

4.2 Introduction

Ageing is implicated as a risk factor in multiple chronic diseases [288]. Understanding how the ageing process leads to deteriorating biological function is now a major research focus. This field has hopes of increasing human longevity and 'healthspan' whilst ameliorating the extensive physical, social and economic costs of these ageing-related disorders [289]. Epigenetic processes, which influence or inform cell-type specific gene expression, are altered with age and are a fundamental hallmark of this progression, indeed they are arguably a hub mediating other hallmarks including stem cell exhaustion, cellular senescence, and mitochondrial dysfunction [11,30].

DNA methylation (DNAm) is the most common epigenetic modification of DNA and age-associated changes in this mark were recognised in mammalian tissues as early as 1983 [138]. Changes in DNAm with age are extensive with thousands of loci affected. Many of these changes represent 'drift' [119] arising from the imperfect maintenance of methylation state. Specific genomic regions show distinct directional changes, with loss of DNA methylation in repetitive or transposable elements [290], and gains in certain promoters, including the targets of polycomb repressor complex [155] as well as bivalent domains [153]. The advent of high-throughput DNAm arrays [291,155,167] has elucidated more detailed patterns of ageing related changes in DNAm. The identification of precise individual CpG sites that exhibit consistent changes with age enabled the construction of predictors of chronological age known as epigenetic or DNAm 'clocks' [167,168,193,292,194]. Furthermore, it was observed that 'acceleration' of this DNAm-derived measure is a biomarker of 'biological' ageing due to associations with morbidity and mortality [293,294]. In a previous investigation of ageing-related DNAm changes within common disease-associated GWAS regions, Bell et al. identified hypermethylation of the specific transfer RNA gene, tRNA-iMet-CAT-1-4 [209]. The initiator methionine tRNA possesses certain unique properties [295–297], including its capacity to be rate limiting for translation [298], association with the translation initiation factor eIF2 [299], and ability to impact the expression of other tRNA genes [300].

tRNAs are fundamental in the translation process for all domains of life and are thus evolutionarily ancient [301]. This translation machinery and the regulation of protein synthesis are controlled by conserved signalling pathways shown to be modifiable in longevity and ageing interventions [302]. The human tRNAome, comprising all of the genomic locations of tRNA genes, represents an extremely small portion of the genome [303]. There are 416 high confidence tRNA genes when additionally considering tRNA pseudo genes, nuclear encoded mitochondrial tRNA genes and possibly some closely related repetitive sequences the number of sequences closely resembling tRNAs is 610 (gtRNAdb [304]) this extended set covers <46 kb (including introns) and represents <0.002% of the human genome. Despite their small genomic footprint, and the observation that approximately half of all tRNA genes are transcribed at negligible levels if at

all [305], these genes produce the second most abundant RNA species next to ribosomal RNA [306] and are required for the production of all proteins. Mature tRNAs have an L-shaped three dimensional structure arising from a 'clover-leaf' shaped two dimensional structure comprised of three hairpin stem-loop structures (Figure 4.1).

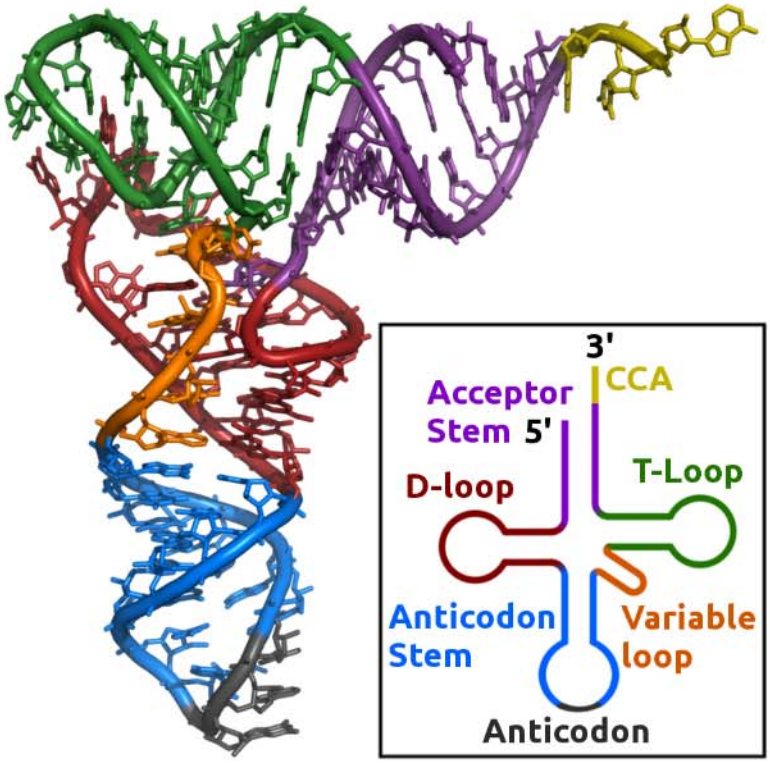

Figure 4.1: **Structure of a mature tRNA** Two and three dimensional representations of tRNA structure with matching colour coding. Adapted from the wikimedia foundation structure based on PDBID: 1ehz. tRNAs are 'charged' when an amino acid is attached at the CCA site at the 3' end.

tRNA genes are transcribed by RNA polymerase III (polIII) [307] and have type II polIII promoters which contain A and B-box internal promoter elements bound by the complex TFIIIC, followed by TFIIIB, and polIII [308] (Figure 4.2). Transcription is terminated by a simple run of Ts and proceeds in rounds of fast re-initiation where the same polIII molecule is preferentially re-used [309]. tRNA gene expression is modulated by the polIII specific transcription factor Maf1 a highly conserved factor which represses tRNA transcription [310–312]. The activity of Maf1 is modulated by the Target of Rapamycin Kinase Complex 1 (TORC1) [313], a highly conserved hub for signals that modulate ageing [314]. Several general transcription factors also influence tRNA gene expression; the tumour suppressors p53 [315] and Rb [316] both negatively regulate tRNA expression and c-Myc upregulates tRNA gene expression, all act via TFIIIB [317]. tRNAs are dysregulated in cancer, and have potential utility as prognostic markers [318]. tRNA regulation may play an important role in cancer [319]. DNAm is able to repress the expres-

sion of tRNA genes experimentally [320], in plasmid expression systems, but may also represent co-ordination with the local repressive chromatin state [321]. In addition, recent results from Gerber et al. [322] show a mechanism by which polII can function in the regulation of certain polIII transcribed loci including several tRNA genes.

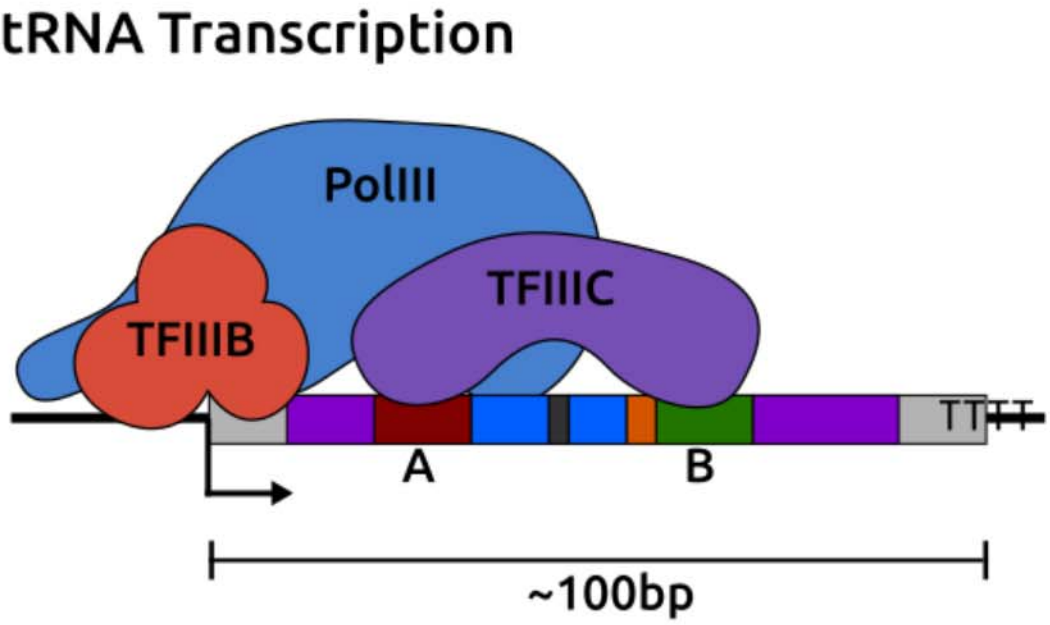

Figure 4.2: **tRNA Transcription** Cartoon representation of the RNA polIII transcription initiation complex and structure of the type II RNA polymerase III promoter. Colour coding here corresponds to that in figure 4.1 illustrating that promoter is internal as the A box corresponds approximately to the D-loop and the B box to the T-loop in the tRNA structure.

Assuming similar rates of transcription one would expect that the more frequently an amino acid is used in the exome, the more copies of that tRNA gene there would be in the genome [323]. Indeed, tRNA gene dosage is quite closely matched to amino acid usage frequency in the human exome, though the correlation is less strong for codon usage (Figure 4.3). The imperfect nature of this correlation suggests that there may be regulation of tRNA expression beyond simply having copy numbers proportionate to usage frequency.

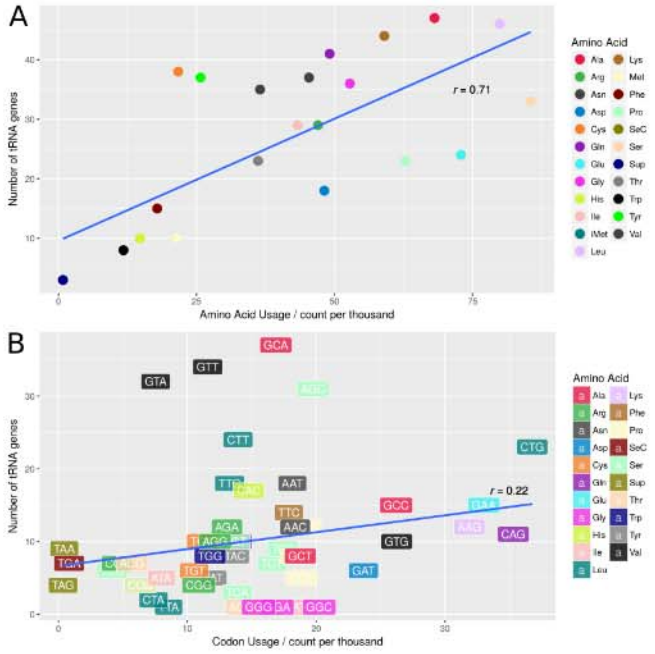

Figure 4.3: **tRNA gene copy number is imperfectly correlated with amino acid and codon usage frequency A)** Amino acid usage frequency in the human exome vs tRNA gene copy number **B)** Codon usage frequency in the human exome vs tRNA gene copy number Usage Frequency Data [324], tRNA gene count data from GtRNAdb [304].

A variety of observations have indicated that many tRNA genes are expressed in a tissue specific fashion in diverse organisms [325,326]. These include the tissue specific manifestations of diseases caused by mutations involving tRNAs and their usage [327],[328]. Though detailed characterisation of tissue specific tRNA expression patterns are still lacking. Considering the broad similarity of transcriptome codon usage across time and tissues and the purported functional equivalence in translation of tRNAs with the same anticodon, there is an open question as to the reason for tissue specificity of tRNA expression. The tissue specific expression of tRNAs whilst substantially maintaining codon ratios would require highly coordinated tRNA gene regulation to obtain similar expression levels of putatively functionally equivalent genes. This situation is strongly suggestive of as yet undiscovered function.

Additionally, beyond their core role in the information flow from DNA to protein sequence, tRNAs can fragment into numerous tRNA-derived small RNAs (tsRNAs). There are a number of types of tRNA derived small RNAs (tsRNAs) [329] which serve as signalling molecules [330] (Figure 4.4). Four main types of tsRNAs have been identified, these are 5' tRNA halves, 3' tRNA halves, and the two main types of tRNA derived fragments (tRFs), the ~18bp and the ~22bp sizes [331–333]. tRNAs and pre-tRNAs also give rise to piRNAs [334–336] and internal tRNA fragments or itRFs [337]. The terminology and ontology of tsRNAs is still stabilising as our understanding of these RNA species evolves. Some of these tsRNAs feedback on protein

synthesis by regulating ribosome biogenesis [338], others have diverse regulatory functions such as targeting transposable element transcripts [339]. The extent of the functional significance of tsRNAs is also an open question [340].

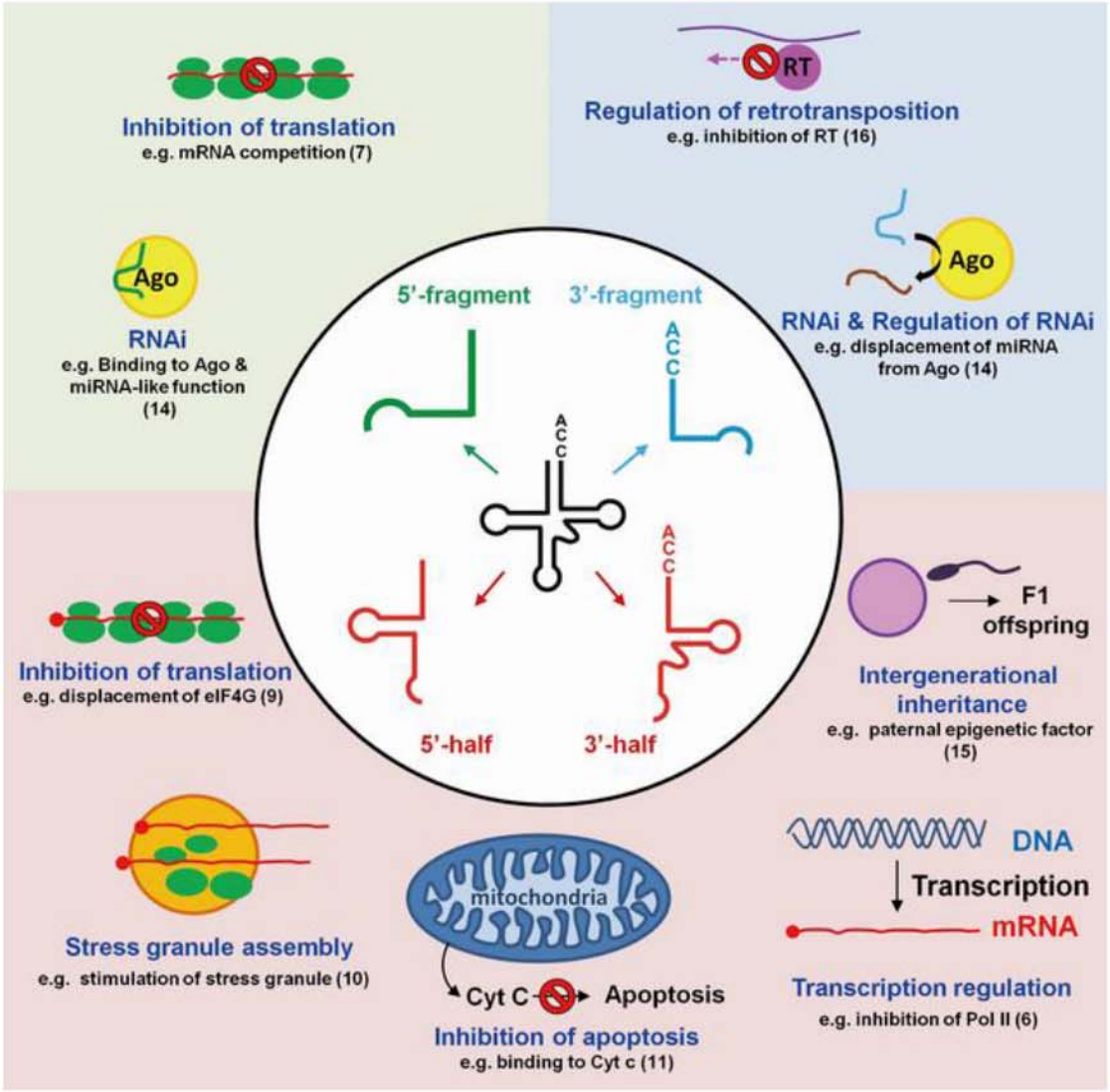

Figure 4.4: **The Types and Functions of tRNA derived small RNAs** Reproduced from Cristodero et al. [341].

Ageing is linked to core aspects of metabolic regulation, with nutrient sensing and stress response acting as major modulators of ageing. tRNAs as well as tsRNAs are integral to the regulation of protein synthesis and stress response. Protein synthesis represents a substantial proportion of total cellular energy expenditure and this fraction can vary considerably with nutrient availability [342]. Metabolic processes are also recognised to modulate the age estimates of DNAm clocks [343]. Partial inhibition of translation increases lifespan in multiple model organisms [344] and polIII inhibition increases longevity acting downstream of TORC1 [345]. Furthermore, 5' tRNA halves circulating in serum are modulated by ageing and caloric restriction [346].

DNAm is able to repress the expression of tRNA genes on methylated expression vectors [320]. Whilst the broader chromatin milieu affects tRNA transcription, tRNA genes are unusual in that they are sufficiently short to fit within a single nucleosome. tRNA genes are generally 'nucleosome free' and precise placement of the nucleosome immediately upstream of the transcriptional start site is of importance for their expression [347]. A recent review of the epigenetic regulation of the polIII transcriptome [348] noted the very limited CpG methylation data available at polIII loci. Transcription by RNA polymerase III at SINE loci is suppressed by histone methylation but not by DNA methylation [321], indicating that DNA methylation may not directly influence the expression of tRNA genes but may do so by influencing the surrounding chromatin state. Standard RNA-seq is of limited utility in examining tRNA expression due to the issues of mappability, in addition much RNA-seq data is generated with size and polyA selection methods which would exclude tRNA derived transcripts. Also, tRNAs are a major target for 'epitranscriptomic' modification, with an average of 11-13 modifications per tRNA [349,340], some of which stop polymerases from elongating or alter base pairing, creating further mapping challenges. Thus, variants on standard RNA-seq procedures have been developed [350–352]. polIII ChIP-seq has been used as a proxy for tRNA gene expression and, unlike RNA-seq based methods, generates reads from uniquely identifiable flanking regions which map to a known locus of origin [308]. This same advantage exists for the MeDIP-seq data used in this study.

tRNA gene loci may also play a role in large scale genome organisation. tRNA gene clusters act as insulators [353], and have extensive long-range chromatin interactions with other tRNA gene loci [354]. The coordinated transcription of tRNAs at subnuclear foci may represent an organising principle for 3D-chromatin by providing spatial constraints. In both budding and fission yeast tRNA genes localise to the nucleolus. In fission yeast, a subset of B-box sequence elements are bound by TFIIIC and not polIII serving as chromatin anchors to the nuclear periphery and acting as boundaries between euchromatic and heterochromatic regions [355]. It is unclear to what extent tRNA genes may play similar roles in large scale chromatin organisation in other organisms.

In this study ageing-related changes in the epigenetic DNA methylation state of the entire tRNAome were directly investigated, facilitated by the availability of a large-scale MeDIP-seq dataset. Arrays poorly cover this portion of genome. The 450k and EPIC arrays have 110 robust probes covering 84 tRNAs and 129 robust probes covering 89 tRNAs respectively, thus even the latest EPIC arrays cover <15% of the tRNA genes, with robust probes, and in total only ~4.7% of all the tRNA gene CpGs [223].

tRNA genes sit at the heart not only of the core biological process of translation but at a nexus of signalling networks operating in several different paradigms, from small RNA signalling to large scale chromatin organisation [354]. In summation tRNA biology, protein synthesis, nutrient sensing, stress response and ageing are intimately interlinked.

4.3 Primary questions

Are there additional loci which show the age-related DNA methylation changes observed in tRNA-iMet-CAT-1-4?

Is there an overall pattern of age-related hypermethylation in the tRNAome?

Can age-related DNA methylation changes at tRNA loci be validated with different technology to MeDIP-seq in which was initially identified?

Can age-related DNA methylation changes at tRNA loci be replicated in a cohort independent of that in which it was originally identified?

Can age-related DNA methylation changes at tRNA loci be observed in other organisms?

Can tissue specific differences in age-related DNA methylation changes at tRNA loci be observed?

4.4 Methods

Code for analyses performed in this chapter can be found here: https://github.com/RichardJActon/tRNA_paper_code

4.4.1 Participants

4.4.1.1 MeDIP-seq DNA methylomes

Participants in the 'EpiTwins' study are adult volunteers from the TwinsUK Register. The participants were aged between 16 and 82 years, with a median of age at sampling of ~55 years for 4350 methylomes taken from 3652 unique participants some of whom where sampled more than once. (cohort profile [232]). The majority of samples 4054 are female and 270 are male. 3001 have full blood counts. These individuals originate from 1933 unique families. There are 1234 monozygotic (MZ) twin pairs (2468 individuals), and 458 dizygotic (DZ) twin pairs (916 individuals). See also the participants section in general Methods 2.2.1. Ethics for the collection of these data were approved by Guy's & St Thomas' NHS Foundation Trust Ethics Committee (EC04/015—15-Mar-04) and written informed consent was obtained from all participants.

4.4.1.2 Targeted Bisulfite sequencing

Participants for our targeted bisulfite sequencing of select tRNA loci were drawn from two studies. Samples from participants aged 4 and 28 years are from the MAVIDOS [207] study, a randomised controlled trial of the effect of maternal vitamin D supplementation of the bone health of their offspring. Participants aged 63 and 78 years are from the Hertfordshire cohort study, which consists of individuals originally recruited in the 1980's for a study of the impact of birth weight on heart attack risk because high quality records of their birth weights and other useful data had been recorded in the 1920s & 30s [356]. Due to a limited number of available samples, the two 4 year old pools contained DNA from 20 individuals each, with all other pools having 25 contributing individuals. Pool 1, the first 4 year old pool used DNA from all male samples, with all other pools using all female samples. Thus, the total number of participants was 190 (see Table 4.1). Samples from the 28 year old time point are all from pregnant women at ~11 weeks gestation.

Richard Acton selected the loci to target, designed the primers and selected the samples to go in each pool. Selected DNA samples were pooled by Nikki Graham at the institute of developmental sciences at the university of Southampton. With Fluidigm amplification and sequencing performed at Bart's and the London genome centre, the genomics core facility at Queen Mary, University of London. The targeted BiS-sequencing data is available at: https:

//www.ncbi.nlm.nih.gov/bioproject/PRJNA635108.

4.4.2 tRNA Gene coordinates and other annotation information

Genomic coordinates of all tRNA genes including some likely pseudogenes were downloaded from GtRNAdb [304]. The 2 tRNAs located in chr1_gl000192_random are tRNA-Gly-CCC-8-1 & tRNA-Asn-ATT-1-2 (The tRNA annotation used is provided in appendix 6.6). The 213 probes overlapping tRNA genes were derived from intersecting the tRNA gene annotation data from gtRNAdb with the Illumina 450k array manifest annotation for the hg19 genome build using bedtools v2.17.0 [231]. 107 tRNA genes were excluded from blacklisted regions of hg19 [357], that is to say regions of the genome which have "anomalous, unstructured, or high signal in next-generation sequencing experiments independent of cell line or experiment."

4.4.2.1 tRNA Gene Clustering

To explore the genomic spatial distribution of the tRNA genes, the tRNA loci were clustered by grouping together all tRNAs within 5Mb of one another using the `bedtools merge` tool v2.17.0 [231]. A command of the form: `bedtools merge -c 4 -o collapse -d <N> -i hg19-tRNAs.bed` was used, where `<N>` is the binsize. The binsize was varied and 5Mb was selected as it is at approximately this size that number of clusters with more than one tRNAs exceeds the number of singleton tRNAs (Figure 4.5). The further requirements that these groupings contain at least 5 tRNA genes with a density of at least 5 tRNA genes per Mb were added for these groups to be considered clusters.

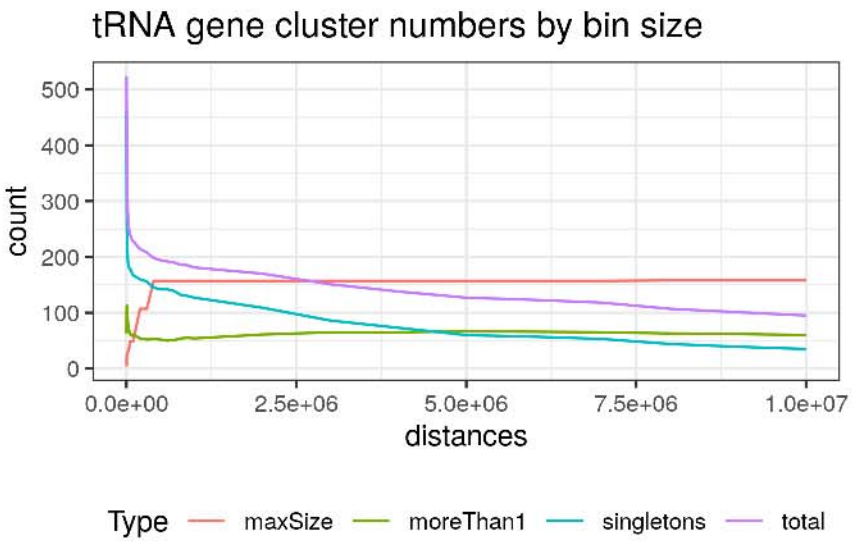

Figure 4.5: **tRNA gene cluster numbers at different bin sizes** total: total number of tRNA clusters. singletons: number of tRNAs in clusters alone. moreThan1: number of tRNAs in clusters with more than one tRNA. maxSize: the number of tRNAs in the largest cluster.

4.4.2.2 tRNA gene mappability assessment

To assess the mappability of tRNA gene regions a value for mappability score density was computed to facilitate comparisons of regions of the genome. Mappability score density is computed as the area under the encode mappability tracks [358] over the length of the region.

4.4.3 DNA methylome data

4.4.3.1 TwinsUK MeDIP-seq methylomes

The Methylated DNA Immunoprecipitation sequencing (MeDIP-seq) data was processed as previously described [209,107] and detailed in Methods 2.2. These processed data are available from the European Genome-phenome Archive (EGA) (https://www.ebi.ac.uk/ega) under study number EGAS00001001910 and dataset EGAD00010000983 and were generated by BGI Shenzen for TwinsUK. The dataset used in this work consists of 4350 whole blood methylomes with age data. This data consists of a matrix of RPM (reads per million base pairs) values for overlapping 500 base pair windows of the genome with a 250bp slide. This is approximately 12 million windows for 4350 samples, or ~52 billion data points.

4.4.3.2 Analysis of DNA methylome data for Significant Ageing-related changes

All analysis was performed in R/3.5.2. Linear models were fitted to age using the MeDIP-seq DNA methylome data, as quantile normalised RPM scores at each 500bp window. Quantile normalisation was performed with the `qqnorm` R function with the theoretical quantiles of the RPM values at each window used in subsequent analysis.

Models were fitted with:

1. Uncorrected, simply modelling age from DNA methylation
2. Batch information as a fixed effect
3. Blood cell-type counts for neutrophils, monocytes, eosinophils, and lymphocytes as fixed effects
4. Batch and Blood Cell counts as fixed effects

- Models 1 & 2 were fitted on the full set of 4350 as batch information was available for all samples but blood cell count data was only available for a subset of 3001 methylomes.
- Models 1 & 2 fitted in the n=3001 subset were similar to those fitted in the complete set of 4350.
- Models 3 & 4 were fitted in the n=3001 subset with full covariate information and sets of significant tRNAs identified at study-wide and genome wide levels in model 4 were used in subsequent analyses.

Models were also fitted for two unrelated subsets created by selecting one twin from each pair (Monozygotic or Dizygotic), yielding sets with n = 1198 & 1206 DNA methylomes, which time-point was used when more than one sample was available for an individual was selected arbitrarily. One additional model was fitted for longitudinal analysis, samples were selected by identifying individuals with a DNA methylome at more than one time point and filtering for only those with a minimum of 5 years between samples. This yielded 658 methylomes from 329 individuals with age differences of 5-16.1 years, median 7.6 years. Models for this set included participant identifier as a fixed effect in addition to blood cell counts and batch information.

4.4.3.3 Permutation Analysis for Enrichment with Age-related Changes

Permutation analysis was performed to determine whether the CpG distribution of sets of the tRNAome was the principle driver of the ageing-related changes observed. Windows overlapping tRNAs have a higher proportion of windows with a greater CpG density than their surrounding sequences (see Figure 4.6). CpGs residing within moderate CpG density loci are the most dynamic in the genome [86] and CpG dense CpG island regions include specific ageing-related changes [155,153,209]. For comparison the permutation was also performed in the CGI regions from the Polycomb group protein target promoters in Teschendorff *et al.* [155] and bivalent loci

from ENCODE ChromHMM 'Poised Promoter' classification in the GM12878 cell-line [359]. A random set of 500bp windows representing an equivalent CpG density distribution of the feature set in question were selected from the genome-wide data. Above a certain CpG density there are insufficient windows to sample without replacement within a permutation. Furthermore, above ~$\geq$ 18% CpG density CpG Islands become increasingly likely to hypomethylated [360]. Therefore, all windows with a CpG density of $\geq$ 18% (45 CpGs per 500bp) were grouped and sampled from the same pool. i.e. a window overlapping a tRNA gene which had a 20% density could be represented in permutation by one with any density $\geq$ 18%. This permutation was performed 1,000 times to determine an Empirical p value by calculating the number of times the permutation result exceeded the observed number of significant windows in the feature set. $Empirical\ p-value = \frac{r+1}{N+1}$, where r is the sum of significantly hypermethylating windows in all permutations and N is number of permutations [361].

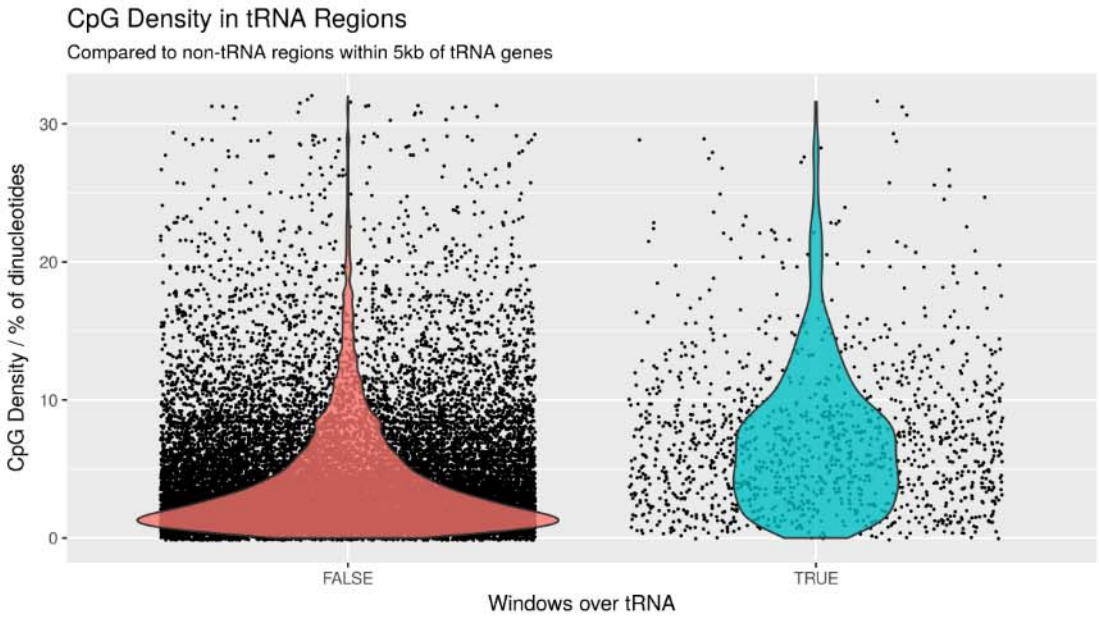

Figure 4.6: CpG Density is higher in windows directly overlapping tRNA genes compared to that of non-tRNA overlapping windows in arbitrary flanking sequences (+/-5kb). This difference in CpG density between tRNA loci and other regions of the genome is a potential source of bias if age related DNA methylation changes vary with CpG density which they may as baseline DNA methylation levels also vary with CpG density.

4.4.3.4 Neonate and Centenarian Whole Genome Bisulfite Sequencing

DNA methylation calls were downloaded from GEO: GSE31263 and intersected with tRNA genes using bedtools v2.17.0 [231].

Table 4.1: Summary information on participants in each pool.

Pool	Mean Age	Sex	Min Age	Max Age	n
Pool 1	4.07	Male	3.99	4.38	20
Pool 2	4.09	Female	3.99	4.36	20
Pool 3	28.07	Female	25.87	29.80	25
Pool 4	28.23	Female	26.05	30.01	25
Pool 5	63.40	Female	62.80	63.80	25
Pool 6	63.26	Female	62.70	63.70	25
Pool 7	77.96	Female	75.50	80.50	25
Pool 8	77.22	Female	74.40	80.10	25

4.4.3.5 Sample pooling and EPIC array

An Illumina Infinium DNA methylation EPIC array ((C) Illumina) and targeted bisulfite sequencing of select tRNA gene loci were performed. Here DNA was extracted from whole blood and combined into 8 pools from unrelated individuals at 4 time-points with 2 pools at each time-point. The individual were age matched at approximately 4, 28, 63, and 78 year timepoints (Table 4.1) Using the EPIC array DNAm age was estimated using the Horvath clock [168] and blood cell-type-composition with the Houseman method [272].

4.4.3.6 Targeted Bisulfite Sequencing

tRNA loci were selected for targeted sequencing which exhibited changes and DNAm with age along with closely related tRNAs in which changes were not observed. Primer design was performed using 'methPrimer' [237] (Primer sequences are provided in appendix 6.7). A total of 84 tRNA loci were targeted in 2 rounds of sequencing, 79 subsequently generated reliable results post-QC. The targeted tRNAs covered a total of 723 CpGs with a median of 8 CpGs per tRNA (range 1-13), data passing QC was generated for 458 CpGs, median 6 (range 1-9) per tRNA.

Initial QC for readcount, base call quality and base composition was caried out with `FastQC` [229] and `multiqc` [362] for combined visualisation of QC outputs. Quality based trimming was performed with `Trim Galore` [363] and target specific primer trimming with `cutadapt` [364] and custom `perl 5` scripts. Alignment and methylation calling was performed with `Bismark` (v0.20.0) [236] making use of `bowtie2` [365].

The alignment was performed against both the whole hg19 genome and just the tRNAome +/- 100bp to assess the possible impact of off-target mapping. Mapping to the whole genome did produce purported methylation calls at a larger number of loci than mapping just to the tRNAome (683,783 vs 45,861 respectively). Introducing a minimum coverage threshold of 25 reads dramatically reduced this and brought the number of sites into line with that in the tRNAome set (36,065 vs 33,664 respectively) suggesting a small number of ambiguously mapping

reads. All subsequent analysis was performed using the alignment to just the tRNAome with a minimum coverage of 25 reads.

Pairwise differential methylation analysis of the tRNA genes at the different time points was performed using `RnBeads` [366] with `limma` [367] and a minimum coverage of 25 reads. Linear regression was used to predict age from DNA methylation at the targeted tRNA sites, permitting us to compare rates of increase with age. For the linear regression, only CpG sites with more than 25 reads mapped to the regions of the genome targeted for amplification were used.

4.4.3.7 TwinsUK Illumina 450k array methylomes

Illumina Infinium DNA methylation 450k arrays ((C) Illumina) were also performed on TwinsUK participants, in 770 Blood-derived DNA samples which had matched MeDIP-seq data (Methods 2.1). These data were available for this analysis in a pre-processed form, Methylation 'beta' values subject to beta-mixture quantile normalisation (BMIQ) as previously described [209,107]. Cell-type correction was performed using cell-count data and the following model: `lm(age ~ beta + eosinophils + lymphocytes + monocytes + neutrophils)`.

4.4.4 Chromatin Segmentation Data

Epilogos chromatin segmentation data [368,359] was downloaded for the tRNA gene regions +/- 200bp from https://explore.altius.org/tabix/epilogos/hg19.15.Blood_T-cell.KL.gz using the `tabix` utility. The data used was the 'Blood & T-cell' 15 State model based on segmentation of 14 cell-types. This data was manipulated and visualised with `R` and `ggplot2`. tRNAs were assigned a predominant chromatin state base on the state with the highest score over that tRNA gene.

4.4.5 Isolated Blood Cell Type Specific Data

Data from 7 cell-type fractions from 6 Male individuals was downloaded from GSE35069 [369] using `GEOquery` [370]. Five of the 6 top age hypermethylating tRNAs are covered by this array dataset.

4.4.6 Cancer and Tissue Specific Methylation Data

Data was downloaded from the TCGA (The Cancer Genome Atlas) via the GDC (genomic data commons) data portal [371] using the `GenomicDataCommons` `R` package. Data from foetal tissue [372,373] was downloaded from GEO (GSE72867, GSE30654). From the TCGA, samples were selected for which DNAm data was available from both the primary site and normal solid

tissue, and for which an approximate age could be inferred (within one year). Selecting those probes overlapping tRNA genes yielded 73,403 data points across 19 tissues with an age range of 15-90yrs (median 63.4) (A complete list of TCGA sample used is provided in appendix 6.11)

4.4.7 Assaying tRNA expression in blood with MINTmap

Small RNA-seq data from sorted blood cell fractions [374] was used. (GSE100467) and the MINTmap [337] tRNA fragment alignment tool. This dataset covered 42 individuals aged 21-63. A customised MINTmap reference designed to include only fragments which unambiguously map to a single tRNA gene locus and which overlap the 5' or 3' end of the genomic tRNA sequence by at least one base with no mismatches was produced. This reference is intended to capture pre-tRNAs prior to processing and CCA addition operating under the assumption that the levels of pre-tRNAs will be informative about the amount of transcription taking place at the tRNA loci. This approach provides at most a many to one mapping of tRNA fragments to a tRNA genes.

Assaying the expression of tRNA genes presents numerous difficulties [331], and usually requires variants on standard RNA-seq protocols. This custom MINTmap reference build yielded 383 fragments mapping to 92 distinct tRNA loci in this data. To control quality, only fragments with more than 20 total instances in the dataset, and present in more than 20 individuals were considered.

The maximum length of a fragment was limited to 50nt, due to the read length of the small RNA-seq data.

4.4.8 Mouse RRBS Analysis

Methylation calls and coverage information resulting from RRBS performed by Petkovich *et al.* [375] were downloaded from GEO using `GEOquery` [370] GSE80672. These data from 152 mice covered 68 tRNA and 436 CpGs after QC requiring >50 reads per CpG and >10 data points per tRNA. 5 tRNAs were excluded for being located within blacklisted regions of mm10 [357]. After QC there were 58 tRNA genes and 385 CpGs. Simple linear modelling to predict age (in months) from methylation level at each tRNA and each CpG were performed in `R`.

4.5 Results

4.5.1 DNA Methylation of Specific tRNA Gene Loci Changes with Age

Due to tRNAs critical role in translation and evidence of their modulation in ageing and longevity-related pathways, these genes were interrogated for evidence of ageing-related DNA methylation changes. The discovery set was a large-scale peripheral blood-derived DNA methylome dataset comprising of 4,350 samples (see Figure 4.7).

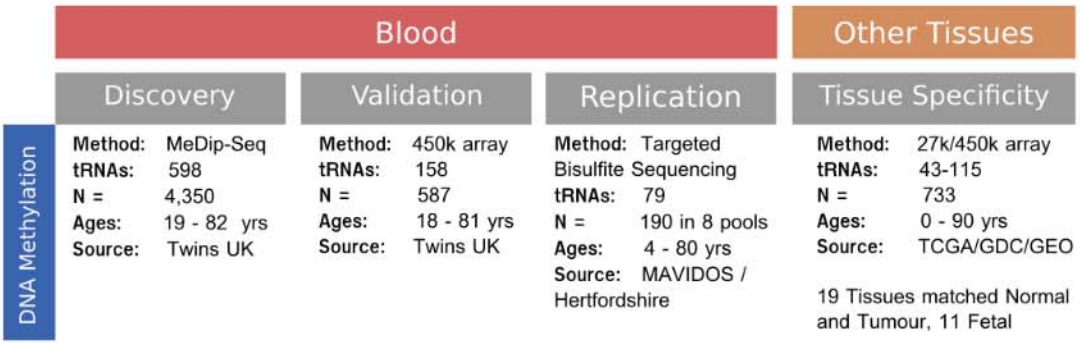

Figure 4.7: **Study Structure** tRNAs differentially methylated with age initially identified in MeDIP-seq, validated (where covered in 450k array) and replicated in targeted bisulfite sequencing of pooled samples. Tissue specificity of these effects was explored in TCGA and foetal tissue data.

This sequencing-based dataset had been generated by Methylated DNA Immunoprecipitation (MeDIP-seq) [127], which relies on the enrichment of methylated fragments of 200-500 bp to give a regional DNAm assessment (500 bp semi-overlapping windows, see Methods 4.4.3.1). In total the extended human tRNAome is comprised of 610 tRNAs and closely related sequences (gtRNAdb)(see Figure 4.8), though only 492 are autosomal and do not reside in blacklisted regions of the genome [357].

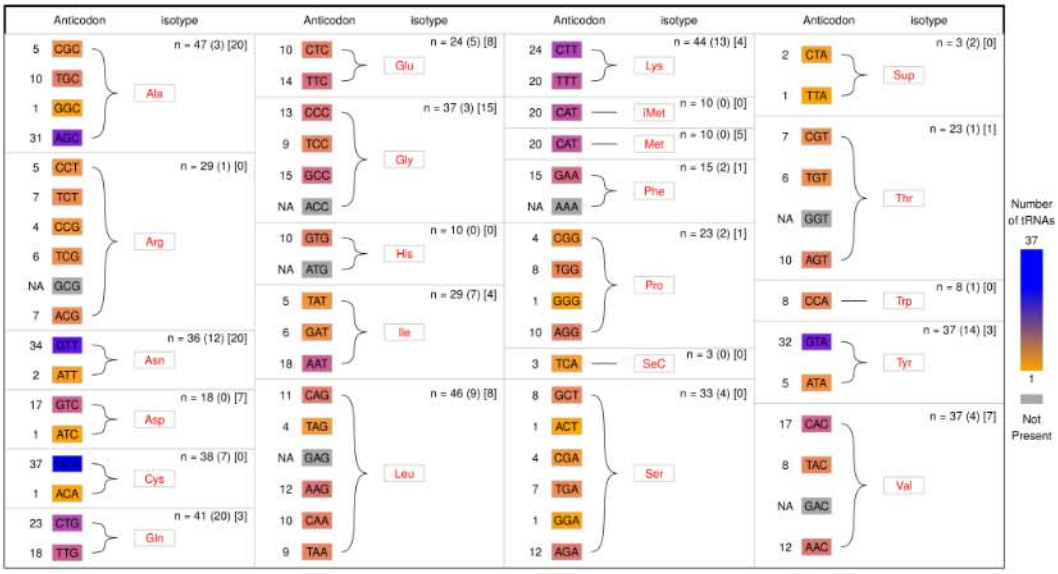

Figure 4.8: **The genetic code as instantiated in the human tRNAome.** The triplet genetic code leads to the incorporation of specific amino acids into an elongating protein via corresponding tRNAs. n is the number of tRNA genes which encode a given amino acid, the number in parentheses is how many of those may be pseudogenes based on their tRNAscan score [376], and the number in square brackets is the number in blacklisted regions [357]. There are a total of 610 tRNAs and closely related sequences in GtRNAdb [304], 416 of which are high confidence tRNAs, 116 of which are potential pseudogenes, and 107 are in blacklisted regions [357]. Notably 7 of the 61 non-STOP codons are missing from the human tRNAome therefore these codons are handled by wobble base matching (*e.g.* GCG Arg, ACC Gly). Also of note are the suppressor and selenocysteine tRNAs. The 20 methionine tRNAs are split equally between initiator methionine and internally incorporating methionine tRNAs, which are structurally distinct [299]. There are also 23 nuclear encoded mitochondrial tRNAs.

Due to the small size of these tRNAs (60-86bp, median 73bp, excluding introns which are present in ~30 tRNAs with sizes from 10-99bp, median 19bp), this fragment-based method enabled a robust examination of the epigenetic state of these highly similar sequences. This was supported by a mappability assessment. The median mappability score density for the tRNAome was 0.90 for 50mers when considering tRNA genes $\pm 500bp$ reflecting the regional nature of the MeDIP-seq assay Methods 4.4.2.2. In contrast the 50mer mappability density is 0.68 for the tRNA gene sequences alone without flanking sequences. Excluding the flanking region is representative of the mappability of reads generated using a technique such as whole-genome bisulfite sequencing. This is because there is no IP fragment so reads mapping to adjacent more mappable sequences do not convey information about the methylation state of sites in the same fragment but to which it is harder to map (see Figures 4.9 & 4.10).

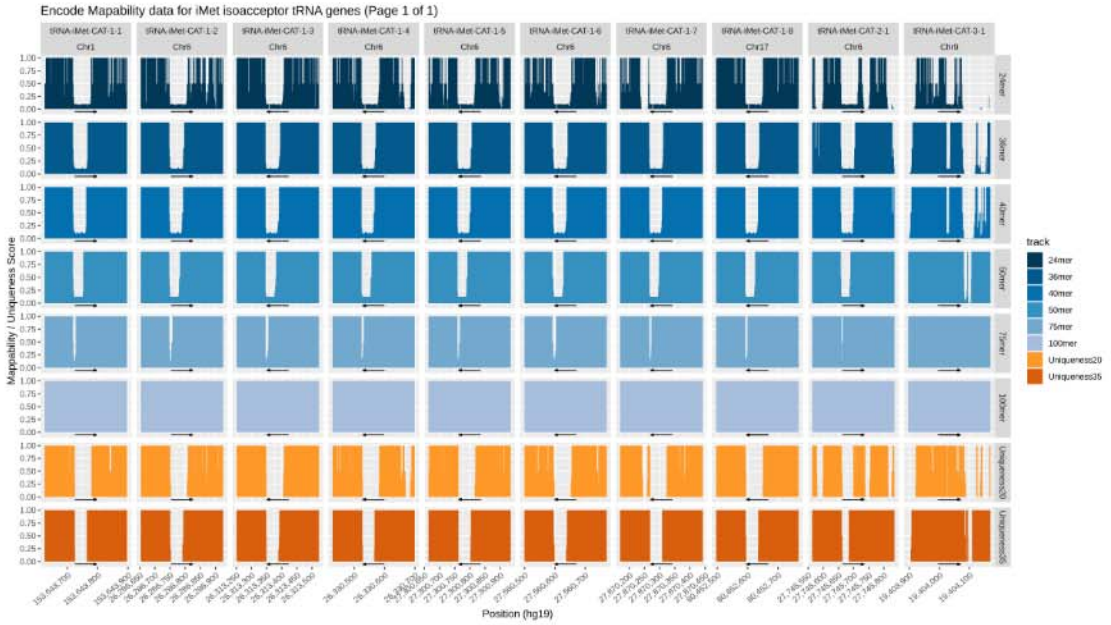

Figure 4.9: Example of mappability data from the encode mappability tracks [358] for the initiator methionine tRNA genes.

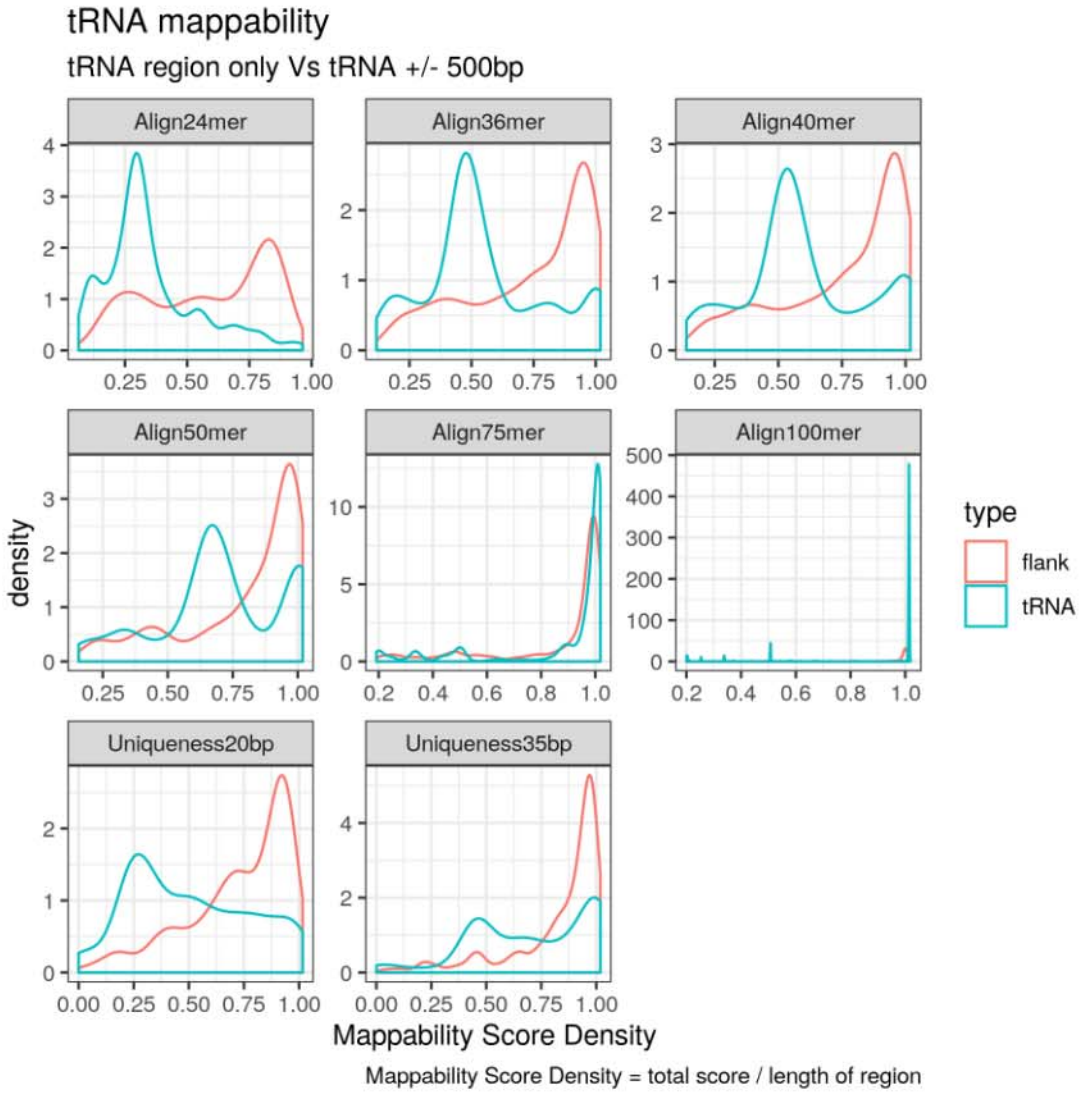

Figure 4.10: Mappability score density of the tRNAome increases with read length and is greater when flanking regions ($\pm 500bp$) are included. Mappability score density is computed as the area under the encode mappability tracks [358] over the length of the region.

21 genome-wide significant and 44 study-wide significant results were identified via linear regression, ($p < 4.34 \times 10^{-9}$ and 8.36×10^{-5}, respectively; see Methods 4.4.3.2, batch corrected n=4350). Study-wide significance was calculated conservatively using the Bonferroni method for all 598 autosomal tRNAs. Two distinct significance thresholds are used here, the study-wide correcting for the number of tRNA loci and the genome-wide correcting for the total number of loci in the genome. This is intended to permit the framing of statistical questions about the tRNA genes relative to other tRNAs genes (study-wide) or relative to the rest of the genome (genome-wide). There was a strong directional trend with all results at both significance levels being due to increases in DNA methylation. Age-related changes in cell type proportion are strong in heterogeneous peripheral blood, and include a myeloid skew, loss of naive T cells and

Table 4.2: Study-wide significantly hypermethylating tRNAs in blood cell-type and batch corrected model MeDIP-seq. With Corresponding results in Twins UK 450k array (blood cell-type corrected) and targeted bisulfite sequencing results. Age models for array and targeted BiS-seq were calculated using all probes / CpGs overlapping the indicated tRNA gene. 'Slope' corresponds to the beta value for methylation in the linear model, orange colouring indicates hypermethylation and blue hypomethylation. p-values coloured such that low values are dark blue and high values are yellow. blank grey cells indicate missing data.

		MeDIP		450k Array		Targeted BiS-seq	
tRNA	Window	Slope	p-value	Slope	p-value	Slope	p-value
tRNA-Gln-CTG-7-1	Chr1:147,800,750-147,801,250	0.84	2.60e-05				
tRNA-Glu-TTC-1-1	Chr2:131,094,500-131,095,000	1.11	4.64e-09				
	Chr2:131,094,250-131,094,750	1.00	1.12e-07				
	Chr2:131,094,750-131,095,250	1.00	3.28e-07				
tRNA-His-GTG-1-2	Chr1:146,544,500-146,545,000	0.92	1.38e-06				
tRNA-His-GTG-2-1	Chr1:149,155,750-149,156,250	1.05	2.98e-08				
	Chr1:149,155,500-149,156,000	0.83	1.37e-05				
tRNA-Ile-AAT-10-1	Chr6: 27,251,500- 27,252,000	1.07	1.45e-08			1.30	1.22e-03
	Chr6: 27,251,750- 27,252,250	0.90	1.86e-06			1.30	1.22e-03
tRNA-Ile-AAT-4-1	Chr17: 8,130,000- 8,130,500	1.19	2.98e-10	19.63	8.92e-06	-0.74	6.88e-04
	Chr17: 8,130,250- 8,130,750	0.77	3.99e-05	19.63	8.92e-06	-0.74	6.88e-04
tRNA-Ile-TAT-2-2	Chr6: 26,987,750- 26,988,250	0.97	7.25e-07	4.16	1.17e-02	-0.60	3.84e-01
tRNA-iMet-CAT-1-4	Chr6: 26,330,500- 26,331,000	1.28	2.83e-11	13.01	6.07e-06	4.54	9.35e-04
	Chr6: 26,330,250- 26,330,750	1.13	2.89e-09	13.01	6.07e-06	4.54	9.35e-04
tRNA-Leu-TAG-2-1	Chr14: 21,093,250- 21,093,750	1.04	9.38e-08			2.49	8.77e-03
	Chr14: 21,093,500- 21,094,000	0.94	8.50e-07			2.49	8.77e-03
tRNA-Pro-AGG-2-2	Chr6: 26,555,500- 26,556,000	1.04	3.97e-08				
	Chr6: 26,555,250- 26,555,750	1.01	9.58e-08				
tRNA-Ser-ACT-1-1	Chr6: 27,261,250- 27,261,750	0.97	3.53e-07			0.66	1.45e-01
tRNA-Ser-AGA-2-6	Chr17: 8,129,750- 8,130,250	1.21	1.16e-10	20.87	6.72e-05	0.62	4.28e-02
	Chr17: 8,130,000- 8,130,500	1.19	3.03e-10	20.87	6.72e-05	0.62	4.28e-02
tRNA-Ser-TGA-2-1	Chr6: 27,513,000- 27,513,500	0.90	3.58e-06	87.21	1.38e-04	-0.25	5.74e-01
tRNA-Val-AAC-1-2	Chr5:180,590,750-180,591,250	0.91	3.28e-06				
tRNA-Val-AAC-4-1	Chr6: 27,648,500- 27,649,000	1.07	1.25e-08	40.06	9.90e-03		
	Chr6: 27,648,750- 27,649,250	0.95	4.31e-07	40.06	9.90e-03		
tRNA-Val-CAC-2-1	Chr6: 27,247,750- 27,248,250	0.85	2.33e-05	59.16	5.05e-06		

increases in senescent cells [377]. A subset of 3 genome-wide and 16 study-wide significant hypermethylation results remained significant even after correcting for potential cell-type changes by including lymphocytes, monocytes, neutrophils and eosinophil cell count data (n=3001, Listed in Table 4.2, Red in Figure 4.11). The 3 genome wide significant tRNAs are: tRNA-iMet-CAT-1-4, tRNA-Ile-AAT-4-1, and tRNA-Ser-AGA-2-6. tRNA-iMet-CAT-1-4 is located on chromosome 6. tRNA-Ile-AAT-4-1 and tRNA-Ser-AGA-2-6 are neighbours and are located on chromosome 17 within the 3' UTR of *CTC1* (CST Telomere Replication Complex Component 1). Going forward these most robustly corrected sets of 3 and 16 tRNA genes are referred to as the genome-wide and study-wide significant tRNA genes respectively.

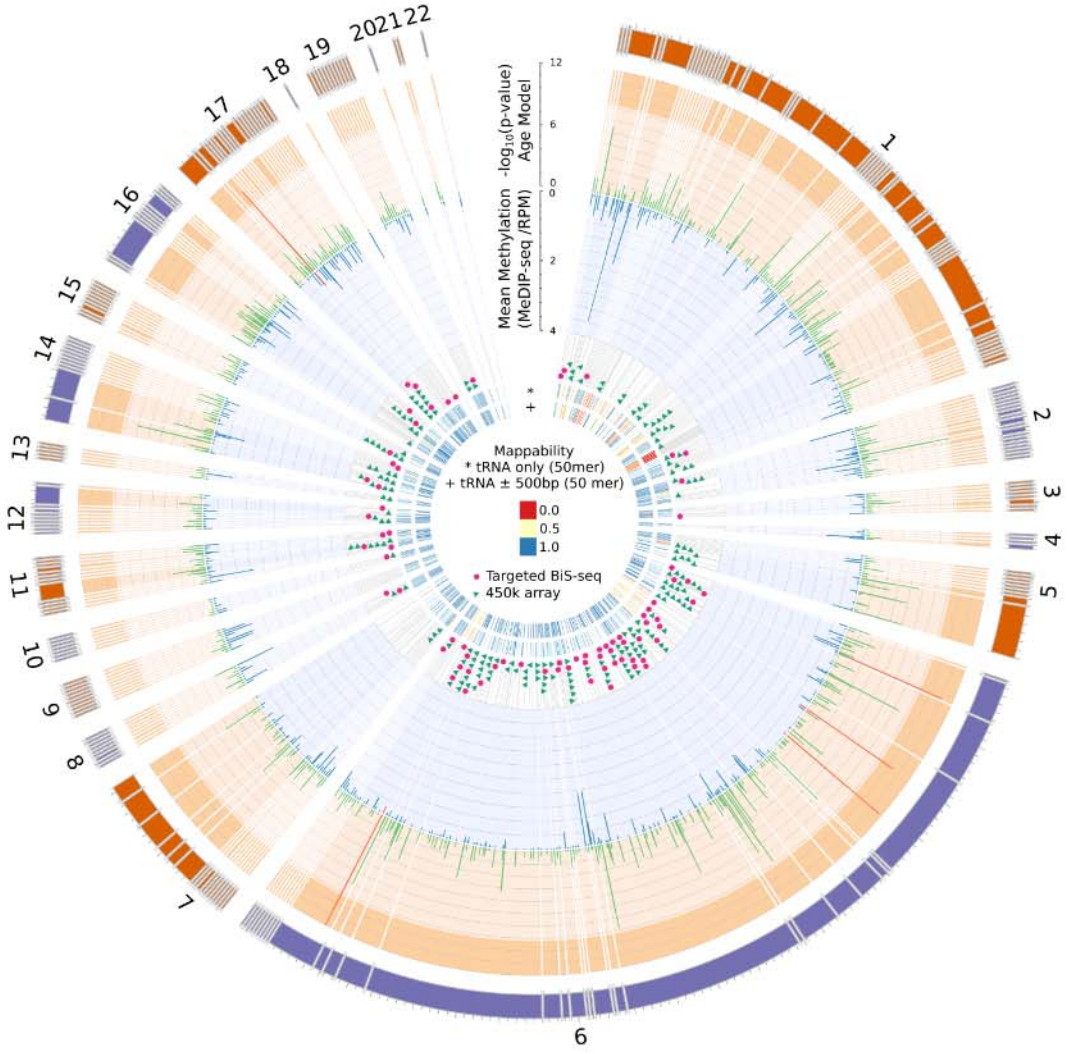

Figure 4.11: Human tRNAome overview. From the outside in: Chromosome ideograms scaled by the number of tRNA genes (total = 598), as excludes chromosome X (10), Y (0) and contig chr1_gl000192_random (2; see Methods 4.4.2). tRNA genes within 20kbp of one another are grouped with breaks inserted between these clusters. Radial grey lines represent the location of tRNA genes in the genome. $-log_{10}(p-value)$ for the blood cell-type and batch corrected age model are shown for each window overlapping a tRNA gene in green. Mean methylation across all samples (n=3001) in RPM (reads per million base pairs) is shown in blue. Genome-wide significant cell-type & batch corrected ($p < 4.34 \times 10^{-9}$) tRNAs show in red. The 158 Loci covered by 213 probes on the 450k array which directly overlap a tRNA gene are shown with green triangles. The 84 loci targeted for bisulfite sequencing in this study are indicated in magenta. Mappability score density is computed as the area under the encode mappability tracks [358] over the length of the region.

Due to the related nature of these twin samples, these data were also analysed in two subsets

of n=1198 & 1206 by selecting one twin from each pair into the separate sets. This analysis also included correction for Batch and Blood Cell counts. Whilst in these smaller datasets no tRNAs were genome-wide significant, 5 and 7 tRNA genes, respectively, reached study-wide significance, with consistent hypermethylation. In these sets 5/5 and 6/7 of these were present in 16 study-wide significant tRNA genes.

Furthermore, a subset of samples with longitudinal data were examined (n=658 methylomes from 329 individuals, median age difference 7.6 yrs). At the nominal significance threshold ($p < 0.05$) this yielded a split of 41 hypermethylating tRNA genes and 22 hypomethylating tRNA genes. Of these hypermethylated tRNAs, 2 are in the previously identified genome-wide significant set of 3 (with tRNA-iMet-CAT-1-4 ranked 3rd by p-value) and 9 are in the study-wide significant set of 16. A full table with model results for all fitted models and all tRNA genes is provided in appendix 6.8.

4.5.2 tRNA Genes are Enriched for Age Related DNA Hypermethylation

Whilst ageing changes are pervasive throughout the DNA methylome, a strong enrichment for such changes occurs within the discrete tRNAome (Fisher's Exact Test $p = 1.05 \times 10^{-27}$) (see Figure 4.12 & appendix 6.10). This is still significant if the 6 of the study-wide significant 16 tRNAs that overlap polycomb or bivalent regions are excluded ($p = 4.66 \times 10^{-15}$)

CpG density is known to have a clear impact on the potential for variability of the DNA methylome as well as ageing-related changes [86,152]. To assess whether this hypermethylation finding was being driven merely by the inherent CpG density of the tRNAome, a CpG density matched permutation analysis was performed (1,000X, see Methods 4.4.3.3). This supported the specific nature of these age-related DNAm changes within the functional tRNAome (Empirical p-value $< 1.0 \times 10^{-3}$, Figure 4.12 b). As a point of comparison for this genomic functional unit, the same permutation analysis was performed for the known age-related changes in the promoters of genes that are polycomb group targets [155] and those with a bivalent chromatin state [153]. The enrichment of the polycomb group targets and bivalent regions (Empirical p-value $< 1.0 \times 10^{-3}$) was reproduced in this dataset.

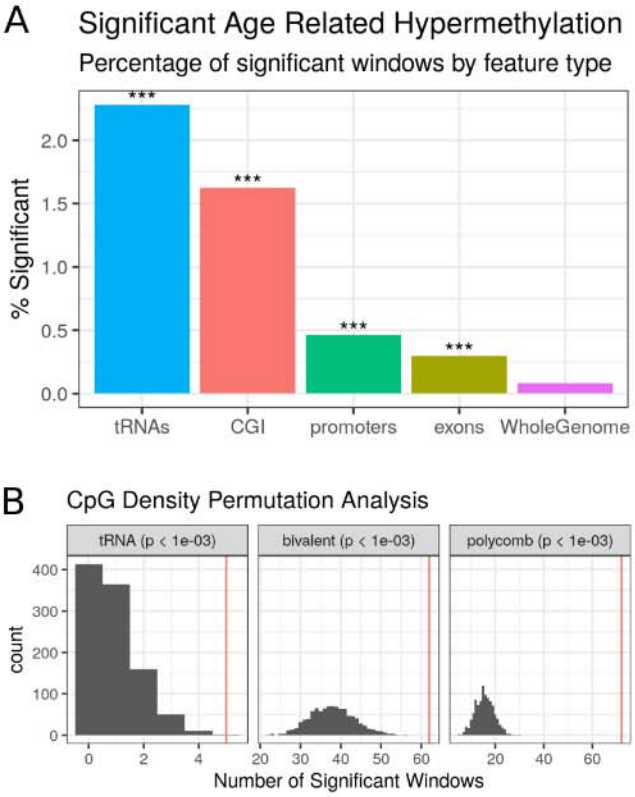

Figure 4.12: **A)** tRNA genes are enriched for age-related hypermethylation compared to the genomic background, (Fisher's Exact Test $p < 1.05 \times 10^{-27}$, $n = 3001$). **B)** tRNA genes show more significant hypermethylations than CpG Density matched permutations. Each permutation represented a random set of windows matching the CpG density of the functional unit (bivalent domains, polycomb group target promoters & the tRNA genes). These are subsequently assessed for signficant age-related DNAm changes (see Methods 4.4.3.3). The red line is the observed number of significant loci.

tRNA-iMet-CAT-1-4 is located in the largest tRNA gene cluster in the human genome at chr6p22.2-1 (Methods 4.4.2.1). This cluster contains 157 tRNA genes spanning the 2.67Mb from tRNA-iMet-CAT-1-2 to tRNA-Leu-AAG-3-1, and also hosts a histone gene microcluster. tRNA-Ile-AAT-4-1 and tRNA-Ser-AGA-2-6 are neighbours and are located on chromosome 17 (Figure 4.13). Notably tRNA-Ile-AAT-4-1 and tRNA-Ser-AGA-2-6 have a third close neighbour tRNA-Thr-AGT-1-2 which does not show significant age-related hypermethylation. A similar pattern of sharp peaks of significance closely localised around the other loci was observed in the study-wide significant set. GENCODE 19 places tRNA-Ile-AAT-4-1 in the 3'UTR of a Nonsense-mediated decay transcript of *CTC1* (CST Telomere Replication Complex Component 1, CTC1-002, ENST00000449476.2) and not of its primary transcript.

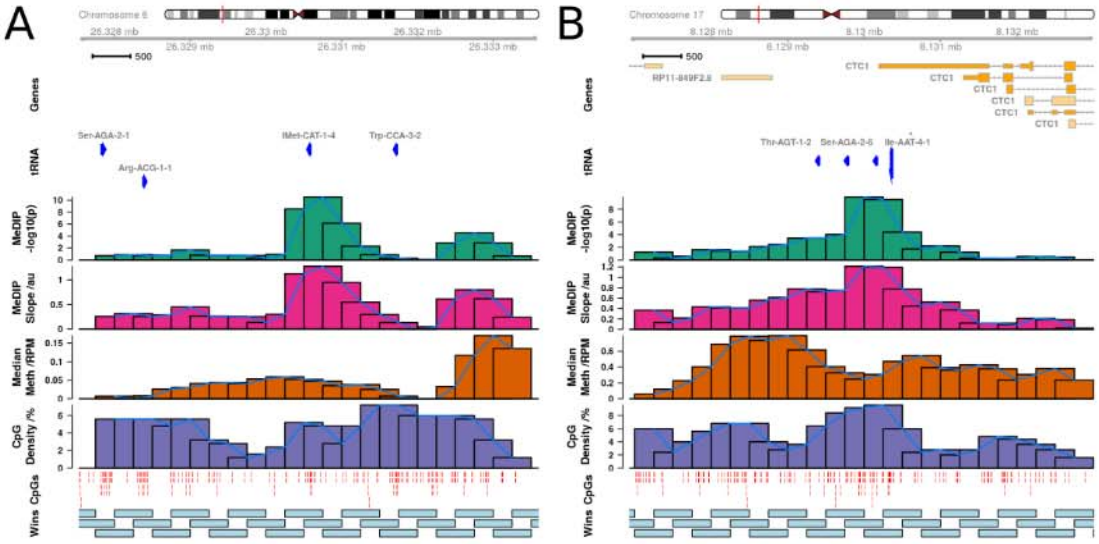

Figure 4.13: **A)** tRNA-iMet-CAT-1-4 as well as **B)** tRNA-Ser-AGA-2-6 and tRNA-Ile-AAT-4-1 exhibit age related DNA hypermethylation in MeDIP-seq data. The sharp peaks suggest that this effect is localised to individual tRNA genes. One of the windows overlapping tRNA-Ile-AAT-4-1 also partially overlaps tRNA-Ser-AGA-2-6. Results shown are from the blood cell-type and batch corrected model. Median Methylation in reads per million is calculated across all samples in the model. The window structure track (Wins) illustrates the tiled nature of the MeDIP-seq windowing. Some annotations of the 3' UTR of CTC1 extend further than illustrated here covering tRNA-Thr-AGT-1-2. CpG density and CpG position are included to illustrate that similarly CpG dense tRNA genes are exhibiting differing age related DNAm change patterns.

4.5.2.1 tRNA gene clustering

To place these hypermethylating tRNA genes in their genomic context the clustering for the extended set of 44 non-blood cell-type corrected study-wide significant tRNAs was examined. The tRNA genes were clustered by grouping together all tRNAs within 5Mb of one another and then required that a cluster contain at least 5 tRNA genes with a density of at least 5 tRNA genes per Mb (see Methods 4.4.2.1). This yielded 12 major tRNA gene clusters containing a total of 353 tRNA genes, 42 of the 44 study-wide significant tRNA genes in the non cell-type corrected age model reside within these clusters (figure 4.14 a). The hypermethylating tRNA genes are spread evenly among these clusters proportionately to their size, (no significant difference in a one-way ANOVA of percentage of significant tRNAs by cluster).

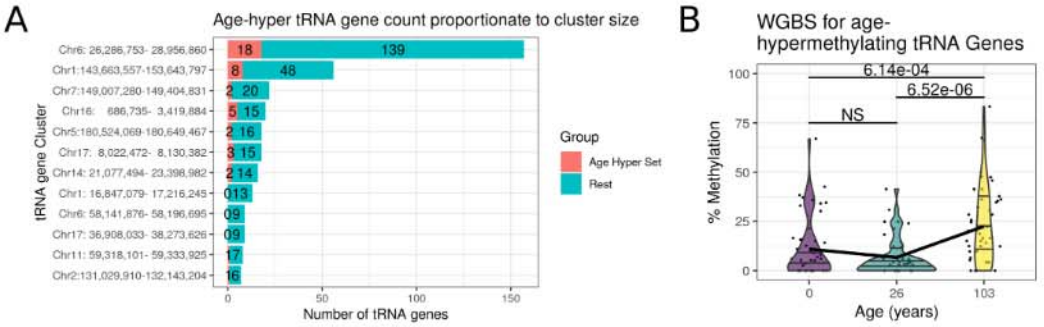

Figure 4.14: **A)** Almost all tRNA genes (42/44, counts shown in red) study-wide significant in the MeDIP batch corrected age model, and all of the genome-wide significant and blood-cell type corrected sets of tRNA genes reside in one of 12 major tRNA gene clusters. (Defined by joining all tRNAs within 5Mb of one another and requiring at least 5 tRNA genes per cluster with a density of at least 5 tRNA genes per Mb). **B)** Available study-wide significant (SWS) tRNA genes (n = 14) are more methylated in a centenarian than in a neonate or a 26 year old. Whole Genome Bisulfite Sequencing Data in a newborn, as adult and a centenarian. Each point represents the methylation level at an individual CpG within a tRNA gene.

4.5.2.2 Age-related tRNAome DNA Hypermethylation is even observed in one Newborn versus one Centenarian

Whole genome bisulfite data was sought out for two reasons: 1) The greater positional resolution of WGBS permits one to check if the age-related hypermethylation could be seen at CpGs located within the tRNA genes themselves. 2) As DNA methylation levels are estimated based on the ratio of methylated to unmethylated reads using WGBS, seeing an increased DNA methylation with age is not likely to be due to a copy number expansion or differential read distribution with age which could confound the MeDIP-seq signal. An available Whole Genome Bisulphite sequencing (WGBS) dataset from Heyn *et al.* [161] was examined (see Methods 4.4). These data consisted of blood-derived DNA WGBS in one newborn child and one 26 year old, and centenarian (103 years). In their analysis, the centenarian was found to have more hypomethylated CpGs than the neonate across all genomic compartments, including promoters, exonic, intronic, and intergenic regions. However, even in this examination of 3 individuals of 3 different ages in the 55% of tRNA genes that possessed coverage, DNA hypermethylation with age was observed among the study-wide significant hypermethylating tRNA genes. The centenarian was significantly more methylated in this set of tRNAs than the neonate (Wilcoxon rank sum test, 6.14% increase (95% CI -Inf - 4.31), W = 717, p = 6.14×10^{-4}, see Figure 4.14 b).

4.5.2.3 Age-related Changes Independently Replicated with Targeted Bisulfite Sequencing

In order to further robustly support these-ageing related changes, an attempt was made to replicate these findings in an independent ageing dataset. Furthermore, a different technology was employed, Targeted bisulfite (BiS) sequencing, to further validate the MeDIP-seq-derived results. These data provide individual CpG resolution to identify what may be driving the regional DNAm changes observed.

This targeted BiS-seq was performed in blood-derived DNA from 8 pools of age-matched individuals at 4 time-points (~4, ~28, ~63, ~78 years) from a total of 190 individuals, as detailed in Table 4.1. A total of 79 tRNA loci generated reliable results post-QC (see Methods 4.4.3.6). These tRNAs covered a total of 458 CpGs with a median of 6 CpGs per tRNA (range 1-9). Median Coverage per site across pools, technical replicates and batches was 679 reads (mean 5902).

Firstly, the 8 Pooled samples were run on the Illumina EPIC (850k) array to confirm that this pooling approach was applicable for DNAm ageing-related evaluation. This showed an $R^2 = 0.98$ between pool mean chronological age and Horvath clock DNAm predicted age [168](see Figure 4.15). This approach permitted the assay tRNA gene DNA methylation across a large number of individuals whilst requiring a minimal amount of DNA from each individual (80-100ng), and costing ~1/24th as much as performing sequencing individually. Therefore, this confirmed the utility of this novel pooling approach. These DNA methylation array derived data were also used to estimate the major blood cell proportions for each of these pools with the Houseman algorithm [272].

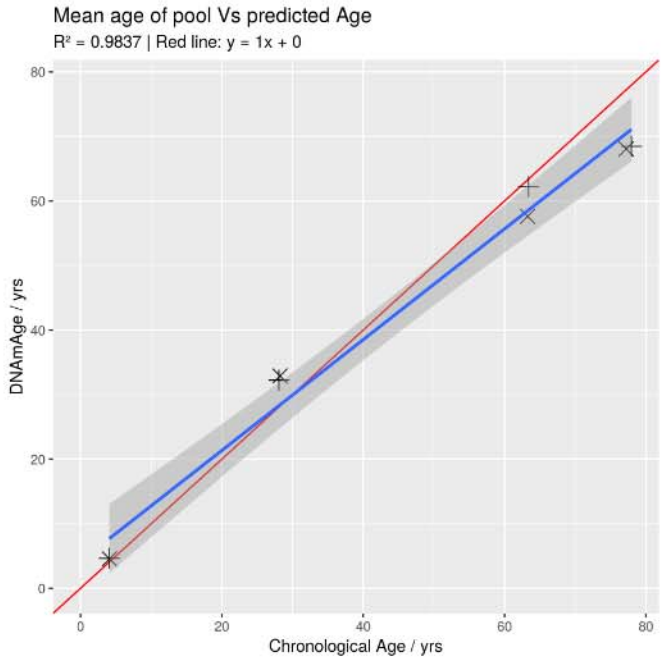

Figure 4.15: Mean chronological age is tightly correlated with DNAm Horvath clock [168] predicted age for the 8 pooled samples. (See Table 4.1 for pool details).

It was observed that individual tRNA loci exhibiting age-related changes in DNAm had duplicate or isodecoder (same anticodon but body sequence variation) sequences in the genome, which despite exact or near sequence identity did not show similar changes. tRNA-iMet-CAT-1-4 for instance is 1 of 8 identical copies in the genome and was the only locus that showed significant changes. The results of pairwise differential methylation tests between age groups for the 6 top tRNAs from the MeDIP-seq models are listed in Table 4.3.

Of the 3 top hits in MeDIP-seq, tRNA-iMet-CAT-1-4 (Figure 4.17c) and tRNA-Ser-AGA-2-6 (Figure 4.17i) exceeded nominal significance (p-values $= 9.35 \times 10^{-4}$ & 4.28×10^{-2} , respectively). tRNA-Leu-TAG-2-1 from the study-wide significant set also showed nominally significant hypermethylation with age (Figure 4.17u). Also, four of the individual CpGs in tRNA-iMet-CAT-1-4 exhibited nominally significant increases in DNAm with Age (Figure 4.16). However, tRNA-Ile-AAT-4-1 (Figure 4.17n) showed a nominal decrease in DNAm with age.

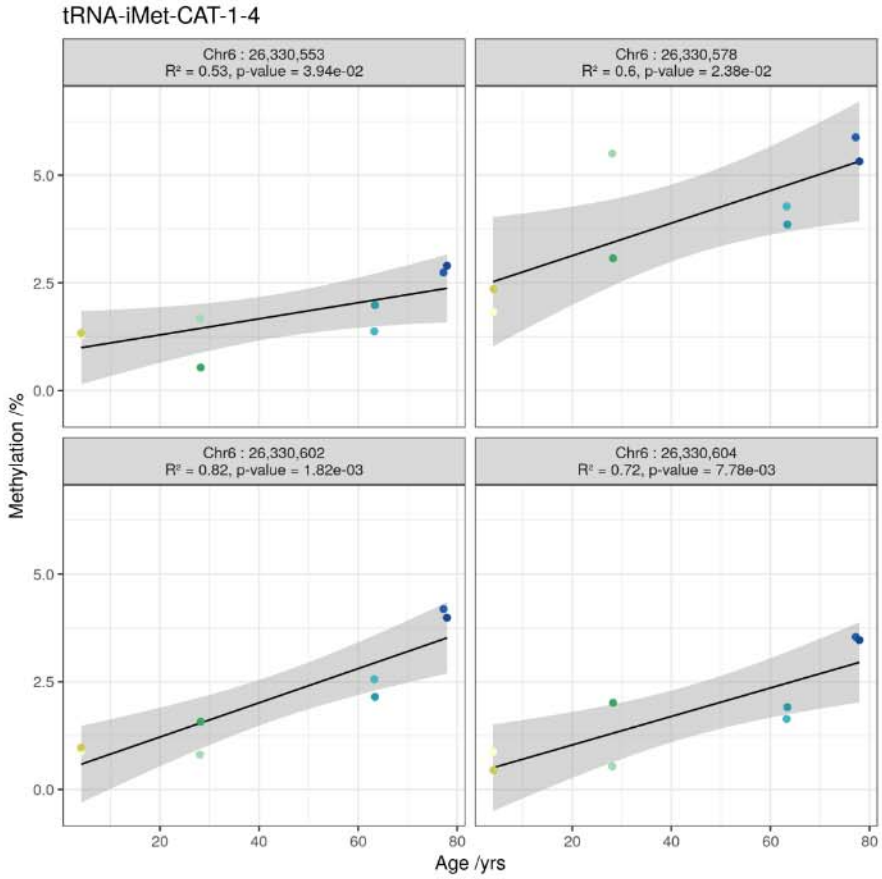

Figure 4.16: Individual CpG methylation increases (nominally significant $p < 0.05$) in tRNA-iMet-CAT-1-4.

Table 4.3: Pairwise Differences in Methylation between Age groups by tRNA. p-values are for pairwise methylation differences (see Methods 4.4.3.6)[366].

tRNA	num. CpGs	comparison	p-value	delta
tRNA-Ile-AAT-4-1	8	4 vs. 28	1.518e-01	-0.2
		4 vs. 63	1.774e-01	-0.234
		4 vs. 78	3.060e-01	0.0113
		28 vs. 63	7.152e-01	-0.0334
		28 vs. 78	1.553e-01	0.212
		63 vs. 78	2.057e-01	0.245
tRNA-iMet-CAT-1-4	5	4 vs. 28	8.403e-02	0.0116
		4 vs. 63	1.716e-01	0.0125
		4 vs. 78	1.997e-04*	0.0368
		28 vs. 63	3.943e-01	0.000869
		28 vs. 78	1.724e-02*	0.0252
		63 vs. 78	6.224e-02	0.0243
tRNA-Ser-AGA-2-6	9	4 vs. 28	4.222e-01	0.0573
		4 vs. 63	3.968e-01	0.0274
		4 vs. 78	4.651e-01	0.0423
		28 vs. 63	1.095e-01	-0.0299
		28 vs. 78	2.126e-01	-0.015
		63 vs. 78	2.201e-01	0.0149

4.5.2.3.1 Select Duplicates & Isodecoders of Hypermethylating tRNA loci remain unchanged A selection of these duplicate and isodecoder loci were targeted for bisulfite sequencing in order to confirm that the identified DNAm changes are specific to a given locus and not general to related tRNAs. Examining the tRNA-iMet-CAT-1 family, only the previously identified 1-4 version confirmed significant hypermethylation (not 1-2, 1-3 or 1-5)(Figure 4.17a-e). Likewise the tRNA-Ser-AGA-2-6 version was supported compared to 2-1,2-4 and 2-5(Figure 4.17f-j)).

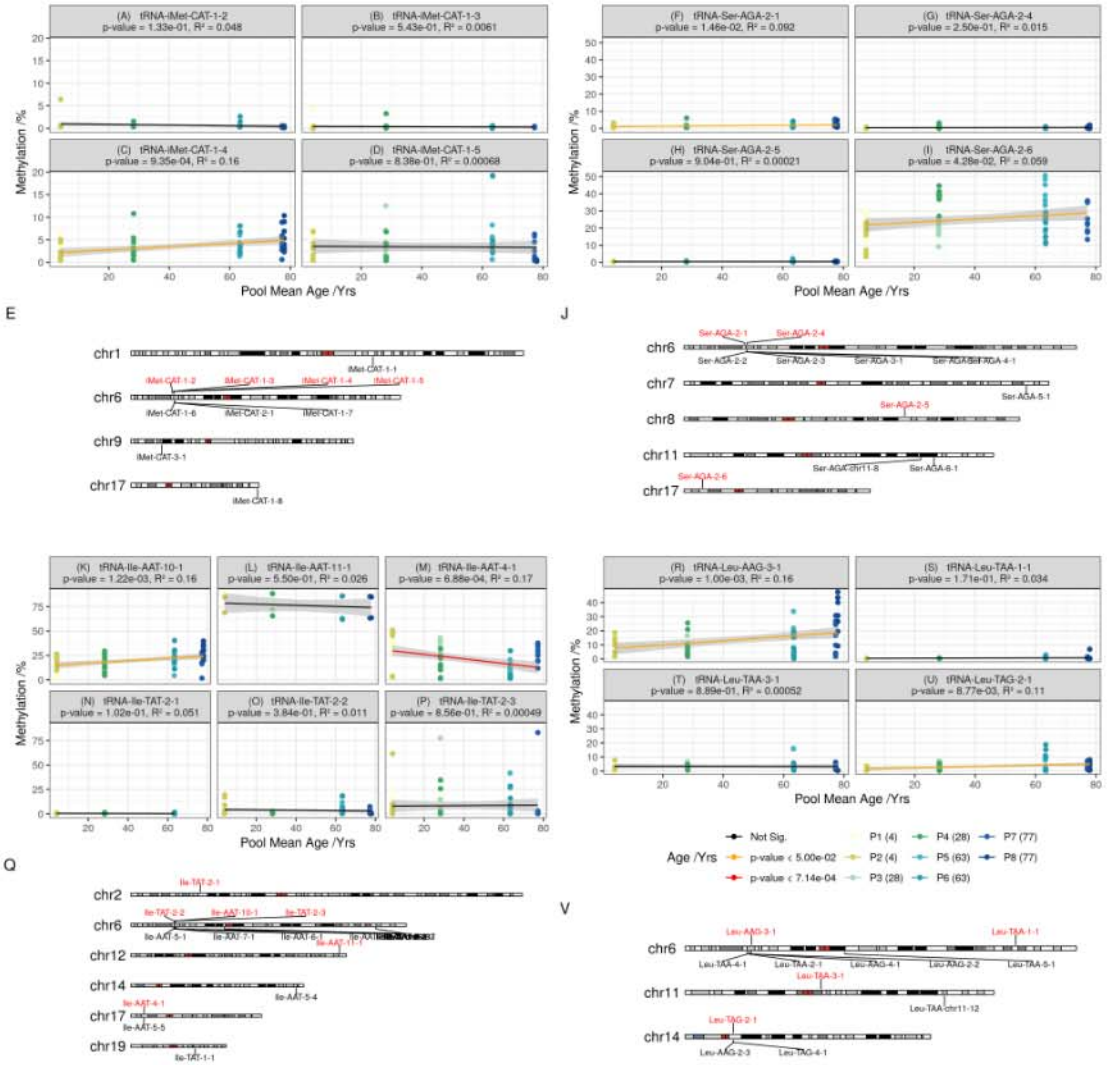

Figure 4.17: Combined CpGs within tRNA loci results (experiment-wide Bonferroni p = 7.14×10^{-4}); **(A-D)** Comparison of select tRNA-iMet-CAT loci: Hypermethylation is specific to iMet-CAT-1-4 (c) not other isodecoders (A, B, & D); **(F-I)** Comparison of select tRNA-Ser-AGA loci: Hypermethylation is specific to Ser-AGA-2-6 A (viii) and to a lesser extent Ser-AGA-2-1 (F), whilst not other isodecoders (G, H); **(K-P)** Comparison of select tRNA-Ile loci: Hypermethylation is specific to Ile-AAT-10-1 (K), Ile-AAT-4-1 (M) displays hypomethylation contrary to previous MeDIP findings, Ile-TAT-2-2 & 2-3 lack hypermethylation (previously non-significant in blood-corrected MeDIP, although significant in uncorrected), whilst no change in Ile-AAT-11-1 (L) and Ile-TAT-2-1 (N); **(R-U)** Comparison of select tRNA-Leu loci: Hypermethylation in Leu-AAG-3-1 (R) consistent with 450k and Leu-TAG-2-1 (U) consistent with MeDIP, whilst no change in Leu-TAA-1-1 (S) & Leu-TAA-3-1 (T).

4.5.2.4 DNA methylation 450k Array Data Validates the MeDIP-seq Results

Although DNA methylation arrays poorly cover the tRNAome, an attempt was made to ascertain if this bisulfite conversion-based but differing and well-established technology was supportive at all of the previous DNA hypermethylation findings. TwinsUK had available 450k array on 587 individuals, and this platform includes 143 probes, covering 103 tRNAs. All the 3 top tRNAs in the MeDIP-seq results were covered by this data set, and 7 of the 16 study-wide significant set. 9 tRNAs show significant ($p < 4.58 \times 10^{-4}$) increases in DNA methylation with age in models corrected for blood cell counts including all 3 of the 3 tRNAs identified in the MeDIP-seq as genome-wide significant and 5 of the 7 study-wide significant set present on the array (Figure 4.18). Although it should be noted that 56 of these 143 probes are within the non-robust set of Zhou et al. [223], including 1 of the genome-wide, and 1 of the study-wide results (covering tRNA-Ile-AAT-4-1 & tRNA-Val-AAC-4-1), respectively. Full model results for the Twins UK 450k array data are provided in appendix 6.12.

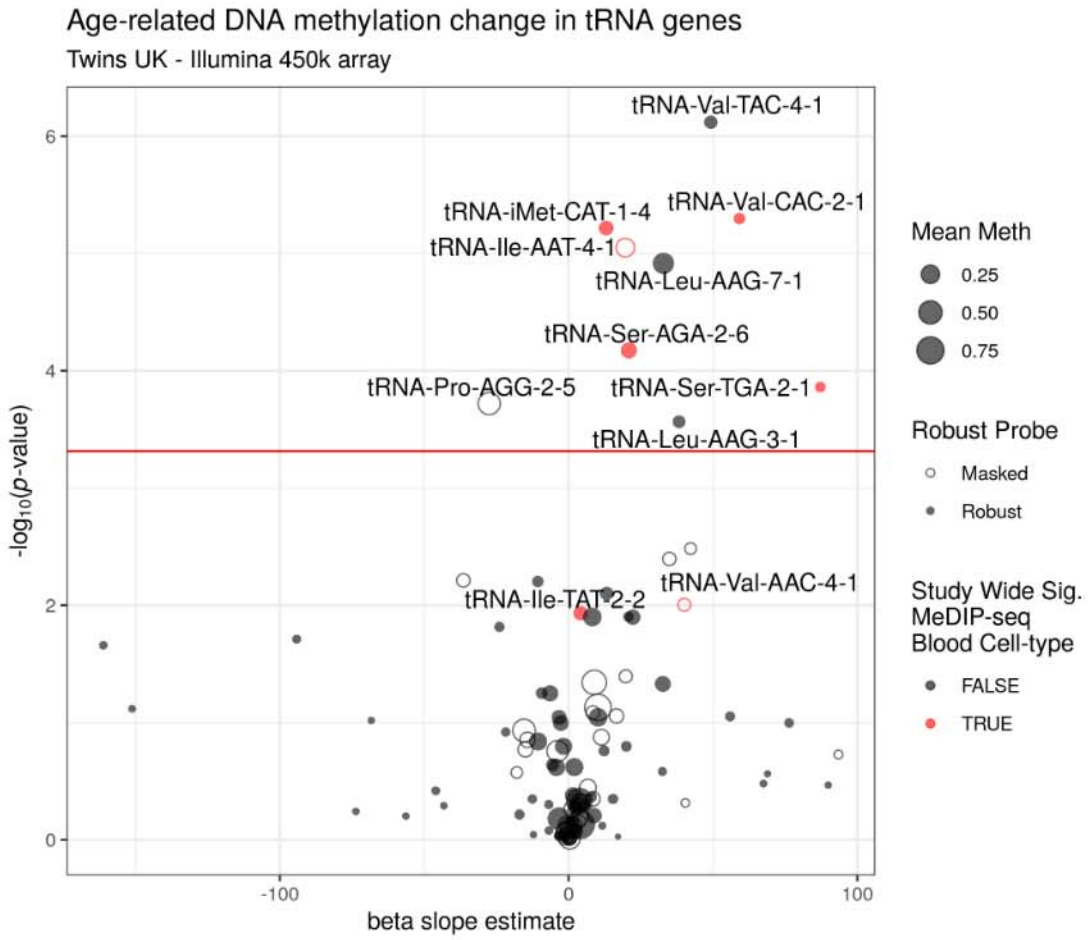

Figure 4.18: Volcano-like plot. tRNAs are labelled if they are significant here or were in the MeDIP-seq data (Red). Model slope: the model coefficient for the methylation values. Unfilled circles indicate those probes in the general mask generated by Zhou et al. [223]. Significance threshold: $0.05/103 \approx 4.58 \times 10^{-4}$ (the number of tRNA genes examined).

4.5.2.5 Ageing-Related tRNA Loci show increased Enhancer-Related Chromatin Signatures

The activity of the tRNAome was further explored using public chromatin segmentation data in blood (Epilogos Blood & T-cells set) [368]. This shows proportionally more Enhancer-related (Enh, EnhBiv & EnhG) chromatin states at tRNA genes hypermethylating with age than the stronger Promoter-related (TSS) in other tRNAs. (Figure 4.19). Whereas these characteristics are less frequently predominant in the rest of the tRNAs (Figure 4.19). Age-hypermethylating tRNA are enriched for enhancer chromatin states compared to the rest of the tRNAome (Fisher's Exact test p = 0.01).

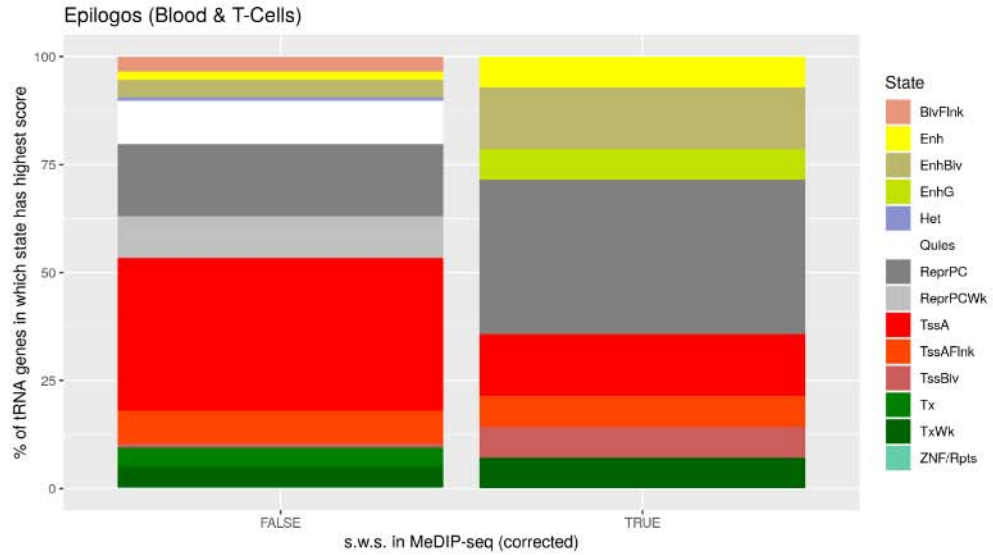

Figure 4.19: Chromatin segmentation data from the Epilogos [368] 'Blood & T-cell' 15 State model (tRNA genes +/- 200bp). Frequency with which a model state was the predominant state at a given tRNA. Proportions of predominant tRNA state for the 14 study-wide significant age-hypermethylating tRNAs covered compared to other 371 available tRNAs.

4.5.2.6 Age Hypermethylating tRNAs are more methylated in Lymphoid than Myeloid cells

Three tRNA genes remained genome-wide significant and 16 study-wide significant following correction for major cell-type fraction. This is suggestive of either cell-type independent change or, presumably less likely, a very large effect in a minor cell-type fraction. tRNAs have exhibited tissue-specific expression [378,325,326] and blood cell-type populations change with age. Specifically, there is shift to favour the production of cells in the myeloid lineage [377]. These points lead us to examine tRNA gene DNAm in sorted cell populations. A publicly available 450k array dataset [369]) that has been used in the construction of cell-type specific DNAm references for cell-type fraction prediction using the Houseman algorithm [272] was used (see Methods 4.4.5). This consists of data from 6 individuals (aged 38 ± 13.6/yrs) from seven isolated cell populations (CD4+ T cells, CD8+ T cells, CD56+ NK cells, CD19+ B cells, CD14+ monocytes, neutrophils, and eosinophils). It was found that tRNA gene DNAm could separate myeloid from lymphoid lineages (Figures 4.20 & 4.21).

Of the eight study-wide significant tRNAs with array coverage, it was identified that collectively these eight are significantly more methylated in the lymphoid than the myeloid lineage (1.1% difference, Wilcoxon rank sum test $p = 1.50 \times 10^{-6}$ 95% CI 0.7%-∞). Thus, any age related increases in myeloid cell proportion would be expected to dampen rather than exaggerate the observed age-related hypermethylation signal. In addition tRNA-Ile-AAT-4-1 and tRNA-Ser-

AGA-2-6 have the highest variance in their DNAm of all 129 tRNAs covered in this dataset. This could represent ageing-related changes as these samples range across almost 3 decades. Another possibility may be that these loci as well as hypermethylating are also increasing their variability with age in a similar fashion to those identified by Slieker *et al.* [120]. In that study they identified that those loci accruing methylomic variability were associated with fundamental ageing mechanisms.

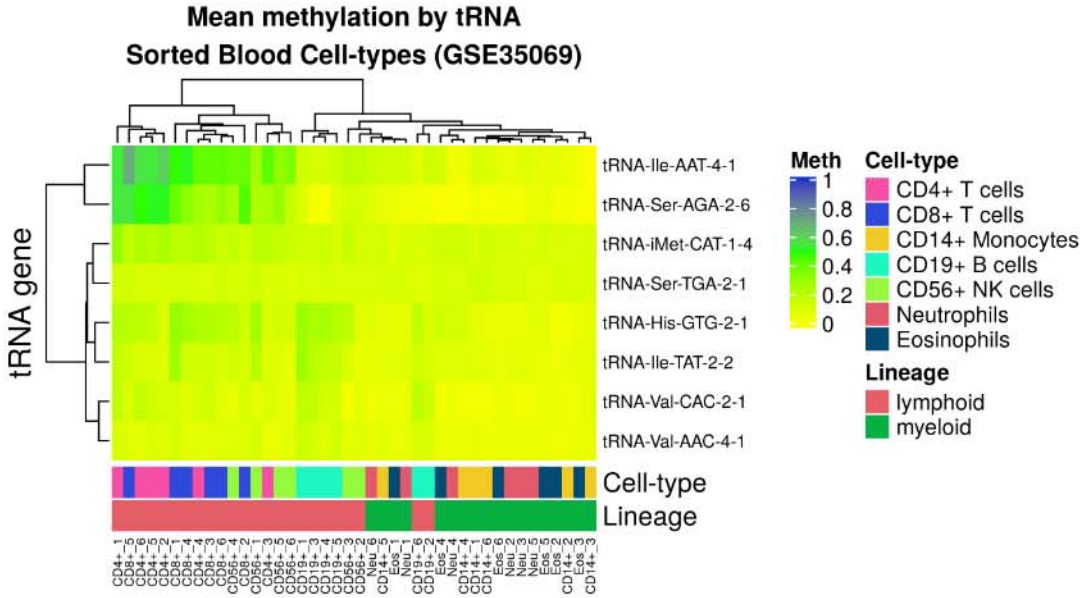

Figure 4.20: Heatmap [379] of mean methylation of probes covering each tRNA in 7 cell-type fractions from 6 Male individuals. Data from GSE35069 [369]. Of the 16 study-wide significant hypermethylating tRNAs, 8 are covered by this dataset.

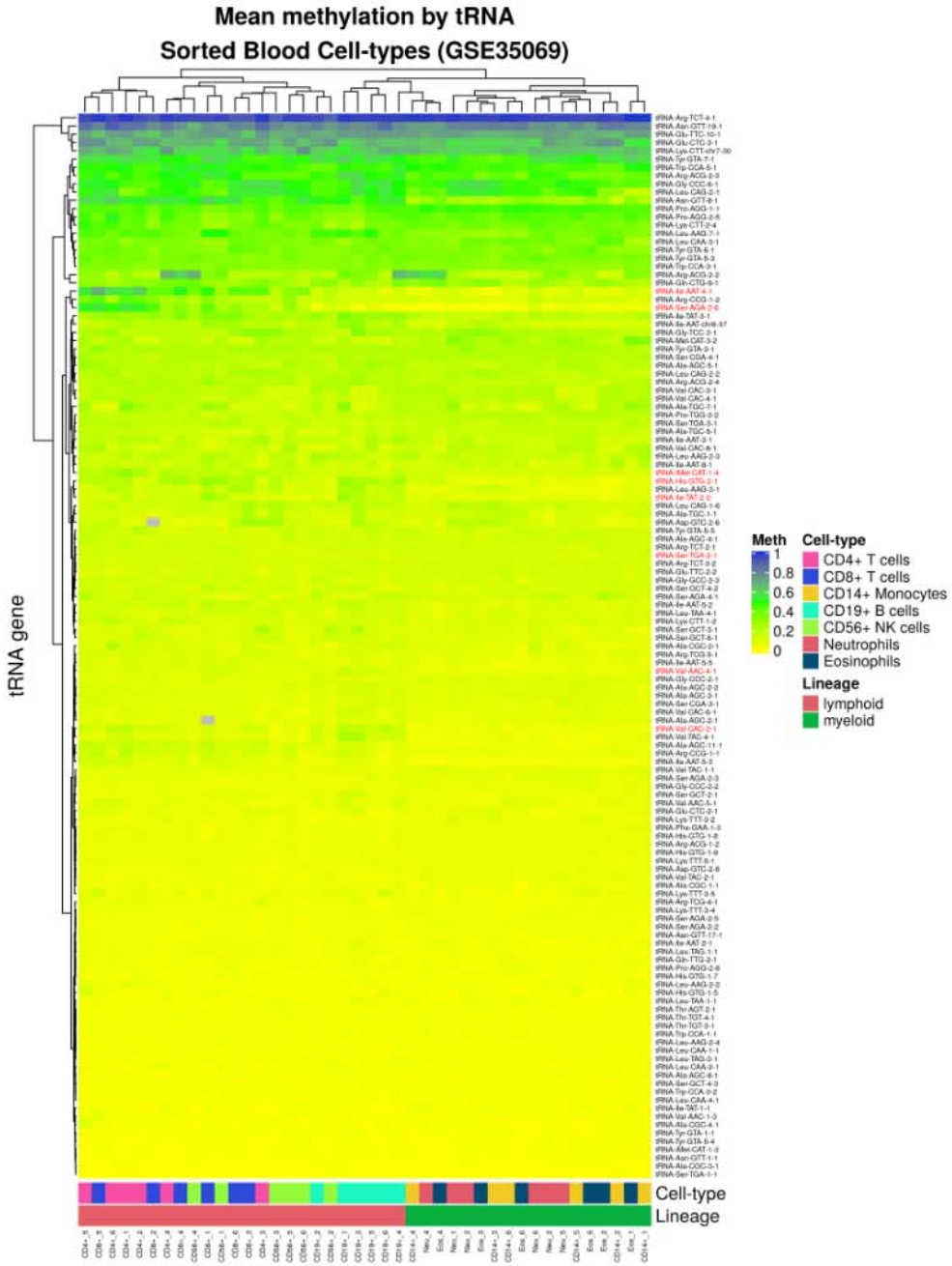

Figure 4.21: Heatmap Mean Methylation of probes covering each tRNA in 7 cell-type fractions from 6 Male individuals. Showing all 150 tRNAs covered by 213 probes on the Illumina 450k array. Data from GSE35069 [369] downloaded using `GEOquery` [370]. Generated with the `ComplexHeatmap` `R` package [379].

4.5.3 tRNA Gene DNA Methylation in Other Tissues

Some tRNA gene expression has been shown to be highly tissue specific [378,325,326]. It follows that our observations of changes in DNAm with age in blood might be specific to that tissue.

A mix of 450k and 27k array data from 'solid tissue normal' samples made available by TCGA (The Cancer Genome Atlas) and data from foetal tissue [372,373] were downloaded from GEO for use in this analysis (see Methods 4.4.6). The samples from TCGA range in age from 15-90 (n = 733). Only 43 tRNA genes had adequate data to compare across tissues in this dataset and 115 in the foetal tissue data.

4.5.3.1 tRNA Genes also Hypermethylate with Age in Solid Tissue

Only 2 of the 3 tRNA genes identified as genome-wide significant and a further 1 of the study-wide significant tRNA genes are present in the set of 45 tRNA genes in the TCGA data, thus limiting our ability to draw conclusions about the tissue specificity of these results. Solid tissue samples have a strong preponderance for low levels of methylation consistent with the active transcription of many tRNA genes and show slight increasing methylation with age but age accounts for very little of the variance (linear regression slope estimate = 1.52; $R^2 = 0.0002$; p-value 1.34×10^{-3} (Figure 4.24d). In a pan-tissue analysis it was found that 10 tRNA genes showed changes in DNAm with age, 9 of which were hypermethylation (p-value $< 1.1 \times 10^{-3}$). One of these tRNA genes, tRNA-Ser-TGA-2-1 was also present in study-wide significant set of tRNA genes. Figures 4.22 & 4.23 illustrate minimal tissue specific differences. Interestingly, however, tRNA-iMet-CAT-1-4 and tRNA-Ser-AGA-2-6 appeared more variable in methylation state than many other tRNAs in the TCGA normal tissue samples (Figure 4.22) and indeed have the highest variance in DNA methylation across tissues (Figure 4.24c).

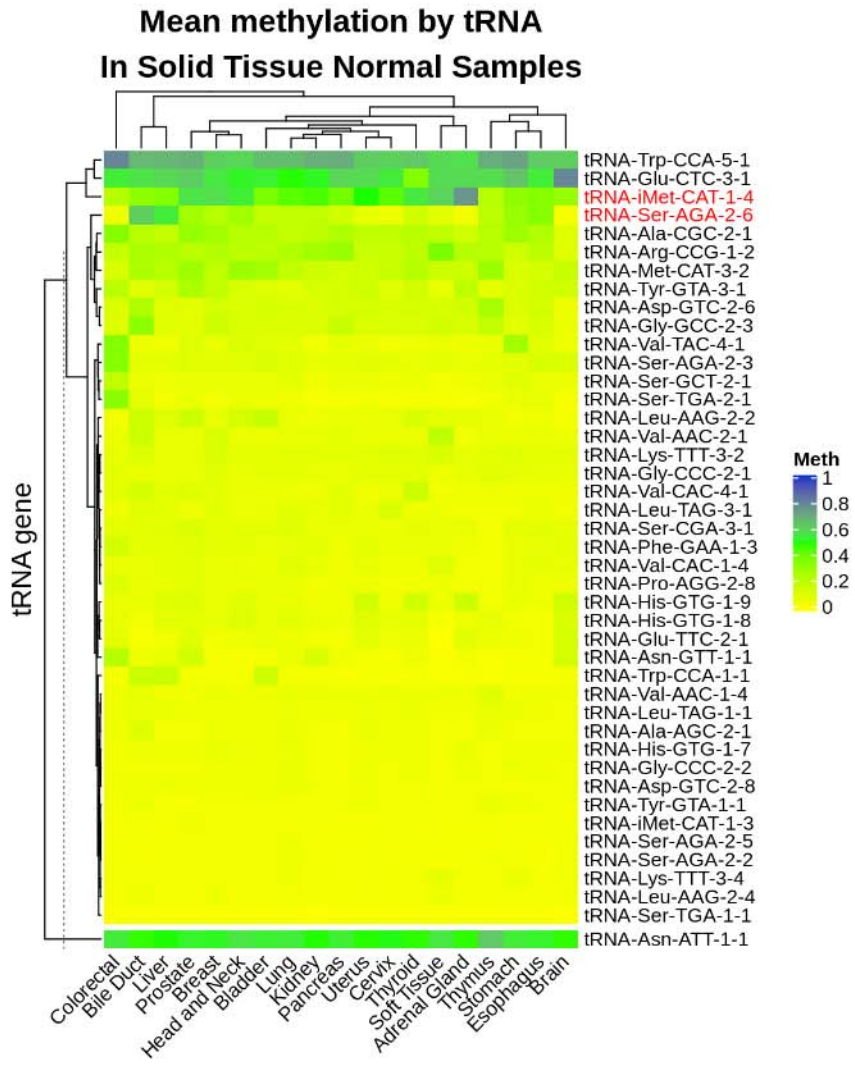

Figure 4.22: Mean Methylation of 43 tRNAs in 19 tissues. Possible pseudogene (tRNA-Asn-ATT-1-1) is shown in a separate cluster beneath the main heatmap [379].

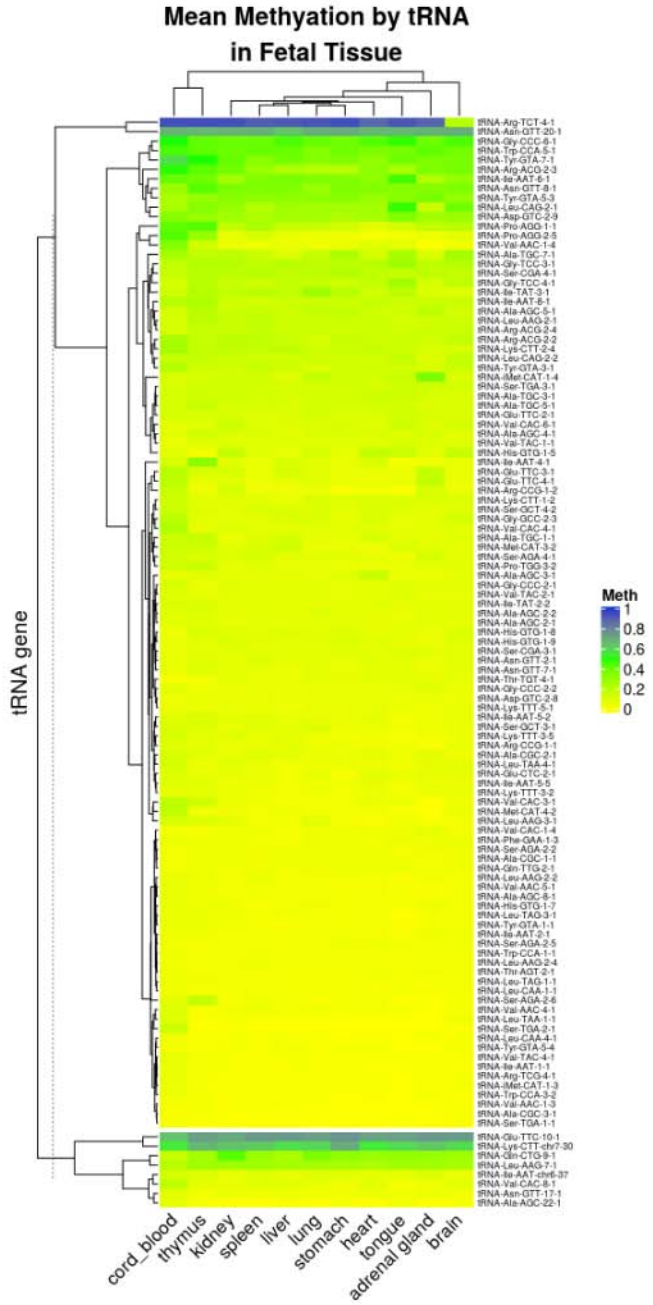

Figure 4.23: Mean Methylation of 115 tRNAs in 11 tissues. Possible pseudogenes are shown in a separate cluster beneath the main heatmap [379].

4.5.3.2 tRNA Gene Methylation in Cancer

A coarse grained examination of the tRNA DNA methylome in tumour samples with matched normal tissue data was performed, considering the 45 tRNA genes and 19 tissues for which data were available. Across multiple tissues it was found that both tumour and 'normal' samples have a strong preponderance for low levels of methylation consistent with the active transcription of

many tRNA genes.

Tumour samples have a higher mean methylation, 0.27% ($95\% ci$ $2.28 \times 10^{-03} - \infty$) greater than that of normal samples (Wilcoxon signed-rank test p-value $= 9.1 \times 10^{-20}$), and a greater variance (Levene test p-value $< 1 \times 10^{-3}$) (Figure 4.24b). Interestingly visual inspection of methylation values in normal tissues (Figure 4.24b & d) suggests a slight bimodality peaking around 50% methylation that is not present in the tumour samples. These differences are small in absolute terms and have very wide error margins but are consistent with a picture of dysregulation in cancer cells [319].

Both Primary Tumour and Solid Tissue Normal samples show slight increasing methylation with age but age accounts for very little of the variance (linear regression slope estimates = 1.65, 1.52; $R^2 = 0.0006$, 0.0002; p-values 1.20×10^{-6}, 1.34×10^{-3} respectively)(Figure 4.24d).

Consistent with Figure 4.22 the tRNAs tRNA-iMet-CAT-1-4 and tRNA-Ser-AGA-2-6 which have the highest ranked variance in normal tissue and also rank highly in tumour samples (Figure 4.24c).

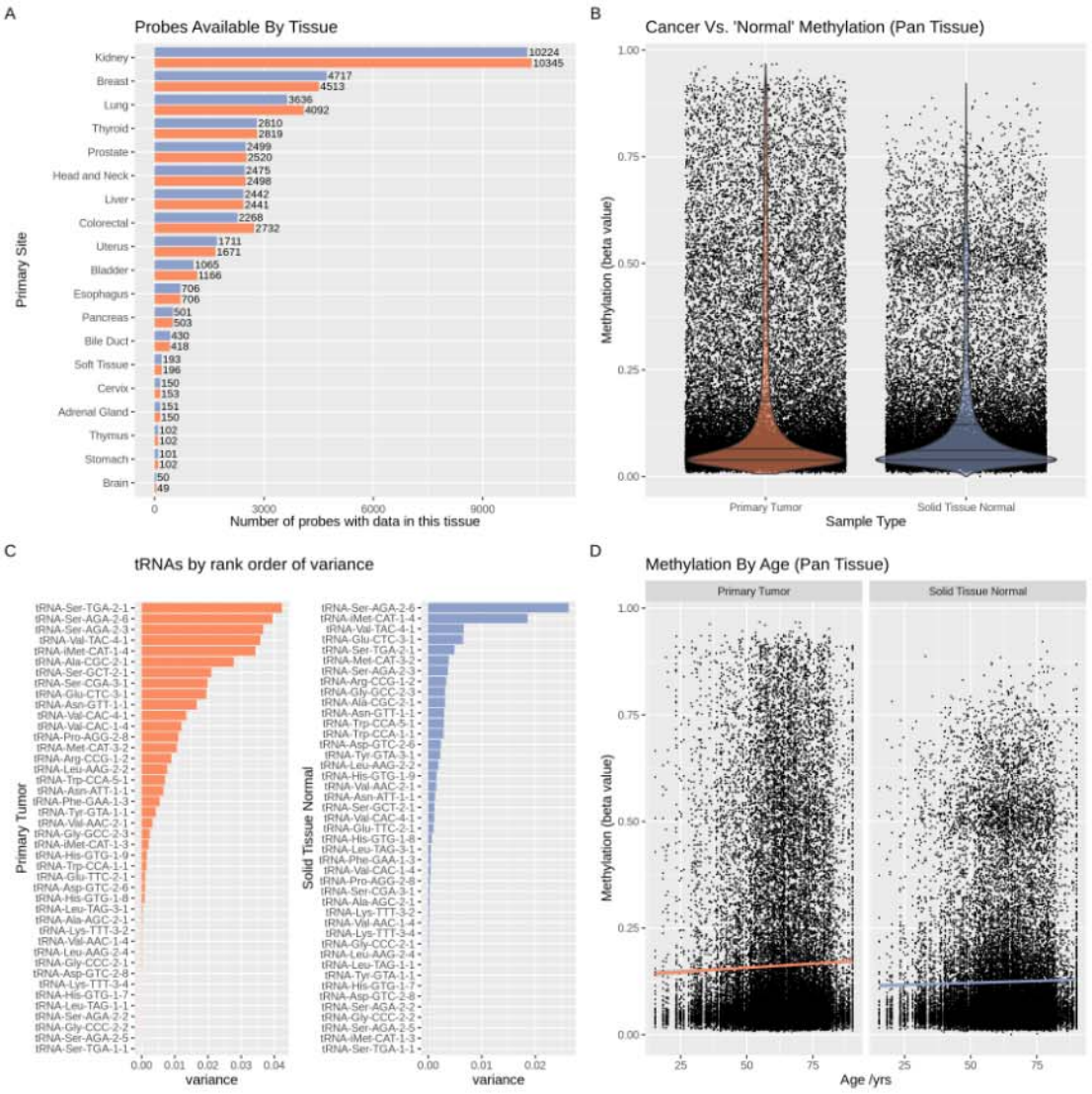

Figure 4.24: Global properties of tRNA methylation data for 45 tRNA genes across 19 tissues with matched normal and tumour samples from 733 cases in TCGA [372,373].

4.5.4 Expression of tRNAs in Blood with Age

Having observed specific tRNA gene isodecoders hypermethylating with age it followed that the expression of tsRNA in blood cell-types would be worth exploring to ascertain if the expression levels of such transcripts declined as DNA methylation levels of the loci which encode them increase. A bioinformatic approach was devised to attempt to assay tRNA transcription in order to use standard publicly available small RNA-seq datasets. A customised MINTmap [337] reference was designed to include only fragments which unambiguously map to a single tRNA gene locus and which overlap the 5' or 3' end of the genomic tRNA sequence by at least one base with no mismatches. This reference is intended to capture pre-tRNAs prior to processing and CCA addition operating under the assumption that the levels of pre-tRNAs will be informative

about the amount of transcription taking place at the tRNA loci (see Methods 4.4.7). This custom MINTmap reference build yielded 383 fragments mapping to 92 distinct tRNA loci in this data. The lack of coverage of age hypermethylating tRNAs by uniquely attributable RNA-seq reads prevented us from drawing any strong conclusions about the relationship between DNAm changes and changes in tRNA transcription.

Using the original MINTmap reference optimised to detect tRNA fragments derived from mature tRNAs there were 5384 unique fragments derived from as many as 417 tRNA loci. However, the mapping between fragments and loci in this reference is many to many, with each tRNA gene able to give rise to many fragments and each fragment attributable to at least 1 and usually many tRNA genes. The examination of these fragments was limited to those with a length of greater than or equal to 40nt to capture reads more likely to be derived from mature tRNAs rather than tRFs or tRNA halves (Figure 4.26). 48 tsRNAs with nominally significant expression changes ($p < 0.05$) were identified, 8 increased and 40 decreased in abundance with age. For example 5 of 6 fragments showing significant age-related expression changes derived from the Gln-CTG family of tRNAs are decreasing with age (Figure 4.25). This is suggestive that expression of some tRNA genes may decline with age but this possibility is in need of additional tRNA expression data before it can be asserted with confidence.

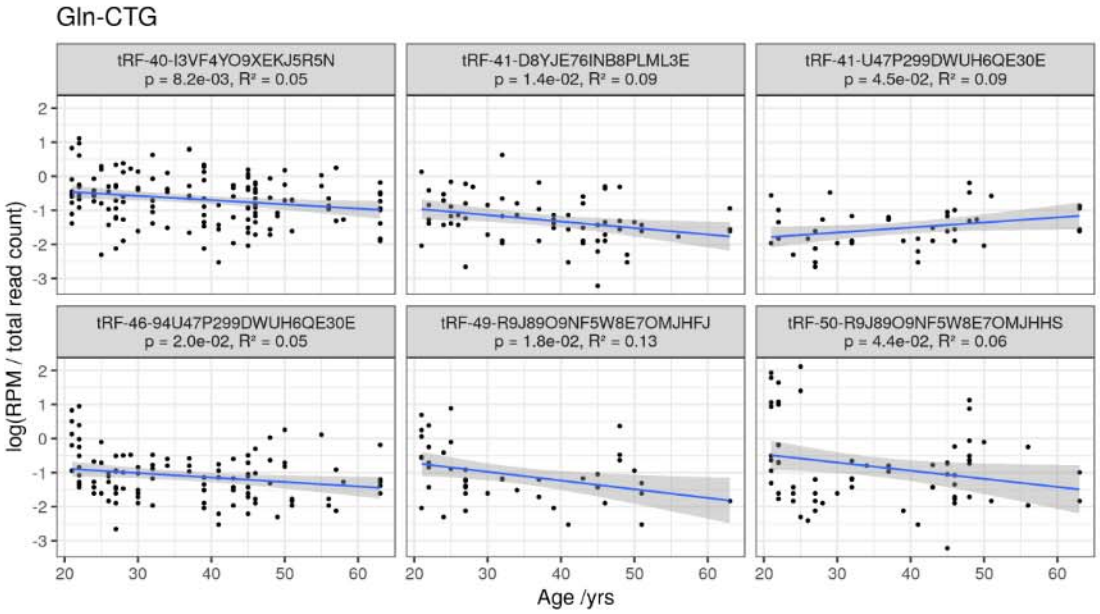

Figure 4.25: tRNA fragments derived from the Gln-CTG family of tRNAs, selected as tRNA-Gln-CTG-7-1 is one of the 16 study-wide significant age-hypermethylating tRNA genes. Pane titles contain the MINTbase Plates, unique identifiers of the tRNA fragments [329].

4.5.4.1 MINTmap reference Fragment distribution

In the original MINTmap reference (Figure 4.26b) there are peaks at around 18nt, 22nt and 32nt. This is consistent with the expected tRNA fragment size distributions with 'tRNA halves' at 30-33nt and other tRFs at 18nt and 22nt. In this custom reference (Figure 4.26a) whilst there is still a peak at ~18nt, with suggestions of peaks near 22nt and 32nt the tRNA fragment length distribution is somewhat different from that of the standard MINTmap reference. There are larger peaks at ~28 and ~40nt consistent with the longer fragments expected given that this reference aimed to target fragments derived from pre-tRNAs not tRFs derived from mature tRNAs.

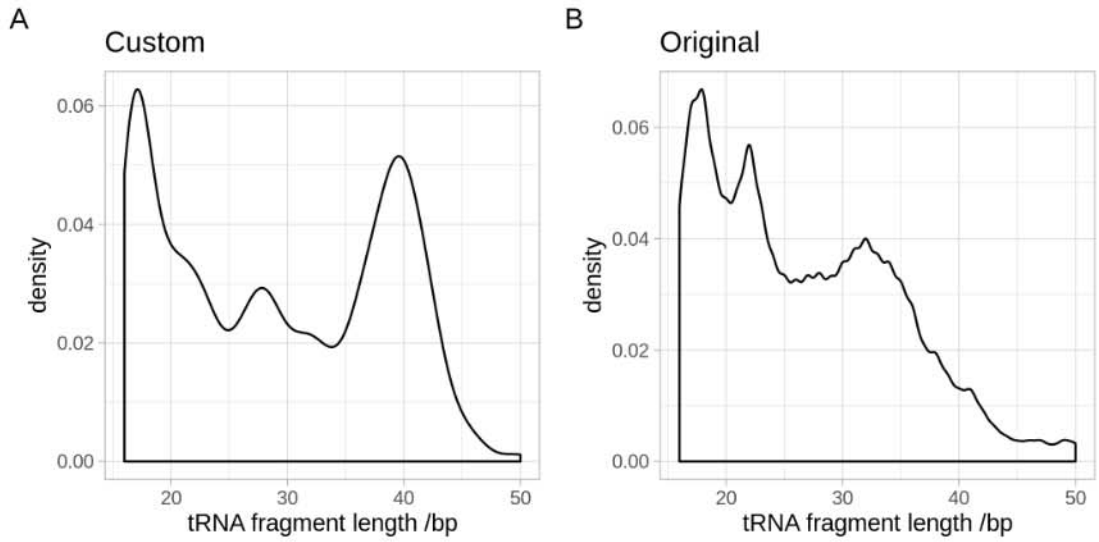

Figure 4.26: Comparison of the fragment size distributions between our custom reference **A)** and the original the MINTmap reference **B)**.

4.5.5 Mice also show age-related tRNA gene DNA hypermethylation

To see if this age-related tRNA gene DNA hypermethylation was present in another mammal, the DNA methylation of the mouse tRNA genes was examined using data from a reduced representation bisulfite sequencing (RRBS) experiment performed by Petkovich et al. [375]. These data from 152 mice covered 51 tRNA genes and 385 CpGs after QC (see Methods 4.4.8), representing ~11% of mouse tRNA genes. The mice ranged in age from 0.67-35 months.

Three of the 51 tRNAs showed Bonferroni significant DNA methylation changes with age (p-value $< 1.08 \times 10^{-4}$) and all were in the hypermethylation direction. These three are tRNA-Asp-GTC-1-12, tRNA-Ile-AAT-1-4, tRNA-Glu-TTC-1-3 (Figure 4.27). Full Age model Results are available in appendix 6.9.

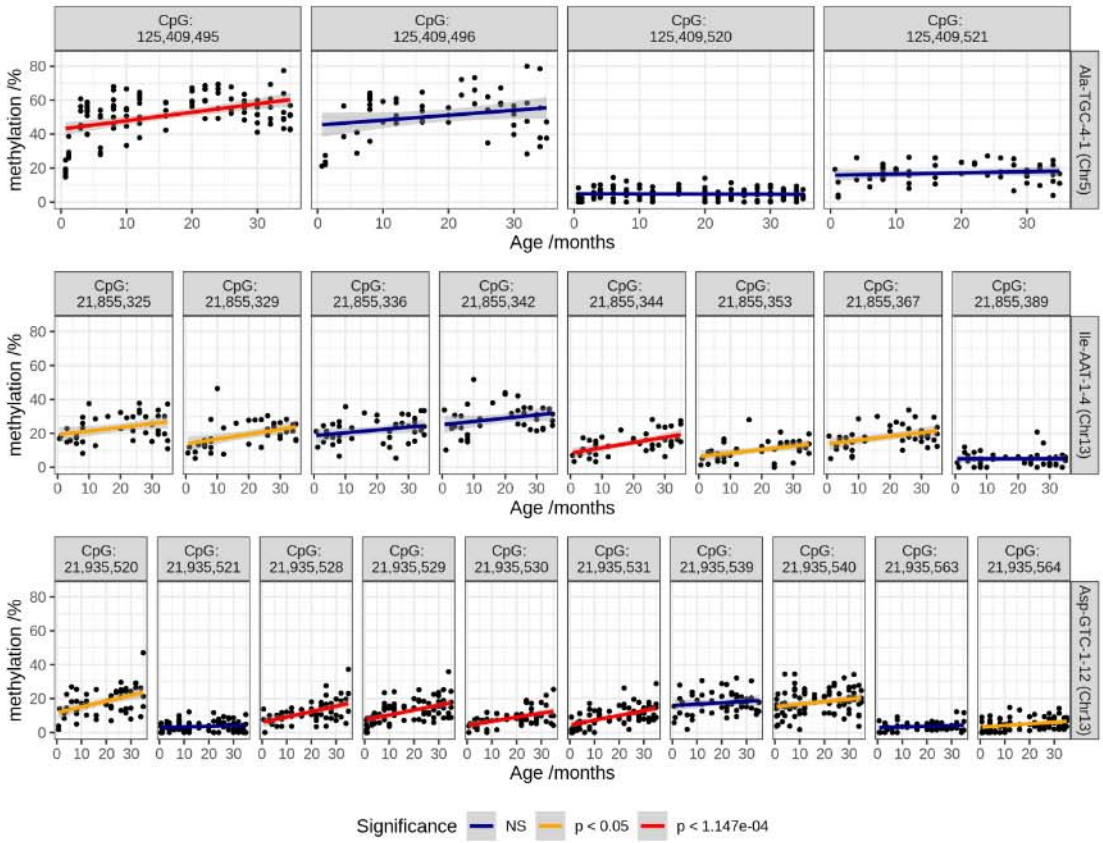

Figure 4.27: DNA methylation of CpGs in 3 tRNA which significantly hypermethylate with age in mice, in data from Petkovich et al. [375]. 6 CpGs reach Bonferroni significance and 7 show nominally significant increases.

In order to investigate the mouse results further data from Thornlow et al. [380], which had established tRNA ortholog sets for 29 mammalian species, were used. They identified 197 mouse tRNAs as having direct human orthologs with 44 of these included in the mouse results from Petkovich et al. [375]. However, unfortunately, although 2 of the top 3 tRNAs (tRNA-Ser-AGA-2-6 & tRNA-Ile-AAT-4-1) have mouse orthologs (tRNA-Ser-AGA-2-2 & tRNA-Ile-AAT-1-1), they are not covered in these mouse data. Furthermore, none of the tRNAs showing significant hypermethylation in mice (at nominal $p<0.05$) had human orthologs in our MeDIP-seq study-wide significant hypermethylating set. Therefore, whilst specific tRNA loci cannot be directly compared due to this lack of coverage, it is interesting that the small number of significant tRNA genes in the mouse data also hypermethylate with age.

4.6 Discussion

This work has identified a previously unknown enrichment for age-related epigenetic changes within the tRNA genes of the human genome. This observation was strongly directional with increasing DNA methylation with age [381]. The MeDIP-seq dataset employed brought advantages in exploring this undefined terrain of the tRNAome. Firstly, being genome-wide it provides much increased access, as these regions are poorly covered by current arrays. Secondly, being a fragment-based regional assessment of DNA methylation, the individual but highly similar small tRNA genes can be surrounded by unique sequence.

It was determined by genome-wide permutation that this strong hypermethylation signal was specific to the tRNAome, and not merely driven by the underlying CpG density of these loci. A targeted BiS-seq experiment validated the defined nature of the tRNA change in an independent dataset, with a pooled sample approach, which may also be useful for other ageing-related targeted DNA methylome evaluations. Additionally, support was gained for these results from limited DNA methylation array data.

Whilst the changes in DNA methylation observed here are relatively small (i.e. 3.7% between 4-year to 78-year pools in tRNA-iMet-CAT-1-4), this is consistent with the effect size seen in the majority of positions differentially methylated with age in many other studies [120,382], except for the noted extreme outliers, such as *EVOLV2* [383]. Furthermore, effect size cut-offs have distinct limitations, they do not generalise well between studies and they have been observed to remove a large fraction of true age-DMPs [384].

Additional exploration of what was driving this age-related phenomenon and its possible biological implications was undertaken. This result was observed in DNA from peripheral blood, a source with a heterogeneous cell-type population [225]. Moreover, there are well known age-related proportional changes in peripheral blood cell composition [377]. The TwinsUK MeDIP-seq and 450k array DNA methylation data included measured haematological values. Therefore, major cell type effects such as a myeloid skew were adjusted for, and distinct tRNAs were still significant. Although, a caveat to this is that one cannot exclude changes in minor specific sub-cell fraction types. However, that these age-related effects were strong enough to be observed in both a regional MeDIP-seq assessment, DNA methylation arrays with minimal coverage of the tRNA genes, and a pooled sequencing approach, implies that they not extremely subtle. Age-related tRNA gene DNA methylation changes were observed in the limited subset of mouse tRNA genes covered in publicly available RRBS data (~13%) and were able to identify tRNAs exhibiting DNA hypermethylation with age in this set. This suggests that age-related tRNA gene hypermethylation may not be unique to humans, and may extend at least across mammals.

The high number of hypermethylating tRNAs prior to cell-type correction pointed to the possibility that the epigenetic state of this small tRNAome fraction of genome could capture cell-type.

tRNA gene DNA methylation was found to be capable of separating myeloid from lymphoid lineages. There also was some suggestion of more fine-grained blood cell-type signatures in tRNA DNAm, such as the separation of CD19+ B cells from CD4/8+ T cells. Ageing is also known to lead to an increase in senescent cells (*e.g.* CD8+ CD28- cells). Whether these epigenetic changes in the tRNAome uniquely represent these cell-types will require technical advances to enable future single cell DNA methylome analysis to accurately assess these regions. If further supported, the epigenetic state of these loci may aid the taxonomy of cell-type definition.

This signal within the tRNA families was observed to occur at specific Isodecoders. After correcting for major cell types, two tRNA genes, tRNA-iMet-CAT-1-4 and tRNA-Ser-AGA-2-6, were identified which had the most consistent hypermethylation across 3 different assays. Isodecoders expand in number with organismal complexity and the high prevalence in mammals has been suggested to be due to their additional regulatory functionality [385,334]. They also have distinct translational efficiency [386], which can also have consequences in human disease [327]. Furthermore, there is great complexity to the fragmentation of tRNA [340], with physiological processes such as stress shown to induce fragment production [387]. These resultant tsRNAs can feedback on protein synthesis by regulating ribosome biogenesis [338] and others have diverse regulatory functions such as targeting transposable element transcripts [339]. They are also observed to circulate in the blood in a cell-free fashion, and fragment levels can be modulated by ageing and calorie restriction [346]. The isodecoder specific nature of our findings frame a possible hypothesis for regulatory change with age and future work will be required to unravel this potential.

There was an inconsistent result for tRNA-Ile-AAT-4-1, it is covered by a MeDIP-seq 500bp window which exhibited genome-wide significant hypermethylation, but also partially overlaps tRNA-Ser-AGA-2-6 (Figure 3B). Whilst the 450k array probe overlapping tRNA-Ile-AAT-4-1 (cg06382303) appears to replicate this hypermethylation, it is a borderline case for exclusion flagged by Zhou et al. [223] due to non-unique 30bp 3' subsequence. In the targeted Bisulfite sequencing data, tRNA-Ile-AAT-4-1 exhibited a loss of methylation. Therefore, this may suggest that the hypermethylation signal observed at this locus in the MeDIP-seq data could have been 'pulled up' by the neighbouring tRNA-Ser-AGA-2-6 hypermethylation signal. Of the 16 study-wide significant tRNA genes, only these two have a shared significant window, furthermore, in the expanded set of 44 only tRNA-Thr-AGT-1-2 could be similarly affected.

Codon usage frequencies have been claimed to be mostly invariant in the transcriptomes of a wide range of tissues, as well as across developmental time [378], Although, others have found the majority of isodecoders are transcribed in different cell types [303].

However, differences in the codon usage of genes with highly tissue-specific expression patterns have been observed [388]. Transcriptome codon usage frequency and corresponding tRNA gene expression levels have been claimed to vary with several other factors; notably, the replicative

state of cells, as well as across developmental time [389]. Experimental stress-related states have revealed changes with an over-representation of codons that are translated by rare tRNAs [323]. Differences in the codon usage of genes with highly tissue-specifically expressed genes have been observed [390]. It has however been argued that these differences are substantially explained by variation in GC content [391], and the extent to which codon usage plays a role in regulation of translation remains contested. Changes in the epigenetic state of specific tRNAs could be modulating transcription efficiency or even codon availability in the ageing cell. tRNA gene dosage is quite closely matched to amino acid usage frequency in the human exome.

The location of tRNA-Ser-AGA-2-6 and tRNA-Ile-AAT-4-1 immediately downstream of *CTC1* and of tRNA-Ile-AAT-4-1 within the 3'UTR of an alternate isoform of *CTC1*, which undergoes nonsense mediated decay raises the possibility that the gene body epigenetic regulation of *CTC1* may impact on the state of these tRNA genes. *CTC1*'s function in telomere maintenance [392], DNA replication licensing [393], and it's role in a rare progeroid condition [394] indicate that it has ageing-relevant biology. The possible relationship between the regulation of *CTC1* and that of the tRNA genes downstream of it warrants further study.

Several transcription factors acting via TFIIIB [317] have a negative (the tumour suppressors p53 [315] and Rb [316]) or positive (the proto-oncogene c-Myc) influence [317]. Regulatory sequence in the flanking or the internal regions of tRNA genes do not explain tRNA expression variation [389]. Whilst DNAm is able to repress the expression of tRNA genes [320] in a plasmid expression system, the broader chromatin environment also affects tRNA transcription. Due to the co-ordinated nature of epigenomic modifications, it may also be revealing to evaluate ageing-related histone modification in these tRNA loci.

tRNA sequences themselves are under strong structural (both secondary and tertiary) [385] as well as functional constraint, which leads to an order of magnitude reduction in variation compared the background genomic mutation rate [303]. However, polymorphic tRNA could be another potential caveat to this work. Although, there is no significant population variation in, for example, tRNA iMet sequences in 1,000 Genomes data. Indeed, there are only 11 new isodecoder sequences with high confidence (tRNAscan scores ≥ 50) at >1% population frequency [303]. Despite strong purifying selection maintaining very low variation in tRNA gene bodies, tRNA genes are subject to high levels of transcription-associated mutagenesis (TAM) leading to elevated mutation rates over evolutionary time in their immediate flanking sequences [395]. There is also some evidence for tRNA copy number variation at specific loci, although this remains under-characterised [396,397]. Another potential cause considered was whether age-related somatic copy number increases could be occurring in these loci. Population or somatic copy number expansions could lead to increased methylated reads in MeDIP-seq without any epigenetic state change. However, this would not be consistent with the targeted and array BiS conversion methodologies, where the proportion of methylated to unmethylated reads would still

be constant.

It is worth noting the parallels with known cancer and ageing epigenetic changes, and that tRNAs are also dysregulated in cancer [319], with proposed utility as prognostic markers [318]. Furthermore, the early replicating state of tRNA loci, potentially associated with high expression [398], may make them prone to hypermethylate, as is observed in early replicating loci in both cancer [399] and senescent cells [400]. Interestingly, tRNA gene loci may also play a role in local as well as large scale genome organisation [Hamdani2019; [354]]. tRNA gene clusters act as insulators [353] and have extensive long-range chromatin interactions with other tRNA gene loci [354]. The coordinated transcription of tRNAs at subnuclear foci and the B-box sequence elements bound by TFIIIC and not polIII may represent an organising principle for 3D-chromatin by providing spatial constraints [355]. Therefore, these tRNA epigenetic changes could contribute to the structural changes that are also observed in ageing [401].

A predominantly unmethylated state across fetal (Figure 4.23) and adult tissues (Figure 4.22) was observed at tRNA gene loci, consistent with the high rate of transcription at many tRNA gene loci. We suspect that the tRNA genes largely remain unmethylated through development and that the moderate increases in DNAm that are observed with age at these loci are being driven by changes arising primarily in older individuals. Distinct biological changes have been observed recently in aged individuals [402,403]. This would also be consistent with the lack of significant differences in the tRNA loci detected between the neonate and the 26-year-old adult in the Heyn et al. [161] data. This low baseline DNA methylation of the tRNA genes may also explain why the observed age related changes are predominantly hypermethylation. Whether this is driven by mechanisms, such as aberrant DNA methylation targeting of the tRNA loci or specific sub-celltype effects with age, will require further experimental investigation.

The attempt to estimate tRNA transcription by identifying fragments which may be derived from pre-tRNA sequences has serious limitations. It assumes that pre-tRNA levels reflect the amount of tRNA transcription, however, this may not be the case as pre-tRNAs are rapidly processed to mature tRNAs. The requirement for a read length of at least 40nt only has only a limited ability to distinguish full length tRNAs from tRNA derived small-RNAs, and the short read length of the data prevented the use of a higher threshold. Furthermore the limited number of reads unambiguously mapping to a specific tRNA locus limits the utility of this approach for inferring the effects of epigenetic changes at a specific locus on the expression at that locus. tsRNA abundance has been associated with locus specific tRNA gene expression, in some cases independent of mature tRNA levels [331]. This has important implications for the interpretation of these results given the multi-copy nature of genes like tRNA-iMet-CAT-1-4, as even if expression levels of mature iMet tRNAs are unaffected by changes in one copy's DNA methylation, these changes could still influence the levels of particular tsRNAs derived from specific tRNA loci. The implications of these changes in DNAm levels at tRNA genes for

biological ageing warrant further exploration.

In conclusion, due to the unique challenges that make the tRNAome difficult to examine it has remained epigenetically under-characterised despite its critical importance for cell function. This work directly interrogates the epigenetic DNA methylation state of the functionally important tRNAome, across the age spectrum in a range of datasets as well as methodologies and identified an enrichment for age-related DNA hypermethylation in the human tRNA genes.

Chapter 5

DNA methylation clocks in Alu Elements

5.1 Abstract

Alu repeat elements the most numerous repetitive sequence element in the human genome making up ~10.7% of total genomic sequence. Alus are generally repressed but contain features with the potential to have regulatory impact were they accessible. Loss of the repression of repetitive elements is thought to occur with age, due in part to a reduction in global DNA methylation levels driven by repeats. This change in DNA methylation in repetitive elements with age may capture information about biological ageing not captured by other such DNA methylation based predictors. Age predictors trained using elastic net regression on MeDIP-seq data for Alus (n = 774) were able to predict the age of an unrelated set of (n = 664) individuals with and R of 0.65 and a median absolute error of 8.1 years. The Alu age estimators were prone to overestimating the age of older samples and underestimating the age of younger samples. The difference between predicted age and chronological age, the Alu age acceleration was used to perform GWAS. The top 4 SNPs common between GWAS for 2 different Alu age predictors showed associations with phenotypes plausibly related to biological ageing. However, Alu Age acceleration was observed to be strongly correlated with Chronological age rendering it difficult to separate genetic associations with Alu age acceleration from those with chronological age. This association may be due to the relatively poor quantitation of absolute methylation levels by MeDIP-seq resulting in bias in the age prediction model.

5.2 Introduction

As we age biological systems from the molecular to the macroscopic undergo functional decline. This increasingly impaired function with time results in increased risk for a wide variety of chronic disease states and indeed for co-morbidities with these conditions [288]. Among the molecular changes that occur with age are alterations to the epigenome [30].

Approximately 45% of the human genome is comprised of repetitive or transposable elements [404,405] (Figure 5.1). The repression of these elements to prevent their expansion or the impact of their latent regulatory potential from effecting genomic stability and regulation is a core function of epigenetic mechanisms [406]. The most common DNA modification, 5-methylcytosine (5mC) at CpG loci, exhibits the most well documented age-related changes, and plays a central role in the repression of repetitive and transposable elements [11]. DNA methylation also classically functions to repress gene expression at promoters [407], similar to the repressive role that is plays at repetitive elements. Beyond solely repressive effects, DNA methylation in gene bodies can modulate expression [408] and influence splicing patterns [409].

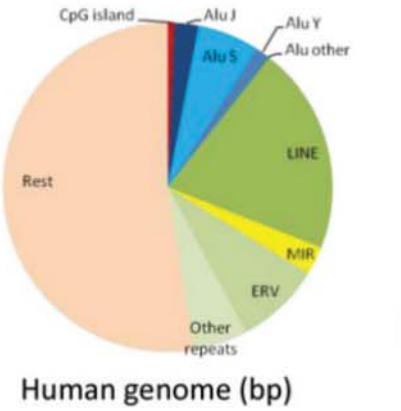

Figure 5.1: **Proportion of the Human genome given over to different repeat classes** 'Rest' indicates the non-repeat features. Reproduced form Jorda et al. [410] Figure 1 b.

Demethylation of repetitive elements is thought to be a major contributor to the global decreases in bulk 5mC with age [138,411]. Some regions of the genome show increasing DNA methylation with age, notably, the promoters of polycomb target genes [155], bivalent chromatin domains [153] and specific tRNA genes Chapter 4 [1].

Alu elements are the single most abundant transposon in the human genome with ~1.19 million copies and comprising ~10.7% of the total genomic sequence [412,413]. Alu elements are members of the short interspersed nuclear elements (SINEs) family of non-LTR retrotransposons, are ~300bp in length, and are derived from the 7SL RNA gene [414]. SINEs are non-autonomous,

unlike LINEs (short interspersed nuclear elements) they do not encode their own reverse transcriptase (RT) but instead replicate making use of an RT encoded in an active LINE element, such as one of the small number of active LINE-1 elements [413,415]. The Alu elements arose from the 7SL sequence via the FAM sequences, which gave rise to truncated variants, FRAM and FLAM the fusion of which yielded the current dimeric Alu sequence [416].

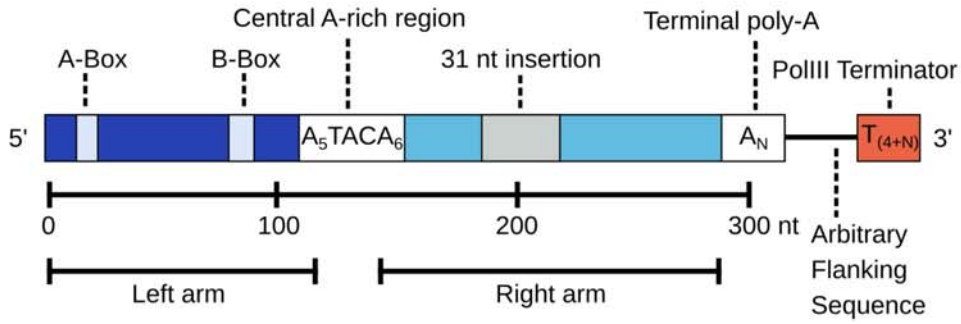

Figure 5.2: **Structure of an Alu element** A complete Alu element contains the A and B box sequence elements from an RNA polIII promoter, a central A rich region, a short insertion in the right arm, and a poly A region involved in retrotransposition and insertion. There is not an explicit polIII terminator in the Alu sequence so transcription continues through arbitrary flanking sequence until a run of at least 4 Ts, is encountered [417,418]. FLAM sequences are the origin of the Left arm and FRAMs the origin of the Right [416].

Alus are primate specific and arose ~65 million years ago (MYA), with a peak in amplification around ~40 MYA (Figure 5.3) [419]. Non-LTR retrotransposons, including Alu, L1 and SVA elements, are active in humans. Mutagenesis arising from Alu element insertion and Alu element mediated recombination is responsible for an estimated 0.4% of all human genetic disorders [405,420]

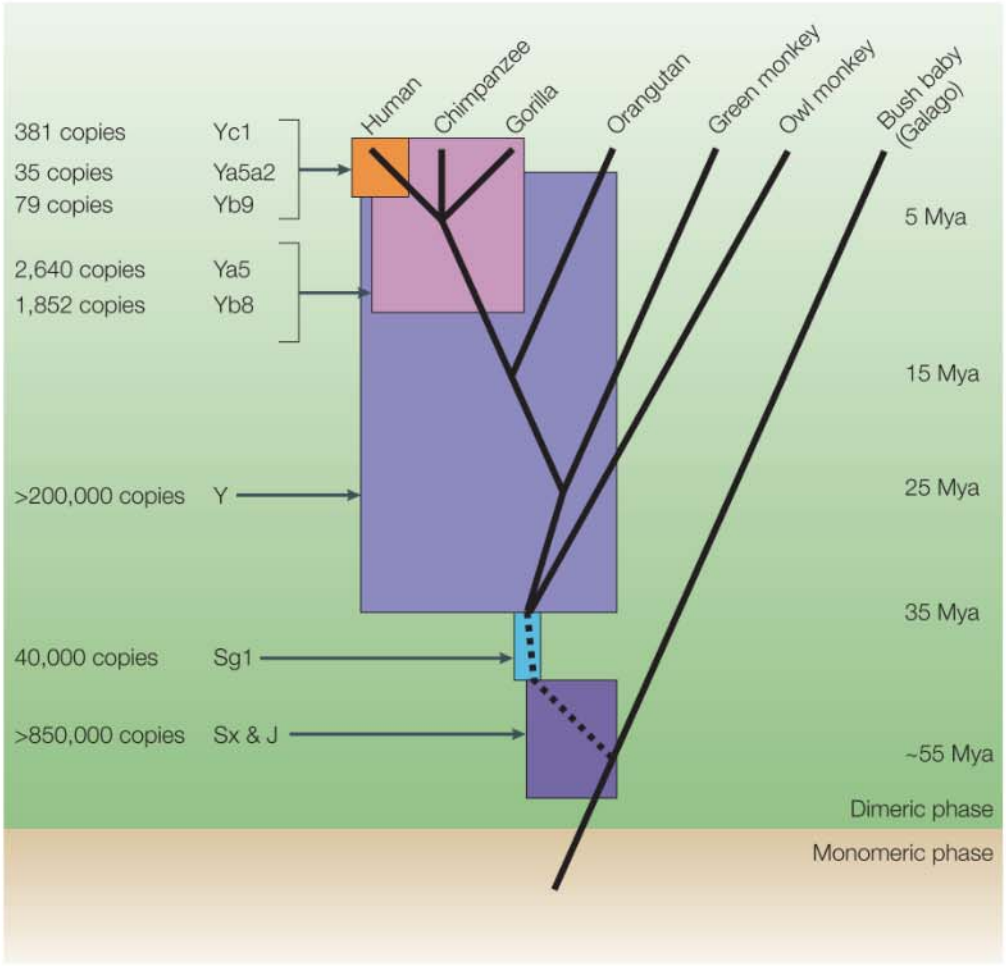

Figure 5.3: **Alu element Age and history of expansion in primates** Mya (milion years ago). Reproduced from Batzer & Deininger 2002 [418].

Alu elements are CpG rich (Figure 5.4), accounting for ~7.5 million CpGs, or ~25% of all human CpG sites (Figure 5.5) [413]. High levels of DNA methylation predominate at Alu elements across normal tissues. Furthermore Alu elements are more resistant to hypomethylation than other repetitive elements in cancer cells. This is indicative of strong pressure to maintain their repressed state [410].

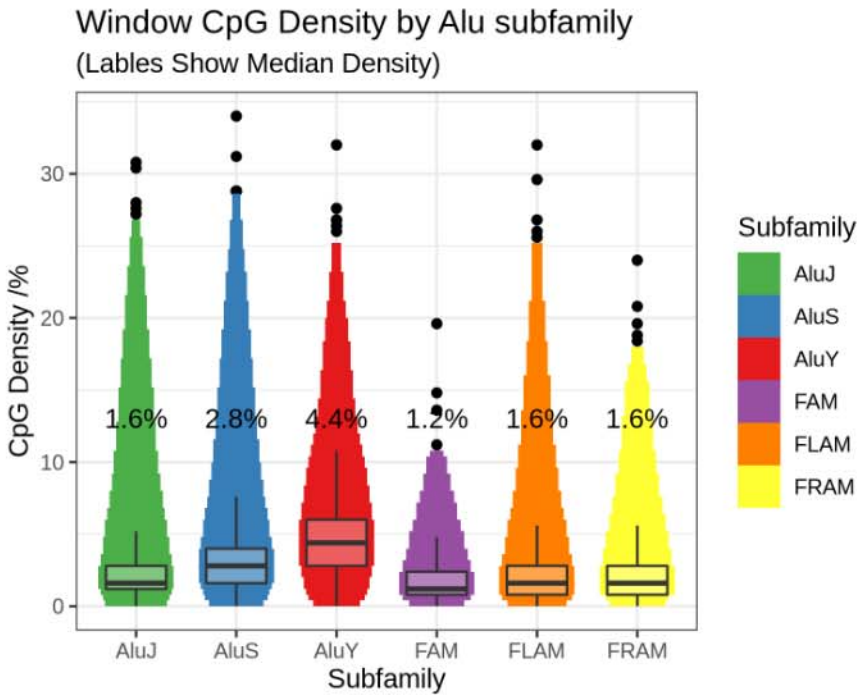

Figure 5.4: **Alu elements are CpG dense** FAM, FLAM and FRAM sequences are the precursors which gave rise to the modern Alu families, they are less CpG dense than the AluS and AluY families and closer to the global average of ~1.8% Chapter 1.4.2.

Alu elements contain many sequences with the potential to effect gene regulation, and indeed do effect gene expression [421,422]. There are numerous TFBSs in the canonical sequences of elements in the Alu family [423], and some may possess enhancer activity [424]. DNA methylation levels at transcription factor binding sites (TFBS) modulate the binding affinity of transcription factors for these sites [425,88]. Thus deterioration in the effectiveness of the repression of Alu elements has the potential to disruptively expose this latent regulatory function [426,290,427,428]. Epigenetic reactivation of cryptic cis-regulatory elements (CREs) within transposons has been show to act as an oncogenic driver [429], including specifically from Alu elements [430]. Alu element methylation has also previously been correlated with age-related phenotypes. Hypomethylation of Alu elements in blood cells has been associated with increasing age and with lower bone mass [431]. Alu and LINE-1 elements were more methylated in the circulating cell-free DNA of women with an older age an menopause [432]. De-repression of Alu elements also has the potential to influence common non-malignant diseases including neurodegenerative and auto immune conditions [290,433].

Statistical methods for sparse regression analysis combined with quantitative assays of DNA methylation with high positional resolution, which cover a large number of loci, have permitted the construction of 'DNA methylation clocks' which predict chronological age with a high degree of accuracy [293,434]. The difference between chronological and predicted age know as the 'age ac-

celeration' has been found to capture aspects of biological ageing in a number of these clocks. For example, Age acceleration is a strong predictor of all-cause mortality [190]. Whilst DNA hypermethylation and hypomethylation are driven by distinct mechanisms [435,101] CpGs undergoing these processes are approximately equally represented in DNA methylation clocks. Interestingly however, hypomethylating CpGs were identified by Zhang et al. to be the most indicative of biological ageing-related all-cause mortality effects in blood [436]. Most DNA methylation clocks constructed in humans have been constructed using the Illumina bead chip array based DNA methylation assays. The coverage of these assays is strongly focused on promoters [154], CpG islands [162], and more recently enhancers [210]. However, array coverage of repetitive elements is lacking due in part to the technical challenges of accurate mapping in these spaces [434]. Despite this, a small number of CpGs in repeat elements are present in existing DNA methylation age clocks (Figure 5.5). Thus, much of the genomic hypomethylation goes uncovered by the assays most commonly used to construct human epigenetic clocks.

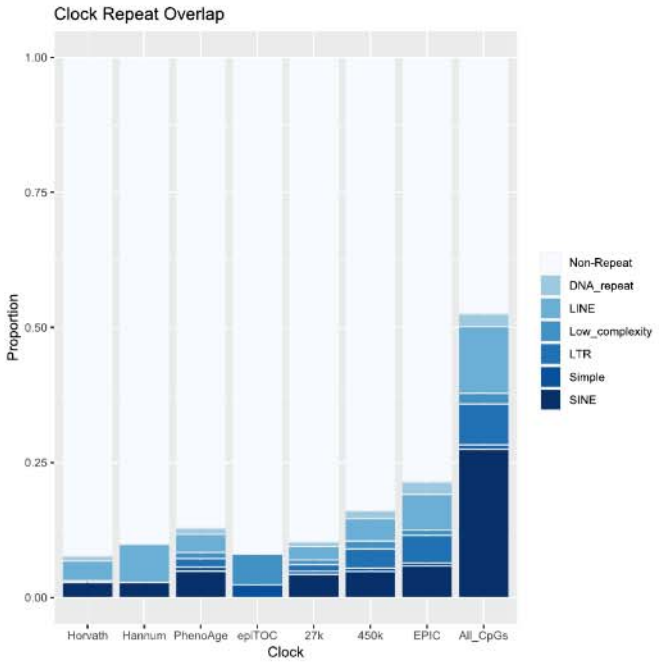

Figure 5.5: Comparison of the overlap of different repeat classes [437] with four DNA methylation clocks (Horvath clock [168], Hannum et al. clock [167], PhenoAge clock [193], epiTOC clock [372]) as well as Illumina array CpGs (27k [154], 450k [162], EPIC [210]) and all CpGs in the genome. Reproduced from Bell et al. [434] figure 2H.

In this study elastic net regression models were used to construct age predictors comprised of the DNA methylation state of Alu elements [438]. DNA methylation data at Alu elements was drawn from a MeDIP-seq dataset described in the Methods Chapter 2. Alu-specific DNA methylation ageing clocks were constructed and compared to other clocks' measures.

5.3 Methods

5.3.1 DNA methylome data

The Methylated DNA Immunoprecipitation sequencing (MeDIP-seq) data was processed as previously described [209,107] and detailed in Methods 2.2. These processed data are available from the European Genome-phenome Archive (EGA) (https://www.ebi.ac.uk/ega) under study number EGAS00001001910 and dataset EGAD00010000983 and were generated by BGI Shenzen for TwinsUK.

`MEDIPS` (v1.0) was used for the MeDIP-seq analysis [128]. This produced reads per million base pairs (RPM) values binned into 500bp windows with a 250bp slide in the BED format, resulting in ~12.8 million windows on the genome.

5.3.2 Alu element annotation and pre-filtering

Alu element annotations were taken from Repeat Masker [437], only Alus on the autosomes were considered. 2,758,588 500bp windows with a 250bp slide overlapped the 1,137,653 Alu elements annotated.

Technically problematic and genetically confounded Alu element loci were removed. Alu elements located in blacklisted 834 regions of the hg19 genome, as identified by Amemiya et al. [357] were excluded as candidate loci when constructing DNAm clocks. Alu elements overlapping known structural variants identified by the 1000 genomes project [439] were excluded as candidate clock loci. Alu insertion variants can alter splicing patterns, this can affect disease risk, for example a particular Alu insertion elevates the risk of multiple sclerosis [440]. Polymorphic Alu elements are disproportionately represented at disease associated loci in GWAS [433] Genetic variants which impact on the rate of biological ageing could confound DNA methylation ageing signal hence the exclusion of know polymorphic Alu Loci. Haplotype specific methylation peaks, in which a common disease associated genetic variant is predictive of DNA methylation state, were also excluded from the clock; Bell et al. identified 7,173 such regions [107]. To exclude loci with low coverage which are likely to be uninformative and noisy a zero filter was used. In quantile normalised data loci in which 0.1% or more of samples have an RPM value of 0 are excluded.

5.3.3 Elastic net regression

`Bigstatsr` was used to fit penalised regression models [438]. Elastic net regression with an $\alpha = 0.5$, and cross-model selection and averaging (CMSA) with 10 fold cross-validation was used. This procedure obviates the necessity of selecting a λ regularisation hyperparameter. CMSA generates model coefficient by averaging across coefficients generated by cross-validation

[438,169,441]. Details of the training and replication set to which elastic net regression was applied are detailed in Sample Selection 5.3.4.

5.3.4 Sample Selection

Two training and two replication sets were created from the twin data so as to permit the training and replication sets to be unrelated. The training sets, which were permitted to contain singletons, contained 1548 individuals and the replication sets contained the 1308 individuals who where twins of those in the replication sets. The 1548 individuals in the training sets were randomly split into 2 groups of 774 the replication set for each training set was comprised of the twins of the other training set such that two training sets with unrelated replication sets could be constructed (Figure 5.6). Giving a total of 2856 individuals in all sets, 2617 were female and 239 male, sex was approximately equally distributed between sets with all sets being ~8% male.

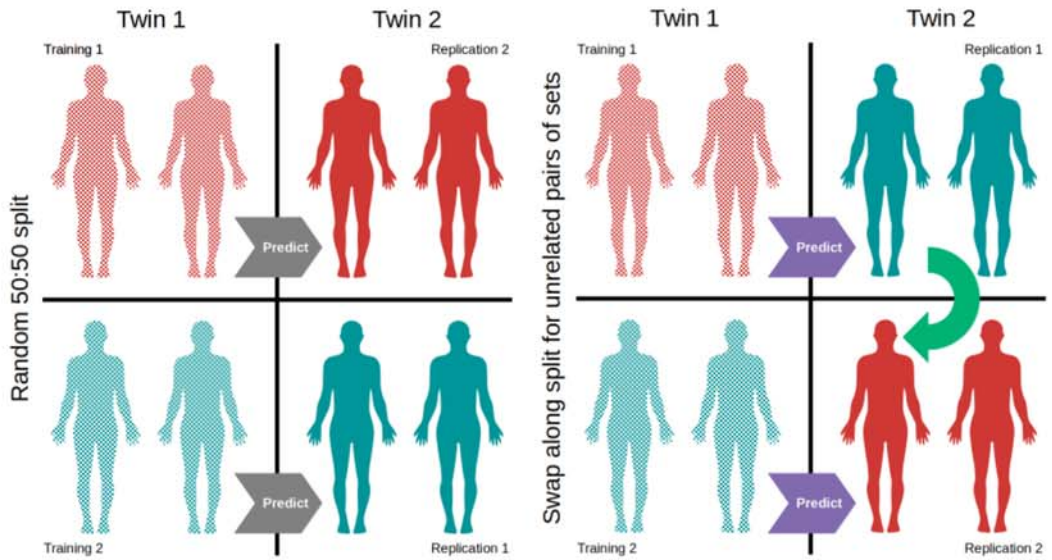

Figure 5.6: **Construction of training and replication sets** Twin pairs were split one twin in each group. The first group of twins was split into two groups with a random number generator. The twins of the individuals in one random half of the group were assigned to be the replication set for the other half, producing two sets of unrelated training and replication groups. Models were also used to predict the age of related individuals to gauge the effect of genetic factors on the predictive power of the models.

The age distribution of these training and replication sets are highly similar (Figure 5.7).

Figure 5.7: **Age distribution of training and replication sets** Training 1 (n = 774, 64M/710F); Training 2 (n = 774, 65M/709F); Replication 1 (n = 664, 55M/609F), unrelated to Training 1; Replication 2 (n = 644, 55M/589F), unrelated to Training 2. M = Male, F = Female

All sets are comprised of ~2/3 Monozygotic (MZ) twins and ~1/3 Dizygotic (DZ) twins, (exact numbers are detailed in Figure 5.8).

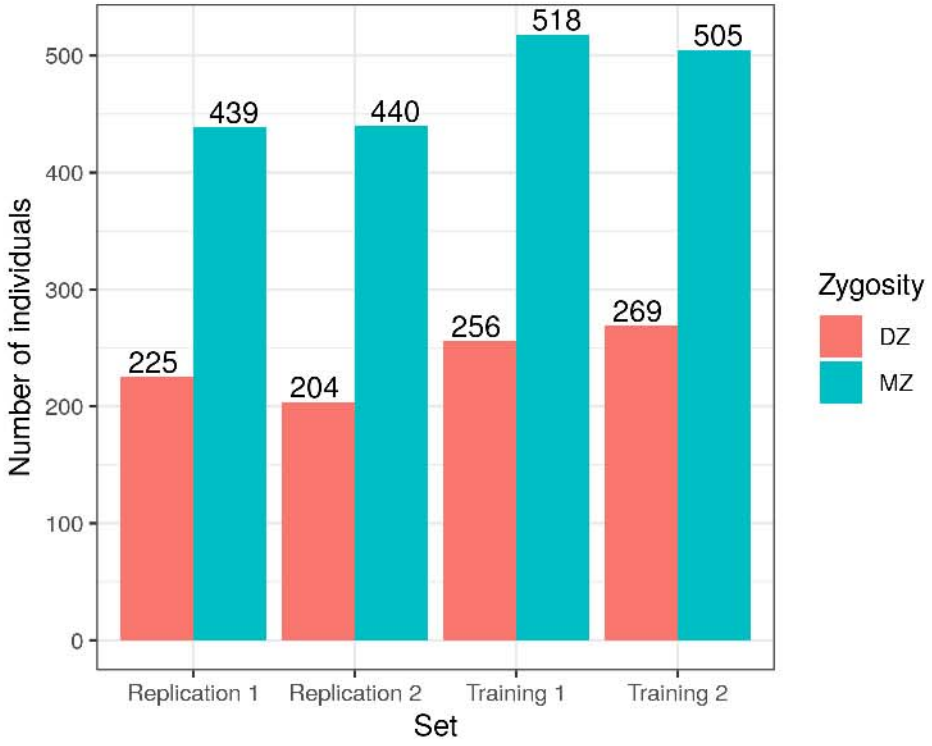

Figure 5.8: **Zygosity Make-up of sets** Replication 1 (n = 664, 225 DZ, 439 MZ); Replication 2 (n = 644, 204 DZ, 440 MZ), Training 1 (n = 774, 256 DZ, 518 MZ); Training 2 (n = 774, 269 DZ, 505 MZ).

A variety of different initial filtering criteria were applied to the Alu elements included in the models, as well as a number of data transformations. Pre-filtering steps were performed partially for quality control and partially to see if subsets of the Alu space with biologically interesting differences could yield clocks of equivalent accuracy but which might capture different aspects of biological ageing related to repeat de-repression.

Preliminary analysis were carried out using RPM values, Quantile normalised RPM values, Absolute methylation value estimates from `MEDIPS` and Binarised DNA methylation data. Quantile normalised RPM values yielded the best and most consistent results and were used in subsequent analyses.

Whilst Human epigenetic age is logarithmic with time across the lifespan the effects of this are more relevant at younger ages [442]. We fitted Alu Clock models on the natural logarithm of age in years this did not improve the quality of predicted to actual age values, likely due in part to the older age distribution of our data.

An Alu element length filter of at least 180bp or approximately two-thirds the length of a full length Alu element (~280bp) was imposed. This minimises the number of reads with the potential to misalign to other Alu elements, reads which, in this context, could lead to reduced accuracy

of positionally specific Alu DNA methylation estimates [443].

CpG density filters were employed to prioritise regions likely to have dynamic and functionally interesting differences in DNA methylation. Filters were used for CpG densities above the genomic background of 2%, as well as for >5% the approximate threshold for CpG island shores and low methylation regions [84,85].

To limit the selection of predictive sites to those likely to have consistent directions of change across samples a clock was constructed with only windows which show nominally significant changes in both members of an MZ twin pair in a consistent direction. Linear models to predict age based on DNA methylation were fitted for 687 pairs of monozygotic (MZ) twins. Batch information and blood cell counts (eosinophils, lymphocytes, monocytes, neutrophils) were included as covariates.

To produce a model that specifically captures age as a function of the de-repression of Alu elements clocks were also constructed with only windows which exhibited nominally significant ($p < 0.05$) decreases in DNAm in both MZ twins. This filtering was based on the model used in selecting windows with a consistent direction of change.

An age predictor was also constructed using only AluY elements. AluY is the youngest (Figure 5.3), and most CpG dense (Figure 5.4) family of Alu elements. Having a CpG density closer to that of typically de-repressed functional elements may make such elements more readily susceptible to stable de-repression. In addition the host genome has had less time to adapt to the presence of these elements so their de-repression may be more disruptive than older elements.

5.3.5 Binarisation

In Alu age prediction model for which Binarisation was used the binarisation of DNA methylation data was achieved by a simple procedure: assigning windows in a given sample a DNA methylation value of 1 if their DNA methylation value was greater than the median RPM value for this window across samples and a 0 if it was less.

Binarisation of gene expression data yielded increased predictive accuracy in gene expression based ageing clocks [444]. The reasoning behind this approach is that the quantitation of changes in gene expression is noisy and binarising the data may reduce the noise whilst retaining the information important for age prediction. Whilst this approach discards information about the level of expression, sacrificing power, its success in transcription data indicate that it makes up for this possibly through greater consistency between samples reducing overfitting to misleading technical noise.

5.3.6 Comparisons to other clocks in TwinsUK Illumina 450k array data

Illumina Infinium DNA methylation 450k arrays ((C) Illumina) were also performed on TwinsUK participants (Methods 2.1). 574 Blood-derived DNA samples had matched MeDIP-seq data not used in the training or replication sets in the Alu clock analysis. These data were available for this analysis in a pre-processed form, Methylation 'beta' values subject to beta-mixture quantile normalisation (BMIQ) as previously described [209,107].

DNA methylation ages were computed for the Horvath [168], Hannum et al. [167] and Levine et al. 'PhenoAge' [193] clocks, using the coefficients, intercept values and transformations specified the original publications, and a model of the form: $DNAmAge = a\beta_{probe\ 1}...b\beta_{probe\ n}+intercept$ Where a, b and *intercept* are model parameters provided by the authors of the respective DNAm age predictors. In the absence of an intercept value 0 was used, in the case of the Hannum clock. All age acceleration values were computed as the predicted age minus the chronological age, including for the Alu age predictors.

5.3.7 Genome Wide Association studies of Alu Age Acceleration

In collaboration with Pirro Hysi at King's College London Genome wide association studies (GWAS) were performed for Alu age acceleration. Genotypic and phenotypeic data from the TwinsUK adult twin registry (see methods 2), with 1108 individuals being analysed in the GWAS. Genotype data was generated with a combination of Illumina HumanHap300 and HumanHap610Q chips [445]. Intensity data for each of the arrays were pooled separately. Genotypes were called with the Illuminus32 calling algorithm, thresholding on a maximum posterior probability of 0.95. Pre-phasing was performed using Eagle v2.4 [446], and imputation was performed using Minimach 3.0 [447], using haplotype information from the HRC r1.1 2016 reference panel [448]. Linear mixed models were fitted to test for associations between the phenotypes and genetic variants using GEMMA v0.94 [449]. Family relatedness arising from the presence of siblings in the cohort was taken into account in constructing linear mixed models of each variant. Associations were considered suggestive at a p-value threshold of $p < 1 \times 10^{-5}$ and significant at $p < 5 \times 10^{-8}$.

Table 5.1: **Alu Clock Model Summaries** *A)* Models fitted with Training 1 set predicting Replication 1 ages. *B)* Models fitted with Training 2 set predicting Replication 2 ages. R is Pearson's correlation coefficient. RMSE = root mean squared Error, MAE = median absolute error. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus.

	Training Set	Prediction Set	R	R^2	RMSE	MAE	Avail. Windows	Clock Windows	Avail. Alu Elems.	Clock Alu Elems.	Description	% Coef. -ve	AluJ	AluS	AluY	FAM	FLAM	FRAM
			0.651	0.424	10.958	8.112	2,646,388	6,942	1,016,610	6,513	No Filters	46.744	24.3	59.4	13	0.389	2.39	0.533
			0.619	0.383	11.201	8.349	1,868,147	6,616	738,953	6,196	>180bp, 99.9% non-zero	47.415	25.4	59.9	11.9	0.378	1.96	0.514
A	Training 1	Replication 1	0.609	0.371	11.165	8.221	981,452	6,293	561,509	5,894	↓ CpG Density >2%	47.624	26.1	58.5	12	0.493	2.38	0.556
			0.591	0.349	11.343	8.360	1,868,147	2,131	738,953	2,066	↓ ln(Age)	46.739	25.5	59	12.6	0.235	2.21	0.516
			0.556	0.31	11.375	8.010	6,000	1,269	5,293	1,040	↓ Consistent Direction	47.518	21	60.3	14.7	0.394	2.76	0.63
			0.524	0.275	11.542	8.541	192,299	4,884	138,506	4,575	↓ CpG Density >3%	48.464	26.4	57	12.8	0.389	2.78	0.696
			0.432	0.186	12.152	9.006	4,389	1,187	3,708	924	↓ Hypomethylation	56.613	21.3	59.2	15.2	0.337	2.95	0.927
			0.292	0.085	12.879	9.614	92,473	1,880	62,231	1,769	↓ Only AluY Elems.	51.330	26.1	57.2	12.4	0.479	3.09	0.638
			0.615	0.378	12.021	8.753	2,646,388	7,091	1,016,610	6,610	No Filters	46.002	25.3	59.2	12.4	0.368	2.31	0.522
			0.582	0.339	11.828	8.442	192,299	5,007	138,506	4,650	↓ CpG Density >3%	48.252	26	58.5	12	0.499	2.52	0.519
B	Training 2	Replication 2	0.576	0.332	12.332	8.903	1,868,147	6,455	738,953	6,021	>180bp, 99.9% non-zero	46.636	24.6	59.3	12.6	0.389	2.09	0.451
			0.572	0.328	11.844	8.928	1,868,147	2,621	738,953	2,546	↓ ln(Age)	44.830	23.9	60.5	12.2	0.496	2.29	0.534
			0.572	0.327	12.296	8.560	981,452	5,518	561,509	5,200	↓ CpG Density >2%	47.191	26	57.9	12.4	0.217	3.03	0.544
			0.520	0.271	12.249	8.186	6,000	1,238	5,293	1,026	↓ Consistent Direction	48.142	22.3	58.6	15.7	0.102	2.58	0.727
			0.404	0.163	12.953	9.330	4,389	1,045	3,708	853	↓ Hypomethylation	58.086	22.4	56.1	17.6	0.191	3.16	0.574
			0.290	0.084	13.657	9.469	92,473	2,656	62,231	2,478	↓ Only AluY Elems.	50.075	26	59.2	11.3	0.301	2.52	0.678

Table 5.2: **Summaries of Alu Clock Models Predicting Twin Groups** for the purpose of comparing predictive of age models in related and unrelated samples. *A)* Models fitted with Training 1 set predicting Replication 2 ages. *B)* Models fitted with Training 2 set predicting Replication 1 ages. R is Pearson's correlation coefficient. RMSE = root mean squared Error, MAE = median absolute error. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus.

	Training Set	Prediction Set	R	R^2	RMSE	MAE	Avail. Windows	Clock Windows	Avail. Alu Elems.	Clock Alu Elems.	Description	% Coef. -ve	AluJ	AluS	AluY	FAM	FLAM	FRAM
			0.677	0.458	11.380	8.191	2,646,388	6,311	1,016,610	5,930	No Filters	46.427	24.7	59.8	12.3	0.38	2.33	0.475
			0.644	0.415	11.626	8.388	1,868,147	6,327	738,953	5,948	>180bp, 99.9% non-zero	47.432	25.1	60.2	11.6	0.253	2.29	0.516
A	Training 1	Replication 2	0.643	0.413	11.600	8.240	981,452	6,610	561,509	6,215	↓ CpG Density >2%	47.428	26.3	58.7	11.8	0.424	2.24	0.454
			0.601	0.361	11.593	7.908	192,299	5,052	138,506	4,707	↓ CpG Density >3%	48.733	25.9	57.6	12.5	0.337	2.99	0.673
			0.575	0.330	11.854	8.313	6,000	1,208	5,293	1,005	↓ Consistent Direction	48.427	20.9	61.3	14.4	0.248	2.73	0.331
			0.569	0.324	11.710	8.531	1,868,147	2,401	738,953	2,334	↓ ln(Age)	46.397	24.5	59.8	12.3	0.25	2.54	0.583
			0.492	0.242	12.359	9.165	4,389	1,297	3,708	1,002	↓ Hypomethylation	54.819	21.7	59.8	14.9	0.463	2.47	0.694
			0.302	0.091	13.450	9.342	92,473	1,951	62,231	1,839	↓ Only AluY Elems.	51.205	25.1	59	11.8	0.564	2.97	0.615
			0.641	0.411	11.171	8.047	2,646,388	6,857	1,016,610	6,406	No Filters	45.909	25.3	59.6	12	0.345	2.06	0.613
			0.556	0.309	11.345	8.216	1,868,147	2,052	738,953	1,986	↓ ln(Age)	43.275	23.6	61.6	12	0.439	1.8	0.439
B	Training 2	Replication 1	0.552	0.305	11.591	8.631	981,452	5,226	561,509	4,893	↓ CpG Density >2%	47.608	25.9	57.7	12.3	0.402	3.12	0.574
			0.574	0.328	11.624	8.658	1,868,147	5,907	738,953	5,545	>180bp, 99.9% non-zero	46.843	25.3	58.2	13	0.423	2.57	0.542
			0.508	0.258	11.342	8.038	192,299	4,821	138,506	4,449	↓ CpG Density >3%	47.832	26.1	57.7	12.6	0.436	2.47	0.622
			0.513	0.263	11.675	8.438	6,000	933	5,293	815	↓ Consistent Direction	48.553	22.3	59.1	15.3	0.322	2.14	0.857
			0.398	0.159	12.437	8.769	92,473	3,158	62,231	2,947	↓ Only AluY Elems.	48.892	26.2	59.8	10.2	0.538	2.47	0.665
			0.329	0.108	12.619	8.752	4,389	752	3,708	641	↓ Hypomethylation	61.057	23.1	54	18.9	0.266	2.93	0.798

5.4 Results

Table 5.1 summarises the details of Alu clocks trained with the various pre-filtering criteria applied to their starting sets. The correlation with the predicted age and the chronological age of samples unrelated to the individuals on which the clock was trained indicate model quality (Figure 5.9 a). The quality of the Alu clock predictions is quite poor in comparison to other DNA methylation clocks with the highest R of 0.65 and a median absolute error of 8.1 years, compared to 0.96 and 3.6 years for Horvath's multi-tissue clock [168].

Table 5.2 shows the model summaries when correlations are computed for for the related sets comprised of the twins of individuals in the training set. The correlations are mostly slightly higher for this prediction than for the prediction of unrelated individuals (Figure 5.9 b). This suggests that there are not large genetic effects on the models as their predictive efficacy is only marginally better when employed to predict the age of individuals closely related to the training set compared to unrelated individuals.

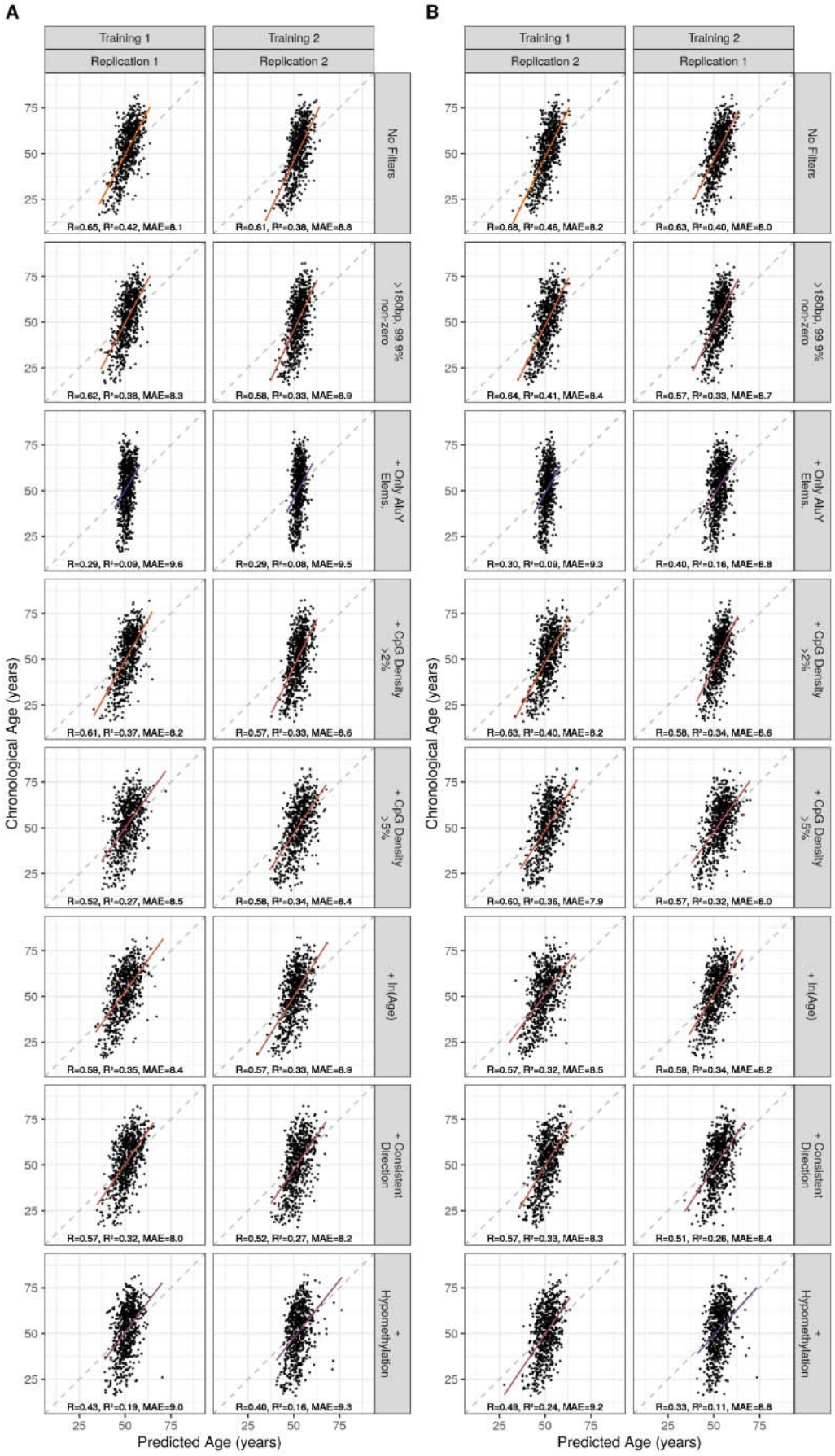

Figure 5.9: **Correlation of Predicted with Chronological Age Across Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus.

5.4.1 Comparison to other DNA methylation clocks

To compare the Alu clock results with those of other DNA methylation age clocks DNAm age was predicted for the: Horvath [168], Hannum et al. [167] and Levine et al. 'PhenoAge' [193] clocks (Methods 5.3.6) using samples for which both DNA methylation array data and MeDIP-seq data was available (n = 574). The array based clocks all perform similarly with R values in excess of 0.8, the Alu clock only attains an R of 0.46, though with a narrower error than PhenoAge. This however, is not entirely surprising given that PhenoAge aims to capture ageing phenotypes and not narrowly chronological age.

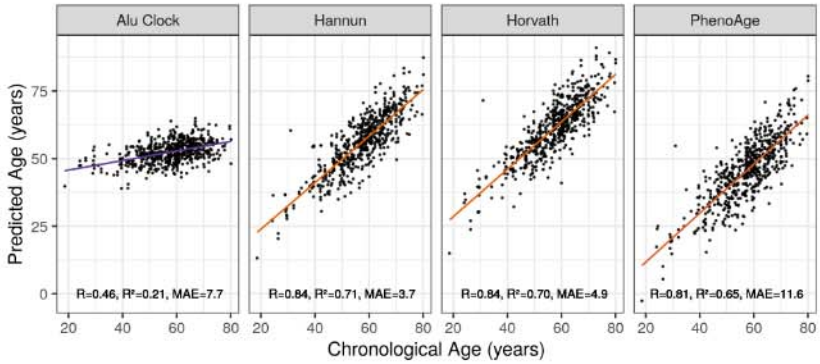

Figure 5.10: **Correlation of Alu Age Acceleration with Chronological Age Across Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. Alu Clock here refers to the unfiltered model trained on the Training 1 set.

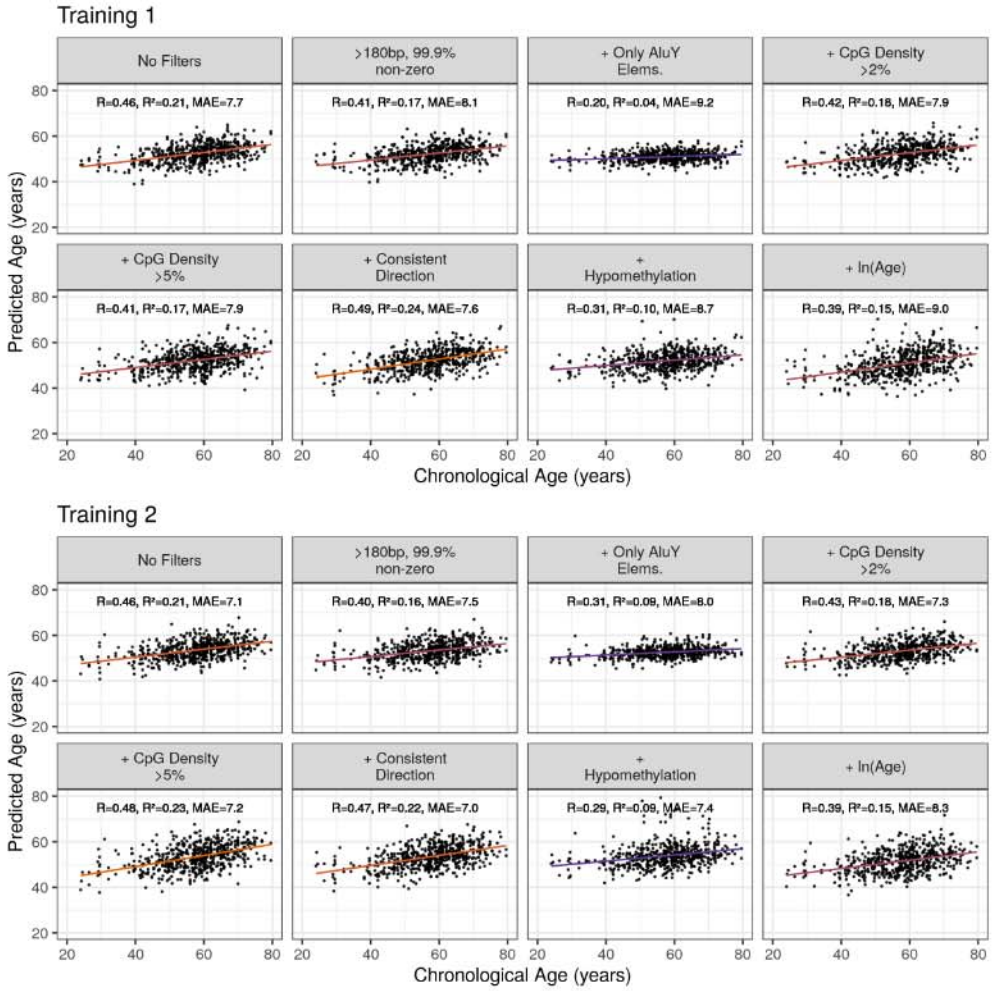

Figure 5.11: **Correlation of Alu Age Acceleration with Chronological Age Across Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. Alu Clock here refers to the unfiltered model trained on the Training 1 set.

5.4.2 Alu Age Acceleration

The age acceleration values (Chronological - Predicted Age) for the Alu clocks are strongly correlated with chronological age (Figure 5.12), unlike the age acceleration of the Horvath DNA methylation clock and others (Figure 5.13). The older the individual the more prone their age is to be overestimated and vice versa. This means that Alu clock age acceleration values cannot be used independently of chronological age to make statements about biological age.

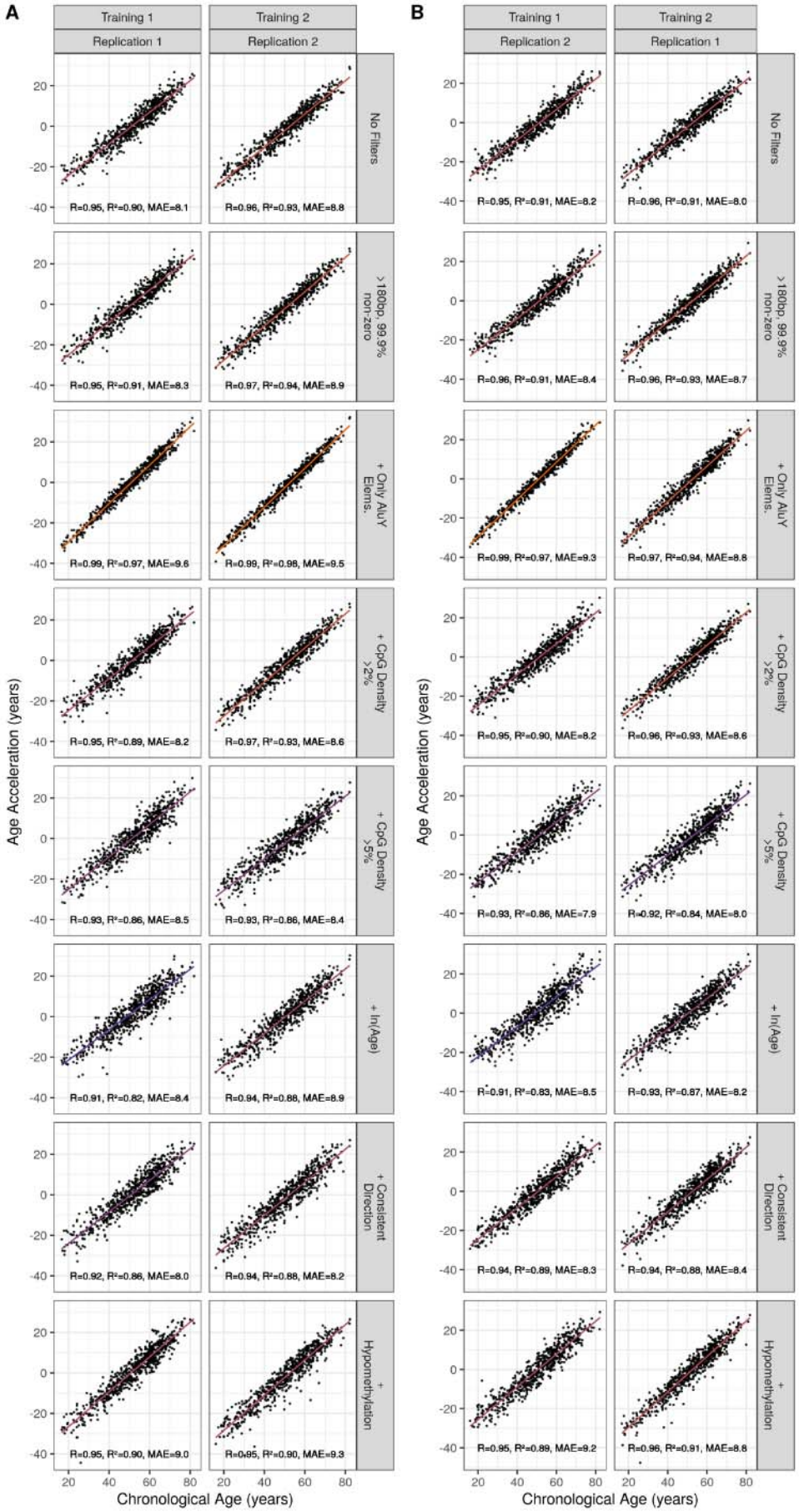

Figure 5.12: **Correlation of Alu Age Acceleration with Chronological Age Across Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus.

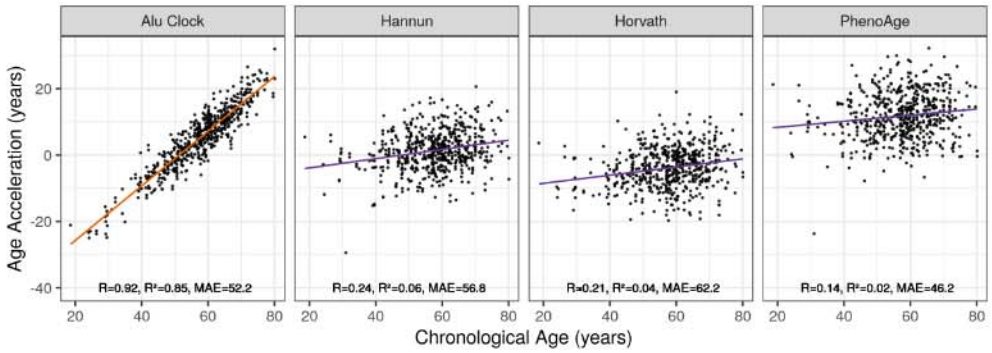

Figure 5.13: **Correlation of Age Acceleration with Chronological Age Across Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. Alu Clock here refers to the unfiltered model trained on the Training 1 set. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus.

Age acceleration in the Alu clocks is not strongly correlated with age acceleration in the Horvath, Hannun or PhenoAge clocks.

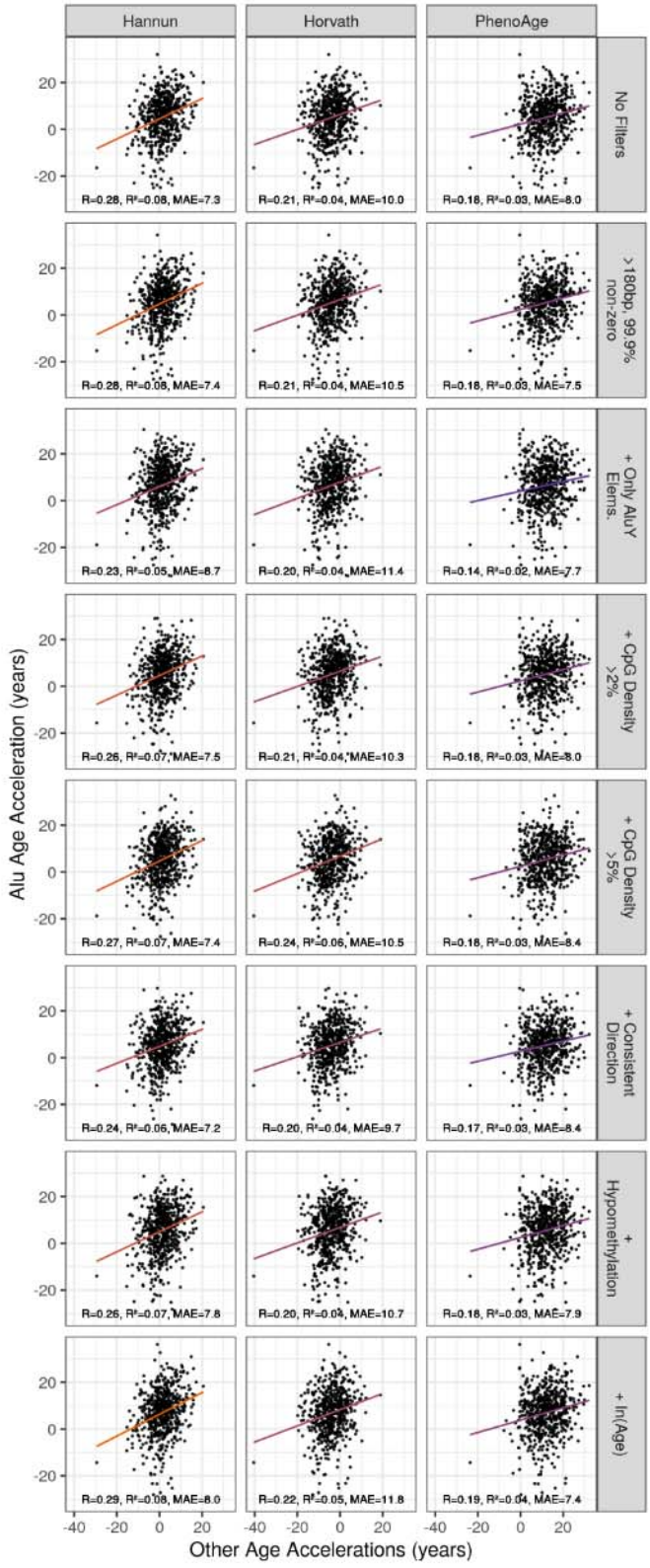

Figure 5.14: **Correlation of Alu Age Accelerations with Age Acceleration from Other Models** MAE = median absolute error. Orange = High R^2, purple = low R^2. Alu Clock here refers to the unfiltered model trained on the Training 1 set. '+' signifies this filter is in addition to element length >180bp and 99.9% non-zero RPM values at this locus.

5.4.3 Top results Genome Wide Association studies of Alu Age Acceleration are shared by Alu Age predictors

Genome wide association studies (GWAS) for Alu age acceleration using acceleration metrics from two models were performed. The models used were the unfiltered (model 1) and the element length and percent zero filtered (model 2) models trained on training set 1. The Manhattan plots in Figures 5.15 & 5.16 illustrate that GWAS for these two models have similar results with the GWAS for model 2 attaining some lower p-values than model 1.

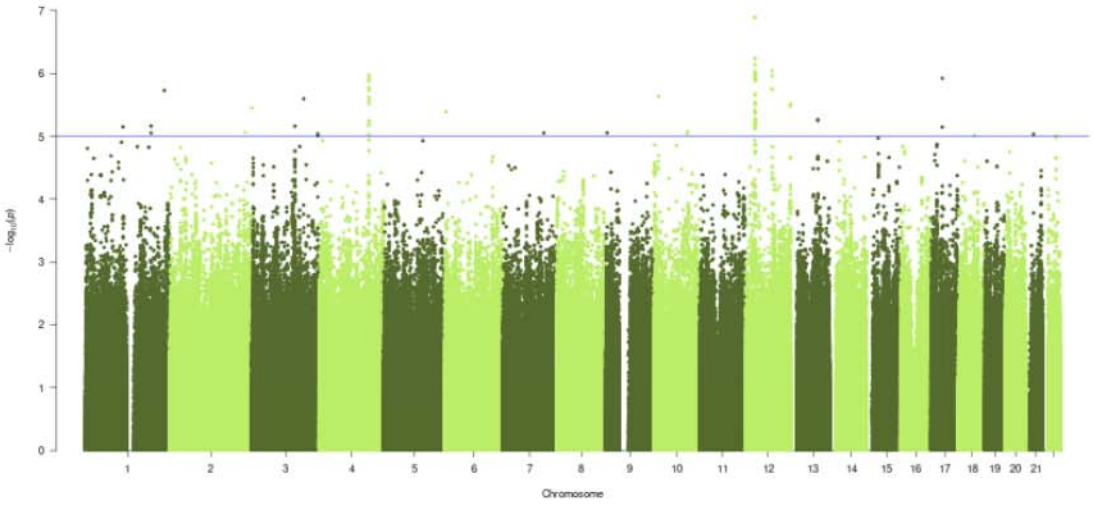

Figure 5.15: Manhattan plot showing the results of a GWAS for Alu Age acceleration computed with the predictor trained using Training set 1 on unfiltered windows.

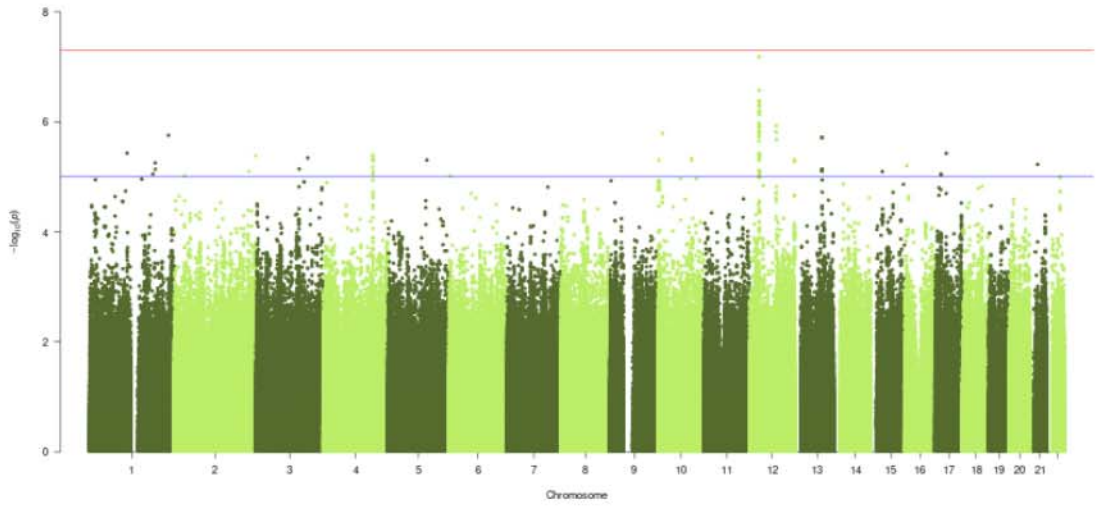

Figure 5.16: Manhattan plot showing the results of a GWAS for Alu Age acceleration computed with the predictor trained using Training set 1 on windows filtered to have an element length >180bp and 99.9% non-zero RPM values at each locus.

Neither model had any associations with a genome wide significant p-value of $p < 5 \times 10^{-8}$, model 1 had 115 and model 2 130 SNPs below the suggestive threshold of $p < 1 \times 10^{-5}$. These sites are listed in appendix 6.14 with the top 10 sites from each GWAS shown in Figures 5.3 & 5.4 respectively. There is substantial overlap between the 2 sets with 143 unique SNPs in the combined list.

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   chr = col_double(),
##   end = col_double(),
##   n_miss = col_double(),
##   alt = col_character(),
##   ref = col_character(),
##   af = col_double(),
##   beta = col_double(),
##   se = col_double(),
##   p_score = col_double(),
##   start = col_double(),
##   rsid = col_character(),
##   vcfstr = col_character()
## )
```

Table 5.3: Top 10 SNPs from model 1.

locus	ref	alt	AF	SNP	p-value
Chr12: 28,247,840	C	T	0.013	rs192929352	1.28e-07
Chr12: 28,249,160	G	A	0.013	rs73080027	1.29e-07
Chr12: 28,275,831	C	T	0.014	rs12320589	5.77e-07
Chr12: 28,275,948	A	G	0.014	rs73081931	5.78e-07
Chr12: 28,437,969	G	A	0.013	rs73083941	7.3e-07
Chr12: 78,820,816	T	C	0.526	rs2060031	9.1e-07
Chr12: 28,329,758	G	C	0.013	rs73081984	9.4e-07
Chr12: 28,334,179	T	G	0.013	rs115527307	9.42e-07
Chr12: 28,319,636	T	C	0.013	rs73081974	9.77e-07
Chr12: 28,311,049	C	T	0.013	rs115986349	9.95e-07

```
##
##
## -- Column specification --------------------------------------------------------
## cols(
##   chr = col_double(),
##   end = col_double(),
##   n_miss = col_double(),
##   alt = col_character(),
##   ref = col_character(),
##   af = col_double(),
##   beta = col_double(),
##   se = col_double(),
##   p_score = col_double(),
##   start = col_double(),
##   rsid = col_character(),
##   vcfstr = col_character()
## )
```

The top 4 sites in both models are the same: rs192929352, rs73080027, rs12320589, & rs73081931 and are in strong linkage disequilibrium with one another. These sites are located in close proximity to one another on chromosome 12 (Figure 5.17).

Table 5.4: Top 10 SNPs from model 2.

locus	ref	alt	AF	SNP	p-value
Chr12: 28,247,840	C	T	0.013	rs192929352	6.55e-08
Chr12: 28,249,160	G	A	0.013	rs73080027	6.57e-08
Chr12: 28,275,831	C	T	0.014	rs12320589	2.66e-07
Chr12: 28,275,948	A	G	0.014	rs73081931	2.66e-07
Chr12: 28,258,339	C	T	0.016	rs35792578	4.18e-07
Chr12: 28,249,612	C	T	0.016	rs11830823	4.25e-07
Chr12: 28,249,623	A	C	0.016	rs11835433	4.25e-07
Chr12: 28,250,701	C	T	0.016	rs7956083	4.25e-07
Chr12: 28,250,549	A	G	0.016	rs7955851	4.25e-07
Chr12: 28,249,196	A	G	0.016	rs11835353	4.26e-07

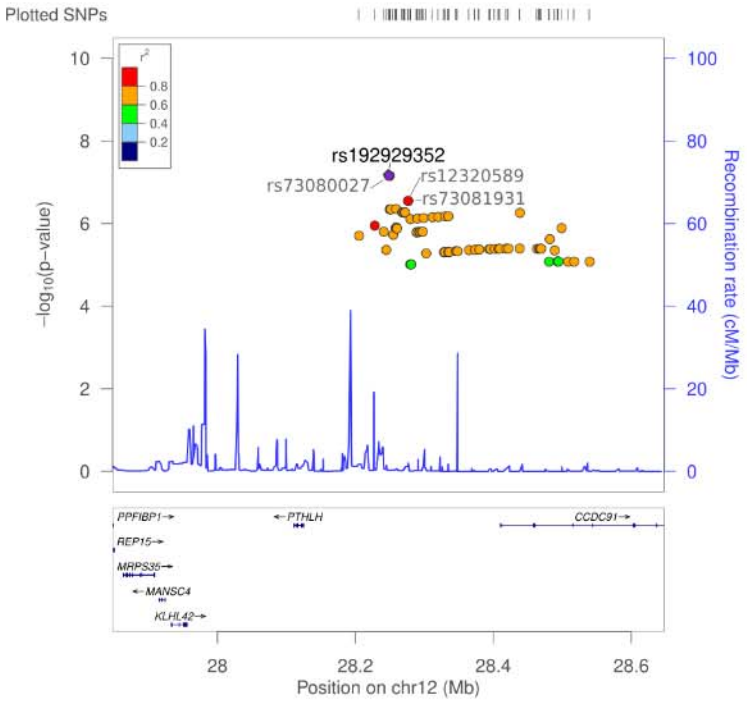

Figure 5.17: Top four shared SNPs are all located just upstream of the the *CCDC91* gene. Adapted from the output of LocusZoom [450].

Further investigation of these top 4 SNPs with `PhenoScanner` [451,452] yielded an association with an increased risk of Treatment with 'oestrogel' topical estradiol (p-value 3.79×10^{-7}, UK biobank, N=337,159) for SNPs rs192929352 & rs73080027. The top phenotypes associated with these SNPs in `PheWeb` [453] also show some overlap (Figure 5.18)

N	Term
4	Other perinatal conditions of fetus or newborn
4	Spermatocele
3	Polyp of corpus uteri
3	Pernicious anemia
2	Acute pharyngitis
2	Sinoatrial node dysfunction (Bradycardia)
2	Pain in limb
2	Sensorineural hearing loss
2	Chronic bronchitis
2	Decreased libido
2	Malignant neoplasm of gallbladder and extrahepatic bile ducts
2	Obstructive chronic bronchitis
2	Congenital anomalies of the integument
2	Cornea replaced by transplant
2	Diabetes or abnormal glucose tolerance complicating pregnancy
1	Polyp of female genital organs
1	Acquired deformities of knee
1	Cramp of limb
1	Abscess or ulceration of vulva

Figure 5.18: *Phenotypes Common to the top 4 shared SNPs.* N = number of the top 4 SNPs in which a term appears in the top 10 Phenotypes Top four that SNP in PheWeb [453].

The risk of having many of these disease phenotypes increases with age. Risk of cancer of the gallbladder for instance increases significantly with age with more than half of new cases in the over 75s and peak incedence at 85-89 years according ot cancer research UK [454]. A common feature of ageing progressive sensorineural hearing loss, or presbycusis [455]. Uterine polyps have a peak prevalence in perimenopausal women but have an increased risk of being maligant with age [456]. It is well known that incidence of cardio-vascular disease increases with age and indeed there are more specific changes in cardiac characteristics with age which could related to sinoatrial node dysfunction. Such as reduced vegal control and altered heart rate variability [457] and, a reduced velocity long-axis systolic shortening in the left ventricle suggestive of impaired ventricular relaxation [458]. These SNP associations have biologically plausible associations with Alu age acceleration, however they are plausibly associated with chronological age and the strong correlation between Alu age acceleration and chronological age means these results cannot be interpreted as due purely to Alu age acceleration.

5.5 Discussion

The Alu Age predictors do not fulfill the Horvath & Raj criteria to be classified as an epigentic clock [293], namely that they should have an r > 0.8.

The age acceleration of the Alu age predictors is strongly correlated with age. This may be due the the relatively poor ability of MeDIP-seq to provide absolute quantitation of DNA methylation. There is substantial technical noise in the value representing DNA methylation at a given locus. The observed systematic bias towards underestimating the ages of the young and overestimating the ages of the old may be explained by a broadly consistent direction with an inconsistent magnitude. Simply scaling the value of given locus by a fixed model coefficient when there is a large amount of variation leads to outsized influence of high magnitude outliers on the age estimates. A DNA methylation age clock trained on the entire genome and not restricted to a particular subset such as the Alu elements should yield the optimal age predictor possible using the MeDIP-seq dataset if there is adequate cross-validation when training the elastic net regression. Comparison to this predictor would reveal if the limitations of the Alu age predictor are specific to it or a more general issue with age predictors based on this dataset.

This strong correlation between Alu age acceleration and chronological age means that age is a substantial confounding factor in the GWAS for Alu age acceleration. Consequently the effect of chronological age cannot be readily disentangled from any signature of biological ageing captured by the Alu repeat based age predictors constructed here.

It was an aim of this work to investigate the genetic influences on the Alu clock based measures of age-acceleration, as has been shown previously with an association in the TERT gene with the Horvath clock [459,460]. Interestingly none of the SNPs identified at $p < 1 \times 10^{-5}$ in either of the two models was present in the list of those identified in a recent GWAS conducted for the age acceleration of several other DNA methylation age clocks [461]. This could suggest that the Alu Age acceleration signature is associating with distinct genetic effects to those of other DNA methylation base age predictors. These results have the potential to reveal mechanisms involved in controlling the epigenetic state of this large portion of the genome. However the strong relationship with Age acceleration and Chronological age in the Alu clock measures of age acceleration renders this difficult as age is a potent confounder of age acceleration in this instance. The results of the genome wide association study whilst they are suggestive of some interesting age related genetic associations do not necessarily represent associations with age acceleration signatures in the Alu repeat regions of the human genome. A GWAS conducted on Chronolgical Age would likely results that differed between the age strata within the cohort which would not be relevant to the question of biological ageing. Thus the associations in the GWAS for Alu age acceleration could be due to population stratification by age in the GWAS study population. GWAS for Longevity have identified very few consistent results with only

APOE and FOXO3 replicating in independent studies [25]. None of 143 unique SNPs from the the 2 GWAS conducted here are shared with the set of top longevity associated SNPs in a meta analysis of longevity GWAS [25].

Direct comparison between DNA methylation array probes and the MeDIP-seq windows used in the Alu clock is not straightforward due to the gaps in coverage of the array, the lack of CpG specific positional resolution and the lack of an absolute unit of methylation level in the MeDIP-seq data. With over a million Alu loci an individually targeted approach to quantifying genome wide Alu element DNA methylation levels would be expensive and impractical. If enzymatic conversion alternatives to bisulfite sequencing methods permit whole genome DNA methylation assays to be performed more cheaply and with longer reads improving mappability [135], then data generated with this approach may permit an Alu DNA methylation clock with greater accuracy in the future.

Chapter 6

Discussion

The epigenomics of ageing has enormous potential for growth in the coming years. DNA methylation based ageing biomarkers have a bright future as a reliable and convenient broad indicators of biological ageing, that have the potential to be used as proxies for longevity in clinical trials intervening in ageing. For instance there is growing interest and investment in drugs with potential anti-ageing activities such as senolytics [462], and in re-purposing of existing drugs like metformin for pro-longevity and increased healthspan interventions [205]. Advances in epigenetic editing [463] promise the possibility of experimentally establishing causal roles of age related epigenetic changes, and the ability to dissect the mechanisms of the involvement of epigenetic changes in the processes of ageing. Improvements are taking plance in DNA methylation assay technologies such as NEBNext Enzymatic Methyl-seq 'EM-seq' [135], better DNA methylation calling in nanopore sequencing [464], advances in single-cell DNA methylation assays [465], and Illumina methylation arrays for sites conserved across mammals [203]. With these tools there are many additional opportunities to characterise the DNA methylomes and capture changes in DNA methylation which were not previously accessible, as well as to begin more mechanistic studies.

6.1 Epigenomic analysis of the developmental origins of long-term bone health

The developmental origins of health and disease hypothesis (DOHaD) holds that early life environmental influences have long term consequences for the risk of developing various pathologies in adulthood and later life. It is with this lens that the EWAS for relationships between umbilical cord DNA methylation and bone health outcomes in Chapter 3 were framed. DNA methylation being an epigenetic mark is influenced by environmental factors and is heritable by subsequent

generations of cells, thus it could be a medium through which environmental factors could act on long term health in accordance with DOHaD. In this chapter two CpGs were identified as having genome wide significant associations with the outcome of interest in their respective EWAS. The first of these was CpG cg26559250 which is located at Chr6:157,653,445-157,653,447 at the *ZDHHC14* (zinc finger DHHC-type palmitoyltransferase 14) gene. This CpG was identified In the EWAS for total bone mineral content minus head at 6 years adjusted for age and sex, with a p-value of 2.52×10^{-8} for an increase of 1.46% per kg in an uncorrected model and a corrected model. The Second was CpG cg22570676 located at Chr19:2,527,492-2,527,494 at the *GNG7* (G protein subunit gamma 7) gene. This CpG was identified on the EWAS for periosteal circumference at 38% from the distal end of the tibia at 6 years (mm) adjusted for age and sex, with a p-value of 4.24×10^{-8} for an increase of 0.370% per mm in an uncorrected model and a corrected model The corrected models included covariates for: blood cell-type composition, maternal age at time of birth (years), sex, maternal BMI at 11 weeks gestation, parity, whether or not the mother smoked during pregnancy, and gestational age.

EWAS were performed for nine different outcomes across three groups of samples, and in each case four different models were fitted. Within a given EWAS the quite stringent Bonferroni standard for multiple testing correction was applied, however conducting multiple EWAS across different groups creates a secondary multiple testing problem. This means that these findings could still be false positives despite the aspiration of family wise correction to minimise type 1 errors. Adjusting for the number of tests performed in a given EWAS should in theory minimise false positives but if several are performed the effective number of tests increases and is not adjusted for increasing the probability that a result could be a false positive from the near zero level family wise correction is aimed at achieving. Confirmation of these associations in another cohort would be necessary in order for the biological reality of these associations to be asserted with confidence. These results are correlational and experimental follow up would be needed to establish any mechanistic or causal relationships between the DNA methylation state at these sites and bone properties in early life. One could for example attempt epigenetic editing of orthologous sites in a model organism and look for an effect on bone measurements [463].

This study did not find significant correlations between the examined bone measurements at *CDKN2A*, where previously an inverse relationship between DNA methylation and bone size, mineral content and mineral density at 4 years had been documented [264]. There was ample opportunity to see changes at this locus in the data, with 95 probe sites in the vicinity of this gene. Nor did this study see significantly reduced DNA methylation at *RXRA* in umbilical cord with maternal vitamin D supplementation or increased circulating vitamin D at the 75 probe sites near *RXRA* [263]. Whilst this study did not replicate these specific findings it has highlighted two new loci with possible relationships to bone health outcomes.

Epigenome wide association studies are here being used as a discovery platform for processes

which may be implicated in the interaction of the *in utero* environment and bone health outcomes as mediated by the epigenome, all of which are complex and multifactorial. There is not strong prior knowledge of the relationships between systems under investigation with which to make precise predictions, the aim is rather to provide a starting point for generating more specific models with which to generate more precise hypotheses. This presents a challenge as there are many sources of noise which could obscure any relationships which do exist between these properties or produce the spurious appearance of a relationship when none may exist. Striking the balance between sensitivity and specificity is particularly challenging in the context. Lowering the threshold for specificity and admitting some type 1 errors might generate sufficient additional hits with which to attempt methods such as gene set enrichment analysis, and related approaches, to identify the biologically relevant systems and processes which may mediate the observed association. However, an excessive number of false positive inputs to such analyses can lead to spuriously identifying associated terms. Simply increasing the sample size of studies to reach the level of power necessary to detect small effect size changes is expensive, impractical and does not help when it comes to analysing existing datasets underpowered for the analysis as initially conceived. The hypothesis free approach has some advantages when attempting to elevate as yet unknown aspects of biology relevant the association being tested to the attention of investigators for further exploration. However, searching for associations between outcomes and particular genomic locations may be of limited effectiveness, even when sufficient statistical power is available to uncover very small effects. As, when individual loci have only very minor contributions to a given effect there are many of them, often spread across many systems [466]. Greater temporal and tissue specificity may reveal larger effect sizes in particular tissues at particular times. Time and tissue specific signals may currently be flattened out in the aggregate signal. The combinatorial complexity of possible times and tissues renders a brute force search impractical, so some prior reason based in biological understanding is likely to be needed to go looking in a particular time and/or tissue for an association.

If the primary interest is in identifying pathways or other functional biological units then making use of dimentionality reduction methods such as weighted correlation network analysis (WGCNA) [467] could potentially help to address some of the power issues faced by these studies. Though this approach also has the limitation that biologically relevant effects may be realised through small perturbations across many systems [466], meaning no whole network may stand out. Grouping the outcomes into effects on correlated gene networks rather than individual genes dramatically reduces the number of statistical tests performed. This approach could be used to narrow the set of tests to perform when looking for gene level associations in other datasets. If changes in gene networks are associated with an outcome of interest in one cohort it is reasonable to take this prior information to a second cohort and test only genes in this network for an association with the outcome of interest dramatically reducing the number of CpG level tests. One could also perform the reciprocal analysis, (dimentionality reduction in the second cohort and CpG

level tests on a reduced set on the first cohort), as a means of validation. An ongoing collaboration with colleagues at MRC-IEU, University of Bristol is including this data in a meta-analysis and is replicating several of these EWAS in other cohorts. This provides an opportunity to attempt to replicate the sites identified here and ascertain if they are sufficiently robust to warrant functional follow-up.

6.2 The implications of ageing-related changes in the epigenetic state of tRNA genes

tRNAs are central to the core cellular process of translation but relatively little is known about the epigenetic state of the genes which encode this essential cellular component. Almost half of all tRNA genes are silent [305], and there is evidence that they are expressed in a tissue specific fashion [325,326], indeed this study found differences in tRNA gene DNA methylation between blood cell types. There are many indicators that there is a great deal of potential biological insight to be gained from a deeper understanding of the epigenetic regulation of tRNA genes.

The finding, in Chapter 4, that the tRNAome is enriched for age related hypermethylation and that there are two loci in which this effect is distinct and replicable is challenging to interpret given the number of different regulatory paradigms in which tRNA genes are involved. There are several ways in which the effects of changes on tRNA DNA methylation could manifest. There is the canonical function of tRNAs in which changes would impact on the expression of the tRNA gene which alters the amount of mature tRNA produced, with potential consequences in translation. There is also action through altering the amount or type of tRNA derived small RNA molecules produced. There is altering the chromatin dynamics of the tRNA gene loci, with potential knock-on effects for other systems given their insulator activity [353] and tendency to cluster in three dimensional space with other tRNA genes [354]. In addition these are not mutually exclusive.

Despite the relatively localised effects on DNA methylation that were observed (Figure 4.13) these changes could still have an effect on larger scale chromatin dynamics. The action of zinc finger CXXC domain proteins binding a unmethylated CpGs and Methyl-CpG-Binding domains at methylated CpGs can alter histone modifications and affect alternate histone usage respectively [468,469]. In addition changes in the nucleosome organisation at tRNA genes may affect genome architecture [470]. So it is hard to rule out effects on chromatin architechture even from localised DNA methylation changes. The circulating tRNA derived small RNAs have been documented as exhibiting age related changes and thus represent an additional signalling vector through which changes in tRNA expression could act [346,471].

The effects of increased DNAm on tRNA genes was to repress their expression in a plasmid

based experimental system [320], however this does not necessarily imply that increased DNA methylation at a tRNA gene locus would result in the reduced expression of tRNA derived small RNAs. Changes in transcriptional dynamics may alter the fate of the transcription product in terms of fragmentation pattern or post transcriptional modifications just as changes in polII transcriptional dynamics impact alternative splicing in protein coding genes [472]. It is plausible that increased DNA methylation at a tRNA gene could cause the tRNA gene to favour production of tRNA derived small RNA transcripts over canonical mature tRNAs rather than silencing it entirely.

tRNAs are a key component of the translation system, a core component of the cell, any age related alterations in this system have the potential to impact on essentially all other systems in the cell if they affect translation. Translation's tight coupling to metabolic regulation and the effect of modulating it on ageing [344] means that any effect of tRNA gene DNAm on translation could be relevant for the regulation of ageing processes.

Unpicking the potential mechanism of action of changes in tRNA gene DNA methylation, if indeed these changes have physiologically relevant impacts, will require greater understanding of a number of aspects of tRNA biology. Our collective understanding of the post-transcriptional modifications of tRNAs and the regulatory functions of tRNA derived small RNAs is still in its infancy [340]. In conclusion with possible effects on chromatin structure, on tRNA derived small RNA signalling and on translation, there are too many mechanisms of action through which changes in epigenetic state of tRNA genes could impact on ageing related systems, and at present insufficient evidence to favour any one particular mechanism. In addition, these possibilities are not mutually exclusive so any effects need not be limited to a single mechanism.

A number of age related tRNA DNA hypermethylation signals persisted after correction for cell-type composition, and there were some differences in tRNA gene DNA methylation between cell-types. This suggests that there are at least some cell-type independent changes in tRNA DNA methylation and leaves open the possibility that there are some changes which are cell-type specific. More detailed characterisation of the cell-type specific activities of tRNA genes may aid in the elucidation of the regulatory functions of tRNAs by permitting the association of the pathways active in particular cell-types. The tRNA-iMet-CAT-1-4 locus which showed the most consistent result has an overall increase of 3.7% from age 4 years to age 78 years in targeted bisulfite sequencing data. The 450k array and Targeted bisulfite sequencing results put the rate of increase in DNA methylation at tRNA-iMet-CAT-1-4 at somewhere in the approximate range of 0.05-0.22 percentage points of DNA methylation increase per year, starting with a baseline methylation near zero at birth. CpGs differentially methylated with age have generally shown changes on the order of 0.1-0.125% per year [120,382], placing the magnitude of the effect seen here in a comparable range to previously observed age related changes in DNA methylation.

The targeted bisulfite sequencing method employed here is quite efficient at capturing DNA

methylation within tRNA genes as a single amplicon is generally sufficient to span a given tRNA gene. This panel could be expanded to most of the human tRNAome though some loci have CpG loci positioned flanking the tRNA genes which prevent the design of primers spanning the tRNA gene locus. It would be interesting to attempt this assay with enzymatic (NEBNext Enzymatic Methyl-seq 'EM-seq') instead of bisulfite conversion as this might permit some inaccessible sites to be targeted with slightly longer amplicons and improve yields due the reduced fragmentation arising from bisulfite treatment [135]. It would be particularly interesting to employ this in mice at tRNA loci not well covered by RRBS, to see if the finding of tRNA gene enrichment for age related hypermethylation is also the case in mice. The RRBS mouse data analysed in this study showed age-related hypermethylation in three of 51 tRNAs. This limited coverage of the 401 high confidence tRNAs in the mouse genome [304] limits the ability to generalise about the enrichment for tRNA gene DNA hypermethylation in mice.

An effective follow up experiment would be to use Pacific Biosciences no amplification CRISPR based targeted sequencing method to examine tRNA gene clusters [473]. A dataset of this type has the potential to elucidate several features of the tRNA genes. Firstly, it would permit a more detailed characterisation of tRNA gene copy number variation, at least those within known clusters, due to the long read sequencing technology. Secondly, the ability to detect DNA methylation, and not simply average levels but patterns of methylation distribution within single molecules, which might reveal additional information about the nature of any age related changes [474]. The long sequencing reads would also permit unambiguous mapping to particular tRNA gene copies which would provide further evidence, in addition to that provided by the targeted bisulfite sequencing analysis, that the observed increase in DNA methylation with age is not somehow an artefact of mapping issues. Coupled with this it would also be illuminating to employ the Hydro-tRNAseq and PAR-CLIP methods employed by Gogakos et al. [352] to perform detailed characterisation of pre-tRNA and mature tRNA transcripts from the same system to permit the relationship between tRNA gene DNA methylation and tRNA transcription levels to be examined simultaneously.

Little was previously known about the epigenetic state of the human tRNA genes [347] and much remains to be characterised. This work represents the first detailed characterisation of the DNA methylation state of the human tRNAome, and revealed a novel pattern of age related hypermethylation in these genes.

6.3 Assessing Biological ageing by DNA methylation changes within Alu repeat elements

Repetitive elements make up some 45% of the human genome [410] and the global hypomethylation observed with age is driven by these repeats [411], but to date limited coverage of these regions [213] has meant that their potential to contain information relevant to biological ageing has gone underexplored.

The best Alu DNA methylation age predictor constructed in Chapter 5 was able to predict chronological age with an R of 0.65 from a training set of n=774 in an unrelated replication set of n=664 with a median absolute error of 8.1 years. Whilst this is less accurate than many of the other DNA methylation based age predictors [434] it was not the primary goal of this predictor to generate the most accurate age predictions but rather to be sufficiently accurate to capture a signature of age acceleration specific to the Alu repeat elements. The difference between the predicted and chronological age, the age acceleration, was strongly correlated with the chronological age such that the age of older individuals is prone to be overestimated and vice versa. One explanation for this may be down to the limitations of DNA methylation quantitation by MeDIP-seq. The elastic net regression may have identified loci which have a consistent direction of change with age but are of variable magnitude. If selecting for sites which had a relatively consistent magnitude in the training set but which varied in a manner which skews higher in the larger population this could lead to the overestimates of the age of older samples and the underestimates of the age of the younger ones. This is speculative and it would be interesting to examine further by looking at the properties of the data in the training and prediction groups of the predictor sites and seeing how they are distributed. In addition it may be possible to study this bias with a simulation approach to see what data properties can produce this pattern of error. This effect in the quantile normalised data was not mitigated by binarising the data by locus, absolute methylation estimates, or in the raw reads per million base pairs data. The difficulty with absolute DNA methylation quantitation by MeDIP-seq is not simple to resolve with common data transformations.

The strong association between Alu age acceleration and chronological age limited the interpretability of the GWAS for age acceleration. This is because associations found here could easily be driven by differences in allele frequencies between different age strata within the GWAS population and not with signal driven by the Alu age acceleration independent of chronological age [475].

The samples used to train and test the models have a median age of approximately 60 years. A larger proportion of samples at a particular part of the age distribution is also a potential source of bias, as poor performance in the lower numbers of older and younger individuals will not be penalised as much as poor performance at the ages with more samples. This also negatively

impacts the ability to determine the quality of the predictor as a good age range is is required to reliably estimate the R of a predictor [293]. Mitigating this by equalising the numbers in certain age groups is a possibility but a substantial number of samples must be left out to achieve this, excessively shrinking the training set. In addition this unlikely to be the sole factor at play in poor prediction performance of these clocks as similarly imperfect age distributions have been used in the training of more performant clocks.

The Twins UK dataset provides a powerful tool for assessing the impact of genetics on age predictors. The age predictors generally performed only marginally better when predicting the ages of the twins of the individuals in the training set than on unrelated individuals, suggesting a minimal impact of genetic factors on the Alu DNA methylation age predictor.

Despite the issues with the correlation of age acceleration with chronological age this work and Wang et al.'s rDNA clock [189] demonstrate that a DNA methylation based age predictors can be trained in repetitive sequence elements. This suggests that constructing DNA methylation based age predictors targeted to particular subsets of the genome is possible. Though, the aim of capturing signatures of biological ageing specific to those subsets is yet to be adequately explored. It may be possible to revisit this approach once large whole genome bisulfite or the enzymatically converted equivalent datasets are generated as it seems unlikely that it would be economical to examine the ~1.1 million Alu elements with any of the available targeted methods. Alternatively if the cost of long-read sequencing drops and the quality of methylated base calls using these methods increases [464] they may also be a viable source of such data. Other possible features on which to attempt to construct age predictors could include: MIR repeat elements some of which have been co-opted as enhancers and which are associated with tissue specific gene expression [476]; Long Terminal Repeat (LTR) elements which are enriched for chromatin marks which characterise active *cis* regulatory elements [477]; Histone genes because of their core function and the wide spread genomic implications of alterations in their availability. They are also located in early-replicating domains [398] so should generally have high fidelity DNA methylation copying during mitosis [123] so any changes observed here are less likely to be the product of epigenetic drift. This would include also alternative histones as they have functions in genome stability and DNA repair [478]. However, Histone genes present a relatively limited set of possible sites with which to predict.

6.4 Conclusion

The unifying theme of this work is the relationship between DNA methylation and healthy ageing. From its possible function as a mediator for the effects of early life environmental influences on long term bone health, through age-related hypermethylation of genes encoding core components of the transcriptional machinery, to signatures of biological ageing in the repetitive regions of

the genome. The epigenome sits upon the genome encoding the annotations to the genome necessary for cells with diverse and dynamic functions to arise from a singular set of genetic information. The ability to construct epigenetic clocks reveals that this layer of information storage and processing contains much that is important for understanding the molecular and cellular processes of ageing. The environmental malleability of the epigenome is its core strength, it is this plasticity to adopt multiple roles that permits multicellularity [479], this malleability both leaves the epigenome open to disruption and presents the possibility of correcting any errant changes. The integrative understanding of epigenomics has the potential to contribute many novel scientific insights into the fundamental mechanisms of ageing in the years to come with profound impacts for our ability to ameliorate chronic and ageing related conditions by intervening in their underlying causes to increase longevity and healthspan.

References

00 1. Acton RJ, Bourne E, Bell J, Lillycrop K, Wang J, Dennison E, Harvey N, Spector TD, Cooper C, Bell CG: **The Genomic Loci of Specific Human tRNA Genes Exhibit Ageing-Related DNA Hypermethylation**. *bioRxiv* 2019, :1–4810.1101/870352Available: https://www.biorxiv.org/content/10.1101/870352v1.

2. Kontis V, Bennett JE, Mathers CD, Li G, Foreman K, Ezzati M: **Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble**. *The Lancet* 2017, **389**:1323–133510.1016/S0140-6736(16)32381-9Available: http://dx.doi.org/10.1016/S0140-6736(16)32381-9.

3. Niccoli T, Partridge L: **Ageing as a risk factor for disease**. *Current Biology* 2012, **22**:R741–R75210.1016/j.cub.2012.07.024Available: http://dx.doi.org/10.1016/j.cub.2012.07.024.

4. Belikov AV: **Age-related diseases as vicious cycles**. *Ageing Research Reviews* 2019, **49**:11–2610.1016/j.arr.2018.11.002Available: https://doi.org/10.1016/j.arr.2018.11.002.

5. Fleming DM, Elliot AJ: **The impact of influenza on the health and health care utilisation of elderly people**. *Vaccine* 2005, **23**:S110.1016/j.vaccine.2005.04.018.

6. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B: **Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study**. *The Lancet* 2012, **380**:37–4310.1016/S0140-6736(12)60240-2Available: http://dx.doi.org/10.1016/S0140-6736(12)60240-2.

7. United Nations, Department of Economic and Social Affairs PDivision: **PopulationPyramid.net**. Available: PopulationPyramid.net.

8. Goldman D: **The Economic Promise of Delayed Aging**. *Cold Spring Harbor Perspectives in Medicine* 2016, **6**:a02507210.1101/cshperspect.a025072Available: http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a025072.

9. Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R: **A C. elegans mutant that lives twice as long as wild type**. *Nature* 1993, **366**:461–46410.1038/366461a0Available: http://www.ncbi.nlm.nih.gov/pubmed/8247153.

10. Medvedev ZA: **An Attempt at a Rational Classification of Theories of Ageing**. *Biological Reviews* 1990, **65**:375–39810.1111/j.1469-185X.1990.tb01428.xAvailable: http://doi.wiley.com/10.1111/j.1469-185X.1990.tb01428.x.

11. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G: **The hallmarks of aging.** *Cell* 2013, **153**:1194–21710.1016/j.cell.2013.05.039Available: http://www.ncbi.nlm.nih.gov/pubmed/23746838.

12. Niedernhofer LJ, Gurkar AU, Wang Y, Vijg J, Hoeijmakers JHJ, Robbins PD: **Nuclear Genomic Instability and Aging**. *Annual Review of Biochemistry* 2018, **87**:295–32210.1146/annurev-biochem-062917-012239Available: https://www.annualreviews.org/doi/10.1146/annurev-biochem-062917-012239.

13. Casagrande S, Hau M: **Telomere attrition: metabolic regulation and signalling function?** *Biology Letters* 2019, **15**:2018088510.1098/rsbl.2018.0885Available: https://royalsocietypublishing.org/doi/10.1098/rsbl.2018.0885.

14. Pal S, Tyler JK: **Epigenetics and aging**. *Science Advances* 2016, **2**:e160058410.1126/sciadv.1600584Available: https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.1600584.

15. Santra M, Dill KA, Graff AMR de: **Proteostasis collapse is a driver of cell aging and death**. *Proceedings of the National Academy of Sciences* 2019, **116**:22173–2217810.1073/pnas.1906592116Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1906592116.

16. Johnson SC: **Nutrient Sensing, Signaling and Ageing: The Role of IGF-1 and mTOR in Ageing and Age-Related Disease**. In *Subcellular biochemistry* Vol. 90 2018:49–97. Available: http://link.springer.com/10.1007/978-981-13-2835-0%7B/_%7D3.

17. Soultoukis GA, Partridge L: **Dietary Protein, Metabolism, and Aging**. *Annual Review of Biochemistry* 2016, **85**:5–3410.1146/annurev-biochem-060815-014422Available: http://www.annualreviews.org/doi/10.1146/annurev-biochem-060815-014422.

18. Payne BAI, Chinnery PF: **Mitochondrial dysfunction in aging: Much progress but many unresolved questions**. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 2015, **1847**:1347–135310.1016/j.bbabio.2015.05.022Available: http://dx.doi.org/10.1016/j.bbabio.2015.05.022.

19. Avelar RA, Ortega JG, Tacutu R, Tyler EJ, Bennett D, Binetti P, Budovsky A, Chatsirisupachai K, Johnson E, Murray A, Shields S, Tejada-Martinez D, Thornton D, Fraifeld VE, Bishop CL, Magalhães JP de: **A multidimensional systems biology analysis of cellular senescence in aging and disease**. *Genome Biology* 2020, **21**:9110.1186/s13059-020-01990-9Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-01990-9.

20. Ren R, Ocampo A, Liu G-H, Izpisua Belmonte JC: **Regulation of Stem Cell Aging by**

Metabolism and Epigenetics. *Cell Metabolism* 2017, **26**:460–47410.1016/j.cmet.2017.07.019Available: http://dx.doi.org/10.1016/j.cmet.2017.07.019%20https://linkinghub.elsevier.com/retrieve/pii/S1550413117304849.

21. Franceschi C, Campisi J: **Chronic Inflammation (Inflammaging) and Its Potential Contribution to Age-Associated Diseases**. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 2014, **69**:S4–S910.1093/gerona/glu057Available: https://academic.oup.com/biomedgerontology/article-lookup/doi/10.1093/gerona/glu057.

22. Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R: **A C. elegans mutant that lives twice as long as wild type**. *Nature* 1993, **366**:461–46410.1038/366461a0Available: http://www.nature.com/articles/366461a0.

23. Kenyon CJ: **The genetics of ageing**. *Nature* 2010, **464**:504–51210.1038/nature08980Available: http://www.ncbi.nlm.nih.gov/pubmed/20336132.

24. Singh PP, Demmitt BA, Nath RD, Brunet A: **The Genetics of Aging: A Vertebrate Perspective**. *Cell* 2019, **177**:200–22010.1016/j.cell.2019.02.038Available: https://doi.org/10.1016/j.cell.2019.02.038.

25. Broer L, Buchman AS, Deelen J, Evans DS, Faul JD, Lunetta KL, Sebastiani P, Smith JA, Smith AV, Tanaka T, Yu L, Arnold AM, Aspelund T, Benjamin EJ, De Jager PL, Eirkisdottir G, Evans DA, Garcia ME, Hofman A, Kaplan RC, Kardia SLR, Kiel DP, Oostra BA, Orwoll ES, Parimi N, Psaty BM, Rivadeneira F, Rotter JI, Seshadri S, Singleton A, et al.: **GWAS of Longevity in CHARGE Consortium Confirms APOE and FOXO3 Candidacy**. *The Journals of Gerontology: Series A* 2015, **70**:110–11810.1093/gerona/glu166Available: https://academic.oup.com/biomedgerontology/article/70/1/110/2947666.

26. Zenin A, Tsepilov Y, Sharapov S, Getmantsev E, Menshikov LI, Fedichev PO, Aulchenko Y: **Identification of 12 genetic loci associated with human healthspan**. *Communications Biology* 2019, **2**:410.1038/s42003-019-0290-0Available: http://dx.doi.org/10.1038/s42003-019-0290-0%20http://www.nature.com/articles/s42003-019-0290-0.

27. Melzer D, Pilling LC, Ferrucci L: **The genetics of human ageing**. *Nature Reviews Genetics* 2019, 10.1038/s41576-019-0183-6Available: http://dx.doi.org/10.1038/s41576-019-0183-6.

28. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, Kathiresan S, Kenny EE, Lindgren CM, MacArthur DG, North KN, Plon SE, Rehm HL, Risch N, Rotimi CN, Shendure J, Soranzo N, McCarthy MI: **A brief history of human disease genetics**. *Nature* 2020, **577**:179–18910.1038/s41586-019-1879-7.

29. Shindyapina AV, Zenin AA, Tarkhov AE, Santesmasses D, Fedichev PO, Gladyshev VN: **Germline burden of rare damaging variants negatively affects human healthspan and lifespan**. *eLife* 2020, **9**:1–1810.7554/eLife.53449Available: https://elifesciences.org/articl

es/53449.

30. Booth LN, Brunet A: **The Aging Epigenome**. *Molecular Cell* 2016, **62**:728–744 10.1016/j.molcel.2016.05.013 Available: http://dx.doi.org/10.1016/j.molcel.2016.05.013.

31. Sinclair DA, Oberdoerffer P: **The ageing epigenome: Damaged beyond repair?** *Ageing Research Reviews* 2009, **8**:189–198 10.1016/j.arr.2009.04.004.

32. Hayano M, Yang J-H, Bonkowski MS, Amorim JA, Ross JM, Coppotelli G, Griffin P, Chew YC, Guo W, Yang X, Vera DL, Salfati EL, Das A, Thakur S, Kane AE, Mitchell SJ, Mohri Y, Nishimura EK, Schaevitz L, Garg N, Balta A-M, Rego MA, Gregory-Ksander M, Jakobs TC, Zhong L, Wakimoto H, Mostoslavsky R, Wagers AJ, Tsubota K, Bonasera SJ, et al.: **DNA Break-Induced Epigenetic Drift as a Cause of Mammalian Aging**. *bioRxiv* 2019, 10.1101/808659 Available: https://doi.org/10.1101/808659.

33. Kane AE, Sinclair DA: **Epigenetic changes during aging and their reprogramming potential**. *Critical Reviews in Biochemistry and Molecular Biology* 2019, **54**:61–83 10.1080/10409238.2019.1570075 Available: https://doi.org/10.1080/10409238.2019.1570075.

34. Bird A: **Perceptions of epigenetics.** *Nature* 2007, **447**:396–8 10.1038/nature05913 Available: http://www.ncbi.nlm.nih.gov/pubmed/17522671.

35. Waddington CH: **The epigenotype. 1942.** *International journal of epidemiology* 2012, **41**:10–3 10.1093/ije/dyr184 Available: http://www.ncbi.nlm.nih.gov/pubmed/22186258.

36. Russo V, Martienssen R, Riggs AD: *Epigenetic Mechanisms of Gene Regulation*. Cold Spring Harbor Laboratory Press; 1996 Available: https://cshmonographs.org/index.php/monographs/issue/view/087969490.32.

37. Tabansky I, Stern JNH, Pfaff DW: **Implications of Epigenetic Variability within a Cell Population for "Cell Type" Classification**. *Frontiers in Behavioral Neuroscience* 2015, **9**:1–11 10.3389/fnbeh.2015.00342 Available: http://journal.frontiersin.org/Article/10.3389/fnbeh.2015.00342/abstract.

38. Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, Vitale L, Pelleri MC, Tassani S, Piva F, Perez-Amodio S, Strippoli P, Canaider S: **An estimation of the number of cells in the human body.** *Annals of human biology* 2013, **40**:463–71 10.3109/03014460.2013.807878 Available: http://www.ncbi.nlm.nih.gov/pubmed/23829164.

39. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Göttgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klenerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundeberg J, Majumder P, et al.: **The Human Cell Atlas**. *eLife* 2017, **6**:1–30 10.7554/eLife.27041 Available: https://elifesciences.org/articles/27041.

40. CellSystemsVoices: **What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism?** *Cell Systems* 2017, **4**:255–259 10.1016/j.cels.2017.03.006 Available: https://linkinghub.elsevier.com/retrieve/pii/S2405471217300911.

41. CellOntology: **Cell Ontology**. 2017, 10.5281/zenodo.168254 Available: http://obofoundry.org/ontology/cl.html. Accessed 5 April 2017.

42. Guo JU, Su Y, Shin JH, Shin J, Li H, Xie B, Zhong C, Hu S, Le T, Fan G, Zhu H, Chang Q, Gao Y, Ming G, Song H: **Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain.** *Nature neuroscience* 2014, **17**:215–22 10.1038/nn.3607 Available: http://www.ncbi.nlm.nih.gov/pubmed/24362762.

43. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462**:315–22 10.1038/nature08514 Available: http://www.ncbi.nlm.nih.gov/pubmed/19829295.

44. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, Lin S, Lin Y, Jung I, Schmitt AD, Selvaraj S, Ren B, Sejnowski TJ, Wang W, Ecker JR: **Human body epigenome maps reveal noncanonical DNA methylation variation.** *Nature* 2015, **523**:212–6 10.1038/nature14465 Available: http://www.ncbi.nlm.nih.gov/pubmed/26030523.

45. Penn NW, Suwalski R, O'Riley C, Bojanowski K, Yura R: **The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid.** *Biochemical Journal* 1972, **126**:781–790 10.1042/bj1260781 Available: https://portlandpress.com/biochemj/article/126/4/781/6593/The-presence-of-5hydroxymethylcytosine-in-animal.

46. Kriaucionis S, Heintz N: **The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain.** *Science* 2009, **324**:929–930 10.1126/science.1169786 Available: https://www.sciencemag.org/lookup/doi/10.1126/science.1169786.

47. Pfeifer GP, Kadam S, Jin S-G: **5-hydroxymethylcytosine and its potential roles in development and cancer.** *Epigenetics & chromatin* 2013, **6**:10 10.1186/1756-8935-6-10 Available: http://www.ncbi.nlm.nih.gov/pubmed/23634848.

48. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y: **Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine.** *Science* 2011, **333**:1300–1303 10.1126/science.1210597 Available: https://www.sciencemag.org/lookup/doi/10.1126/science.1210597.

49. Wu TP, Wang T, Seetin MG, Lai Y, Zhu S, Lin K, Liu Y, Byrum SD, Mackintosh SG, Zhong M, Tackett A, Wang G, Hon LS, Fang G, Swenberg Ja, Xiao AZ: **DNA methyla-**

tion on N(6)-adenine in mammalian embryonic stem cells. *Nature* 2016, **532**:329–3310.1038/nature17640Available: http://www.ncbi.nlm.nih.gov/pubmed/27027282.

50. Xie Q, Wu TP, Gimple RC, Li Z, Prager BC, Wu Q, Yu Y, Wang P, Wang Y, Gorkin DU, Zhang C, Dowiak AV, Lin K, Zeng C, Sui Y, Kim LJY, Miller TE, Jiang L, Lee CH, Huang Z, Fang X, Zhai K, Mack SC, Sander M, Bao S, Kerstetter-Fogle AE, Sloan AE, Xiao AZ, Rich JN: **N-methyladenine DNA Modification in Glioblastoma**. *Cell* 2018, **175**:1228–1243.e2010.1016/j.cell.2018.10.006Available: https://linkinghub.elsevier.com/retrieve/pii/S009286741831314X.

51. Losick JDWTABSPBAGMLR: *Molecular Biology of the Gene*. 7th ed. Pearson; 2014.

52. Bannister AJ, Kouzarides T: **Regulation of chromatin by histone modifications**. *Cell Research* 2011, **21**:381–39510.1038/cr.2011.22Available: http://www.ncbi.nlm.nih.gov/pubmed/21321607.

53. Strahl BD, Allis CD: **The language of covalent histone modifications**. *Nature* 2000, **403**:41–4510.1038/47412Available: http://www.nature.com/articles/47412.

54. Schreiber SL, Bernstein BE: **Signaling Network Model of Chromatin**. *Cell* 2002, **111**:771–77810.1016/S0092-8674(02)01196-0Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867402011960.

55. Voigt P, LeRoy G, Drury WJ, Zee BM, Son J, Beck DB, Young NL, Garcia BA, Reinberg D: **Asymmetrically modified nucleosomes.** *Cell* 2012, **151**:181–9310.1016/j.cell.2012.09.002Available: http://www.ncbi.nlm.nih.gov/pubmed/23021224.

56. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, Halow J, Van Nostrand EL, Freese P, Gorkin DU, Shen Y, He Y, Mackiewicz M, Pauli-Behn F, Williams BA, Mortazavi A, Keller CA, Zhang X-O, Elhajjajy SI, Huey J, Dickel DE, Snetkova V, Wei X, Wang X, Rivera-Mulia JC, Rozowsky J, et al.: **Expanded encyclopaedias of DNA elements in the human and mouse genomes**. *Nature* 2020, **583**:699–71010.1038/s41586-020-2493-4Available: http://www.nature.com/articles/s41586-020-2493-4.

57. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes Ja, Noble WS: **Unsupervised pattern discovery in human chromatin structure through genomic segmentation.** *Nature methods* 2012, **9**:473–610.1038/nmeth.1937Available: http://www.ncbi.nlm.nih.gov/pubmed/22426492.

58. Carrillo-de-Santa-Pau E, Juan D, Pancaldi V, Were F, Martin-Subero I, Rico D, Valencia A: **Automatic identification of informative regions with epigenomic changes associated to hematopoiesis**. *Nucleic Acids Research* 2017, **45**:9244–925910.1093/nar/gkx618Available: https://academic.oup.com/nar/article/45/16/9244/3976483.

59. Talbert PB, Ahmad K, Almouzni G, Ausió J, Berger F, Bhalla PL, Bonner WM, Cande W, Chadwick BP, Chan SWL, Cross GAM, Cui L, Dimitrov SI, Doenecke D, Eirin-López JM, Gorovsky MA, Hake SB, Hamkalo BA, Holec S, Jacobsen SE, Kamieniarz K, Khochbin S, Ladurner AG, Landsman D, Latham JA, Loppin B, Malik HS, Marzluff WF, Pehrson JR, Postberg J, et al.: **A unified phylogeny-based nomenclature for histone variants**. *Epigenetics & Chromatin* 2012, **5**:710.1186/1756-8935-5-7Available: http://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/1756-8935-5-7.

60. Weber CM, Henikoff S: **Histone variants: dynamic punctuation in transcription.** *Genes & development* 2014, **28**:672–8210.1101/gad.238873.114Available: http://www.ncbi.nlm.nih.gov/pubmed/24696452.

61. Kaikkonen MU, Lam MTY, Glass CK: **Non-coding RNAs as regulators of gene expression and epigenetics.** *Cardiovascular research* 2011, **90**:430–4010.1093/cvr/cvr097Available: http://www.ncbi.nlm.nih.gov/pubmed/21558279.

62. Cerase A, Pintacuda G, Tattermusch A, Avner P: **Xist localization and function: new insights from multiple levels**. *Genome Biology* 2015, **16**:16610.1186/s13059-015-0733-yAvailable: http://genomebiology.com/2015/16/1/166.

63. Jenuwein T: **Translating the Histone Code**. *Science* 2001, **293**:1074–108010.1126/science.1063127Availabl http://www.ncbi.nlm.nih.gov/pubmed/20653993.

64. Smith RWA, Monroe C, Bolnick DA: **Detection of Cytosine methylation in ancient DNA from five native american populations using bisulfite sequencing.** *PloS one* 2015, **10**:e012534410.1371/journal.pone.0125344Available: http://www.ncbi.nlm.nih.gov/pubmed/26016479.

65. Pedersen JS, Valen E, Velazquez AMV, Parker BJ, Rasmussen M, Lindgreen S, Lilje B, Tobin DJ, Kelly TK, Vang S, Andersson R, Jones PA, Hoover CA, Tikhonov A, Prokhortchouk E, Rubin EM, Sandelin A, Gilbert MTP, Krogh A, Willerslev E, Orlando L: **Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome**. *Genome Research* 2014, **24**:454–46610.1101/gr.163592.113Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.163592.113.

66. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S: **Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA.** *Nucleic acids research* 2010, **38**:e8710.1093/nar/gkp1163Available: http://www.ncbi.nlm.nih.gov/pubmed/20028723.

67. Bauden M, Kristl T, Andersson R, Marko-Varga G, Ansari D: **Characterization of histone-related chemical modifications in formalin-fixed paraffin-embedded and fresh-frozen human pancreatic cancer xenografts using LC-MS/MS**. *Laboratory Investigation* 2017, **97**:279–28810.1038/labinvest.2016.134Available: http://www.nature.com/d

oifinder/10.1038/labinvest.2016.134.

68. Hashimshony T, Zhang J, Keshet I, Bustin M, Cedar H: **The role of DNA methylation in setting up chromatin structure during development**. *Nature Genetics* 2003, **34**:187–19210.1038/ng1158Available: http://www.nature.com/articles/ng1158.

69. Estève P-O, Chin HG, Smallwood A, Feehery GR, Gangisetty O, Karpf AR, Carey MF, Pradhan S: **Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication.** *Genes & development* 2006, **20**:3089–10310.1101/gad.1463706Available: http://www.ncbi.nlm.nih.gov/pubmed/17085482.

70. Rose NR, Klose RJ: **Understanding the relationship between DNA methylation and histone lysine methylation**. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 2014, **1839**:1362–137210.1016/j.bbagrm.2014.02.007Available: http://dx.doi.org/10.1016/j.bbagrm.2014.02.007.

71. Luo Y, Lu X, Xie H: **Dynamic Alu Methylation during Normal Development, Aging, and Tumorigenesis**. *BioMed Research International* 2014, **2014**:1–1210.1155/2014/784706Available: http://www.hindawi.com/journals/bmri/2014/784706/.

72. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–92110.1038/35057062Available: http://www.ncbi.nlm.nih.gov/pubmed/11237011.

73. Duncan BK, Miller JH: **Mutagenic deamination of cytosine residues in DNA.** *Nature* 1980, **287**:560–1Available: http://www.ncbi.nlm.nih.gov/pubmed/6999365.

74. Zhou Y, He F, Pu W, Gu X, Wang J, Su Z: **The Impact of DNA Methylation Dynamics on the Mutation Rate During Human Germline Development**. *G3: Genes|Genomes|Genetics* 2020, **10**:3337–334610.1534/g3.120.401511Available: http://g3journal.org/lookup/doi/10.1534/g3.120.401511.

75. Guo Y, Jamison DC: **The distribution of SNPs in human gene regulatory regions**. *BMC Genomics* 2005, **6**:14010.1186/1471-2164-6-140Available: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-6-140.

76. Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP: **Redefining CpG islands using hidden Markov models**. *Biostatistics* 2010, **11**:499–51410.1093/biostatistics/kxq005Available: https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxq005.

77. Gardiner-Garden M, Frommer M: **CpG Islands in vertebrate genomes**. *Journal of Molecular Biology* 1987, **196**:261–28210.1016/0022-2836(87)90689-9Available: https://linkingh

ub.elsevier.com/retrieve/pii/0022283687906899.

78. 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**:1412–141710.1073/pnas.0510310103Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0510310103.

79. Bell JSK, Vertino PM: **Orphan CpG islands define a novel class of highly active enhancers.** *Epigenetics* 2017, **12**:449–46410.1080/15592294.2017.1297910Available: http://www.ncbi.nlm.nih.gov/pubmed/28448736.

80. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP: **Orphan CpG islands identify numerous conserved promoters in the mammalian genome.** *PLoS genetics* 2010, **6**:e100113410.1371/journal.pgen.1001134Available: http://www.ncbi.nlm.nih.gov/pubmed/20885785.

81. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJM, Haussler D, Marra MA, Hirst M, Wang T, Costello JF: **Conserved role of intragenic DNA methylation in regulating alternative promoters.** *Nature* 2010, **466**:253–710.1038/nature09165Available: http://www.ncbi.nlm.nih.gov/pubmed/20613842.

82. Ehrlich M, Gama-Sosa MA, Huang L-H, Midgett RM, Kuo KC, McCune RA, Gehrke C: **Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells.** *Nucleic Acids Research* 1982, **10**:2709–272110.1093/nar/10.8.2709Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/10.8.2709.

83. Bird A: **DNA methylation patterns and epigenetic memory.** *Genes & development* 2002, **16**:6–2110.1101/gad.947102Available: http://www.ncbi.nlm.nih.gov/pubmed/11782440.

84. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, Nimwegen E van, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schübeler D: **DNA-binding factors shape the mouse methylome at distal regulatory regions.** *Nature* 2011, **480**:490–510.1038/nature10716Available: http://www.ncbi.nlm.nih.gov/pubmed/22170606.

85. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash J, Sabunciyan S, Feinberg AP: **The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.** *Nature genetics* 2009, **41**:178–18610.1038/ng.298Available: http://www.ncbi.nlm.nih.gov/pubmed/19151715.

86. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, De Jager PL, Rosen ED,

Bennett DA, Bernstein BE, Gnirke A, Meissner A: **Charting a dynamic DNA methylation landscape of the human genome.** *Nature* 2013, **500**:477–8110.1038/nature12433Available: http://www.ncbi.nlm.nih.gov/pubmed/23925113.

87. Baubec T, Schübeler D: **Genomic patterns and context specific interpretation of DNA methylation.** *Current Opinion in Genetics & Development* 2014, **25**:85–9210.1016/j.gde.2013.11.015Available: http://www.ncbi.nlm.nih.gov/pubmed/24614011.

88. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, Nitta KR, Taipale M, Popov A, Ginno PA, Domcke S, Yan J, Schübeler D, Vinson C, Taipale J: **Impact of cytosine methylation on DNA binding specificities of human transcription factors.** *Science (New York, N.Y.)* 2017, **356**:eaaj223910.1126/science.aaj2239Available: http://www.ncbi.nlm.nih.gov/pubmed/28473536.

89. Zuo Z, Roy B, Chang YK, Granas D, Stormo GD: **Measuring quantitative effects of methylation on transcription factor–DNA binding affinity.** *Science Advances* 2017, **3**:eaao179910.1126/sciadv.aao1799Available: http://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aao1799.

90. Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones Pa: **Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules.** *Genome Research* 2012, **22**:2497–25010.1101/gr.143008.112Available: http://www.ncbi.nlm.nih.gov/pubmed/22960375.

91. Huff JT, Zilberman D: **Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes.** *Cell* 2014, **156**:1286–12910.1016/j.cell.2014.01.029Available: http://www.ncbi.nlm.nih.gov/pubmed/24630728.

92. Raiber E, Portella G, Martínez Cuesta S, Hardisty R, Murat P, Li Z, Iurlaro M, Dean W, Spindel J, Beraldi D, Liu Z, Dawson MA, Reik W, Balasubramanian S: **5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells.** *Nature Chemistry* 2018, **10**:1258–12610.1038/s41557-018-0149-xAvailable: http://www.nature.com/articles/s41557-018-0149-x.

93. Lyko F: **The DNA methyltransferase family: a versatile toolkit for epigenetic regulation.** *Nature Reviews Genetics* 2017, **19**:81–9210.1038/nrg.2017.80Available: http://www.nature.com/doifinder/10.1038/nrg.2017.80.

94. Vertino PM, Sekowski JA, Coll JM, Applegren N, Han S, Hickey RJ, Malkas LH: **DNMT1 is a component of a multiprotein DNA replication complex.** *Cell cycle (Georgetown, Tex.)* 2002, **1**:416–2310.4161/cc.1.6.270Available: http://www.ncbi.nlm.nih.gov/pubmed/12548018.

95. Bostick M, Kim JK, Esteve P-O, Clark A, Pradhan S, Jacobsen SE: **UHRF1 Plays a**

Role in Maintaining DNA Methylation in Mammalian Cells. *Science* 2007, **317**:1760–176410.1126/science.1147939Available: https://www.sciencemag.org/lookup/doi/10.1126/science.1147939.

96. 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**:248–25110.1038/nature06146Available: http://www.nature.com/articles/nature06146.

97. Kaiser S, Jurkowski TP, Kellner S, Schneider D, Jeltsch A, Helm M: **The RNA methyltransferase Dnmt2 methylates DNA in the structural context of a tRNA**. *RNA Biology* 2017, **14**:1241–125110.1080/15476286.2016.1236170Available: https://www.tandfonline.com/doi/full/10.1080/15476286.2016.1236170.

98. Tahiliani M, Koh KP, Shen Y, Pastor Wa, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A: **Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1**. *Science* 2009, **324**:930–93510.1126/science.1170116Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1170116.

99. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L: **Mutational landscape and significance across 12 major cancer types**. *Nature* 2013, **502**:333–33910.1038/nature12634.

100. Ko M, Huang Y, Jankowska AM, Pape UJ, Tahiliani M, Bandukwala HS, An J, Lamperti ED, Koh KP, Ganetzky R, Liu XS, Aravind L, Agarwal S, Maciejewski JP, Rao A: **Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2**. *Nature* 2010, **468**:839–84310.1038/nature09586Available: http://www.nature.com/articles/nature09586.

101. Wu X, Zhang Y: **TET-mediated active DNA demethylation: mechanism, function and beyond**. *Nature Reviews Genetics* 2017, **18**:517–53410.1038/nrg.2017.33Available: http://www.nature.com/doifinder/10.1038/nrg.2017.33.

102. Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schübeler D: **Identification of genetic elements that autonomously determine DNA methylation states.** *Nature genetics* 2011, **43**:1091–710.1038/ng.946Available: http://www.ncbi.nlm.nih.gov/pubmed/21964573.

103. Boks MP, Derks EM, Weisenberger DJ, Strengman E, Janson E, Sommer IE, Kahn RS, Ophoff RA: **The relationship of DNA methylation with age, gender and genotype in twins and healthy controls.** *PloS one* 2009, **4**:e676710.1371/journal.pone.0006767Available: http://www.ncbi.nlm.nih.gov/pubmed/19774229.

104. Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, Hod E, Li K, Murty VV, Schupf N, Vilain E, Morris M, Haghighi F, Tycko B: **Genomic surveys by methylation-sensitive SNP analysis**

identify sequence-dependent allele-specific DNA methylation. *Nature genetics* 2008, **40**:904–810.1038/ng.174Available: http://www.ncbi.nlm.nih.gov/pubmed/18568024.

105. Schalkwyk LC, Meaburn EL, Smith R, Dempster EL, Jeffries AR, Davies MN, Plomin R, Mill J: **Allelic skewing of DNA methylation is widespread across the genome.** *American journal of human genetics* 2010, **86**:196–21210.1016/j.ajhg.2010.01.014Available: http://www.ncbi.nlm.nih.gov/pubmed/20159110.

106. Shoemaker R, Deng J, Wang W, Zhang K: **Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome.** *Genome research* 2010, **20**:883–910.1101/gr.104695.109Available: http://www.ncbi.nlm.nih.gov/pubmed/20418490.

107. Bell CG, Gao F, Yuan W, Roos L, Acton RJ, Xia Y, Bell J, Ward K, Mangino M, Hysi PG, Wang J, Spector TD: **Obligatory and facilitative allelic variation in the DNA methylome within common disease-associated loci.** *Nature Communications* 2018, **9**:810.1038/s41467-017-01586-1Available: http://www.nature.com/articles/s41467-017-01586-1.

108. Martin-Trujillo A, Vidal E, Monteagudo-Sánchez A, Sanchez-Delgado M, Moran S, Hernandez Mora JR, Heyn H, Guitart M, Esteller M, Monk D: **Copy number rather than epigenetic alterations are the major dictator of imprinted methylation in tumors.** *Nature Communications* 2017, **8**:46710.1038/s41467-017-00639-9Available: http://www.ncbi.nlm.nih.gov/pubmed/28883545.

109. Ciernia AV, LaSalle J: **The landscape of DNA methylation amid a perfect storm of autism aetiologies.** *Nature Reviews Neuroscience* 2016, **17**:411–42310.1038/nrn.2016.41Available: http://dx.doi.org/10.1038/nrn.2016.41.

110. Ushijima T, Watanabe N, Shimizu K, Miyamoto K, Sugimura T, Kaneda A: **Decreased fidelity in replicating CpG methylation patterns in cancer cells.** *Cancer research* 2005, **65**:11–7Available: http://www.ncbi.nlm.nih.gov/pubmed/15665274.

111. Kunkel TA: **DNA Replication Fidelity.** *Journal of Biological Chemistry* 2004, **279**:16895–1689810.1074/jbc.R400006200Available: http://www.jbc.org/lookup/doi/10.1074/jbc.R400006200.

112. Sharif J, Muto M, Takebayashi S, Suetake I, Iwamatsu A, Endo TA, Shinga J, Mizutani-Koseki Y, Toyoda T, Okamura K, Tajima S, Mitsuya K, Okano M, Koseki H: **The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA.** *Nature* 2007, **450**:908–91210.1038/nature06397Available: http://www.nature.com/articles/nature06397.

113. Arita K, Ariyoshi M, Tochio H, Nakamura Y, Shirakawa M: **Recognition of hemimethylated DNA by the SRA protein UHRF1 by a base-flipping mechanism.** *Nature* 2008, **455**:818–82110.1038/nature07249Available: http://www.nature.com/articles/nature0724

9.

114. Liu X, Gao Q, Li P, Zhao Q, Zhang J, Li J, Koseki H, Wong J: **UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9**. *Nature Communications* 2013, **4**:156310.1038/ncomms2562Available: http://www.ncbi.nlm.nih.gov/pubmed/23463006.

115. Laird CD, Pleasant ND, Clark AD, Sneeden JL, Hassan KMA, Manley NC, Vary JC, Morgan T, Hansen RS, Stoger R: **Hairpin-bisulfite PCR: Assessing epigenetic methylation patterns on complementary strands of individual DNA molecules**. *Proceedings of the National Academy of Sciences* 2004, **101**:204–20910.1073/pnas.2536758100Available: http://www.pnas.org/cgi/doi/10.1073/pnas.2536758100.

116. Pfeifer GP, Steigerwald SD, Hansen RS, Gartler SM, Riggs AD: **Polymerase chain reaction-aided genomic sequencing of an X chromosome-linked CpG island: methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability.** *Proceedings of the National Academy of Sciences* 1990, **87**:8252–825610.1073/pnas.87.21.8252Available: http://www.pnas.org/cgi/doi/10.1073/pnas.87.21.8252.

117. Riggs AD, Xiong Z: **Methylation and epigenetic fidelity**. *Proceedings of the National Academy of Sciences* 2004, **101**:4–510.1073/pnas.0307781100Available: http://www.ncbi.nlm.nih.gov/pubmed/14695893.

118. Jenkinson G, Pujadas E, Goutsias J, Feinberg AP: **Potential energy landscapes identify the information-theoretic nature of the epigenome.** *Nature genetics* 2017, **49**:719–72910.1038/ng.3811Available: http://www.ncbi.nlm.nih.gov/pubmed/28346445.

119. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suner D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu Y-Z, Plass C, Esteller M: **From The Cover: Epigenetic differences arise during the lifetime of monozygotic twins**. *Proceedings of the National Academy of Sciences* 2005, **102**:10604–1060910.1073/pnas.0500398102Available: http://www.ncbi.nlm.nih.gov/pubmed/16009939.

120. Slieker RC, Iterson M van, Luijk R, Beekman M, Zhernakova DV, Moed MH, Mei H, Galen M van, Deelen P, Bonder MJ, Zhernakova A, Uitterlinden AG, Tigchelaar EF, Stehouwer CDA, Schalkwijk CG, Kallen CJH van der, Hofman A, Heemst D van, Geus EJ de, Dongen J van, Deelen J, Berg LH van den, Meurs J van, Jansen R, 't Hoen PAC, Franke L, Wijmenga C, Veldink JH, Swertz MA, Greevenbroek MMJ van, et al.: **Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms.** *Genome biology* 2016, **17**:19110.1186/s13059-016-1053-6Available: http://www.ncbi.nlm.nih.gov/pubmed/27654999.

121. Hahn O, Grönke S, Stubbs TM, Ficz G, Hendrich O, Krueger F, Andrews S,

Zhang Q, Wakelam MJ, Beyer A, Reik W, Partridge L: **Dietary restriction protects from age-associated DNA methylation and induces epigenetic reprogramming of lipid metabolism**. *Genome Biology* 2017, **18**:5610.1186/s13059-017-1187-1Available: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1187-1.

122. Zhao L, Sun M, Li Z, Bai X, Yu M, Wang M, Liang L, Shao X, Arnovitz S, Wang Q, He C, Lu X, Chen J, Xie H: **The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation**. *Genome Research* 2014, **24**:1296–130710.1101/gr.163147.113Available: http://genome.cshlp.org/lookup/doi/10.1101/gr.163147.113.

123. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, Laird PW, Berman BP: **DNA methylation loss in late-replicating domains is linked to mitotic cell division**. *Nature Genetics* 2018, **50**:591–60210.1038/s41588-018-0073-4Available: http://dx.doi.org/10.1038/s41588-018-0073-4.

124. Hatada I, Hayashizaki Y, Hirotsune S, Komatsubara H, Mukai T: **A genomic scanning method for higher organisms using restriction sites as landmarks.** *Proceedings of the National Academy of Sciences of the United States of America* 1991, **88**:9523–710.1073/pnas.88.21.9523Available: http://www.ncbi.nlm.nih.gov/pubmed/1946366.

125. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D: **Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells**. *Nature Genetics* 2005, **37**:853–86210.1038/ng1598Available: http://www.ncbi.nlm.nih.gov/pubmed/16007088.

126. Harrison A, Parle-McDermott A: **DNA Methylation: A Timeline of Methods and Applications**. *Frontiers in Genetics* 2011, **2**:632–64910.3389/fgene.2011.00074Available: http://journal.frontiersin.org/article/10.3389/fgene.2011.00074/abstract.

127. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJP, Durbin R, Tavaré S, Beck S: **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis**. *Nature Biotechnology* 2008, **26**:779–78510.1038/nbt1414Available: http://www.nature.com/articles/nbt1414.

128. Lienhard M, Grimm C, Morkel M, Herwig R, Chavez L: **MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments**. *Bioinformatics* 2014, **30**:284–28610.1093/bioinformatics/btt650Available: http://www.ncbi.nlm.nih.gov/pubmed/24227674.

129. Serre D, Lee BH, Ting AH: **MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome**.

Nucleic Acids Research 2010, **38**:391–39910.1093/nar/gkp992Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp992.

130. Li J, Yang J, Zhou P, Le Y, Zhou C, Wang S, Xu D, Lin H-K, Gong Z: **Circular RNAs in cancer: novel insights into origins, properties, functions and implications.** *American journal of cancer research* 2015, **5**:472–80Available: http://www.ncbi.nlm.nih.gov/pubmed/25973291.

131. Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, Jäger N, Gnirke A, Stunnenberg HG, Meissner A: **Quantitative comparison of genome-wide DNA methylation mapping technologies**. *Nature Biotechnology* 2010, **28**:1106–111410.1038/nbt.1681Available: http://www.ncbi.nlm.nih.gov/pubmed/20852634.

132. Hayatsu H, Wataya Y, Kai K, Iida S: **Reaction of sodium bisulfite with uracil, cytosine, and their derivatives**. *Biochemistry* 1970, **9**:2858–286510.1021/bi00816a016Available: http://pubs.acs.org/doi/abs/10.1021/bi00816a016.

133. Adamowicz M, Maratou K, Aitman TJ: **Multiplexed DNA Methylation Analysis of Target Regions Using Microfluidics (Fluidigm)**. In *Methods in molecular biology (clifton, n.j.)* Vol. 1708 2018:349–363. Available: http://www.ncbi.nlm.nih.gov/pubmed/29224153.

134. Skvortsova K, Zotenko E, Luu P-L, Gould CM, Nair SS, Clark SJ, Stirzaker C: **Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA**. *Epigenetics & Chromatin* 2017, **10**:1610.1186/s13072-017-0123-7Available: http://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-017-0123-7.

135. Williams L, Bei Y, Church HE, Dai N, Dimalanta ET, Ettwiller LM, Evans TC, Langhorst BW, Borgaro JG, Guan S, Marks K, Menin JF, Nichols NM, Chaithanya Ponnaluri VK, Saleh L, Samaranayake M, Sexton BS, Sun Z, Tamanaha E, Vaisvila R, Yigit E, Davis TB: **Enzymatic Methyl-seq: The Next Generation of Methylome Analysis BISULFITE SEQUENCING ALTERNATIVE METHODS FOR DETECTING 5mC AND 5hmC**. 2019, :2–5Available: https://international.neb.com/tools-and-resources/feature-articles/enzymatic-methyl-seq-the-next-generation-of-methylome-analysis.

136. Rhoads A, Au KF: **PacBio Sequencing and Its Applications**. *Genomics, Proteomics & Bioinformatics* 2015, **13**:278–28910.1016/j.gpb.2015.08.002Available: https://linkinghub.elsevier.com/retrieve/pii/S1672022915001345.

137. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W: **Detecting DNA cytosine methylation using nanopore sequencing**. *Nature Methods* 2017, **14**:407–41010.1038/nmeth.4184Available: http://www.nature.com/articles/nmeth.4184.

138. Wilson VL, Jones PA: **DNA methylation decreases in aging but not in immortal cells.** *Science (New York, N.Y.)* 1983, **220**:1055–7Available: http://www.ncbi.nlm.nih.gov/pub

med/6844925.

139. Feinberg AP, Vogelstein B: **Hypomethylation distinguishes genes of some human cancers from their normal counterparts**. *Nature* 1983, **301**:89–9210.1038/301089a0Available: http://www.nature.com/articles/301089a0.

140. Romanov GA, Vanyushin BF: **Methylation of reiterated sequences in mammalian DNAs. Effects of the tissue type, age, malignancy and hormonal induction.** *Biochimica et biophysica acta* 1981, **653**:204–18Available: http://www.ncbi.nlm.nih.gov/pubmed/7225396.

141. Berdyshev GD, Korotaev GK, Boiarskikh GV, Vaniushin BF: **Nucleotide composition of DNA and RNA from somatic tissues of humpback and its changes during spawning.** *Biokhimiia (Moscow, Russia)* 1967, **32**:988–93Available: http://www.ncbi.nlm.nih.gov/pubmed/5628601.

142. Wilson VL, Smith RA, Ma S, Cutler RG: **Genomic 5-methyldeoxycytidine decreases with age.** *The Journal of biological chemistry* 1987, **262**:9948–51Available: http://www.ncbi.nlm.nih.gov/pubmed/3611071.

143. Cole JJ, Robertson NA, Rather MI, Thomson JP, McBryan T, Sproul D, Wang T, Brock C, Clark W, Ideker T, Meehan RR, Miller RA, Brown-Borg HM, Adams PD: **Diverse interventions that extend mouse lifespan suppress shared age-associated epigenetic changes at critical gene regulatory regions**. *Genome Biology* 2017, **18**:5810.1186/s13059-017-1185-3Available: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1185-3.

144. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R: **Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.** *Nucleic acids research* 2005, **33**:5868–7710.1093/nar/gki901Available: http://www.ncbi.nlm.nih.gov/pubmed/16224102.

145. Feinberg AP, Tycko B: **The history of cancer epigenetics**. *Nature Reviews Cancer* 2004, **4**:143–15310.1038/nrc1279Available: http://www.nature.com/articles/nrc1279.

146. Knudson AG: **Mutation and cancer: statistical study of retinoblastoma.** *Proceedings of the National Academy of Sciences of the United States of America* 1971, **68**:820–310.1073/pnas.68.4.820Available: http://www.ncbi.nlm.nih.gov/pubmed/5279523.

147. Esteller M, Silva JM, Dominguez G, Bonilla F, Matias-Guiu X, Lerma E, Bussaglia E, Prat J, Harkes IC, Repasky EA, Gabrielson E, Schutte M, Baylin SB, Herman JG: **Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors.** *Journal of the National Cancer Institute* 2000, **92**:564–910.1093/jnci/92.7.564Available: http://www.ncbi.nlm.nih.gov/pubmed/10749912.

148. Feinberg AP: **The Key Role of Epigenetics in Human Disease Prevention and Mit-**

igation. *New England Journal of Medicine* 2018, **378**:1323–133410.1056/NEJMra1402513Available: http://www.ncbi.nlm.nih.gov/pubmed/29617578.

149. Bibikova M: **High-throughput DNA methylation profiling using universal bead arrays**. *Genome Research* 2006, **16**:383–39310.1101/gr.4410706Available: http://www.genome.org/cgi/doi/10.1101/gr.4410706.

150. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S: **DNA methylation profiling of human chromosomes 6, 20 and 22**. *Nature Genetics* 2006, **38**:1378–138510.1038/ng1909Available: http://www.nature.com/doifinder/10.1038/ng1909.

151. Bjornsson HT: **Intra-individual Change Over Time in DNA Methylation With Familial Clustering**. *JAMA* 2008, **299**:287710.1001/jama.299.24.2877Available: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.299.24.2877.

152. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, Nelson HH, Karagas MR, Padbury JF, Bueno R, Sugarbaker DJ, Yeh R-F, Wiencke JK, Kelsey KT: **Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context**. *PLoS Genetics* 2009, **5**:e100060210.1371/journal.pgen.1000602Available: https://dx.plos.org/10.1371/journal.pgen.1000602.

153. Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, Whittaker P, McCann OT, Finer S, Valdes AM, Leslie RD, Deloukas P, Spector TD: **Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains**. *Genome Research* 2010, **20**:434–43910.1101/gr.103101.109Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.103101.109.

154. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, Gunderson KL: **Genome-wide DNA methylation profiling using Infinium assay**. *Epigenomics* 2009, **1**:177–20010.2217/epi.09.14Available: https://www.futuremedicine.com/doi/10.2217/epi.09.14.

155. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, Savage DA, Mueller-Holzner E, Marth C, Kocjan G, Gayther SA, Jones A, Beck S, Wagner W, Laird PW, Jacobs IJ, Widschwendter M: **Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer.** *Genome research* 2010, **20**:440–610.1101/gr.103606.109Available: http://www.ncbi.nlm.nih.gov/pubmed/20219944.

156. Jung M, Pfeifer GP: **Aging and DNA methylation**. *BMC Biology* 2015,

13:710.1186/s12915-015-0118-4Available: http://www.biomedcentral.com/1741-7007/13/7.

157. Bell CG, Teschendorff AE, Rakyan VK, Maxwell AP, Beck S, Savage Da: **Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus.** *BMC Medical Genomics* 2010, **3**:3310.1186/1755-8794-3-33Available: http://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-3-33.

158. Bocklandt S, Lin W, Sehl ME, Sánchez FJ, Sinsheimer JS, Horvath S, Vilain E: **Epigenetic Predictor of Age.** *PLoS ONE* 2011, **6**:e1482110.1371/journal.pone.0014821Available: https://dx.plos.org/10.1371/journal.pone.0014821.

159. Koch CM, Wagner W: **Epigenetic-aging-signature to determine age in different tissues.** *Aging* 2011, **3**:1018–2710.18632/aging.100395Available: http://www.ncbi.nlm.nih.gov/pubmed/22067257.

160. Bell JT, Tsai P-C, Yang T-P, Pidsley R, Nisbet J, Glass D, Mangino M, Zhai G, Zhang F, Valdes A, Shin S-Y, Dempster EL, Murray RM, Grundberg E, Hedman AK, Nica A, Small KS, MuTHER Consortium, Dermitzakis ET, McCarthy MI, Mill J, Spector TD, Deloukas P: **Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population.** *PLoS genetics* 2012, **8**:e100262910.1371/journal.pgen.1002629Available: http://www.ncbi.nlm.nih.gov/pubmed/22532803.

161. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut JV, Setien F, Carmona FJ, Puca AA, Sayols S, Pujana MA, Serra-Musach J, Iglesias-Platas I, Formiga F, Fernandez AF, Fraga MF, Heath SC, Valencia A, Gut IG, Wang J, Esteller M: **Distinct DNA methylomes of newborns and centenarians.** *Proceedings of the National Academy of Sciences* 2012, **109**:10522–1052710.1073/pnas.1120658109Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1120658109.

162. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan J-B, Shen R: **High density DNA methylation array with single CpG site resolution.** *Genomics* 2011, **98**:288–9510.1016/j.ygeno.2011.07.007Available: http://www.ncbi.nlm.nih.gov/pubmed/21839163.

163. Affinito O, Scala G, Palumbo D, Florio E, Monticelli A, Miele G, Avvedimento VE, Usiello A, Chiariotti L, Cocozza S: **Modeling DNA methylation by analyzing the individual configurations of single molecules.** *Epigenetics* 2016, **11**:881–8810.1080/15592294.2016.1246108Available: http://www.ncbi.nlm.nih.gov/pubmed/27748645.

164. Haerter JO, Lövkvist C, Dodd IB, Sneppen K: **Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states.** *Nucleic Acids Research* 2014, **42**:2235–224410.1093/nar/gkt1235Available: http://www.pubmedcentral.nih.go

v/articlerender.fcgi?artid=PMC3936770.

165. Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr ARW, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R, Bird A: **CpG islands influence chromatin structure via the CpG-binding protein Cfp1.** *Nature* 2010, **464**:1082–610.1038/nature08924Available: http://www.ncbi.nlm.nih.gov/pubmed/20393567.

166. Garagnani P, Bacalini MG, Pirazzini C, Gori D, Giuliani C, Mari D, Di Blasio AM, Gentilini D, Vitale G, Collino S, Rezzi S, Castellani G, Capri M, Salvioli S, Franceschi C: **Methylation of ELOVL2 gene as a new epigenetic marker of age.** *Aging cell* 2012, **11**:1132–410.1111/acel.12005Available: http://www.ncbi.nlm.nih.gov/pubmed/23061750.

167. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan J-B, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K: **Genome-wide methylation profiles reveal quantitative views of human aging rates.** *Molecular cell* 2013, **49**:359–36710.1016/j.molcel.2012.10.016Available: http://www.ncbi.nlm.nih.gov/pubmed/23177740.

168. Horvath S: **DNA methylation age of human tissues and cell types.** *Genome biology* 2013, **14**:R11510.1186/gb-2013-14-10-r115Available: http://www.ncbi.nlm.nih.gov/pubmed/24138928.

169. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *Journal of Statistical Software* 2010, **33**:7–1010.18637/jss.v033.i01Available: http://www.jstatsoft.org/v33/i01/.

170. Florath I, Butterbach K, Muller H, Bewerunge-Hudler M, Brenner H: **Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites.** *Human Molecular Genetics* 2014, **23**:1186–120110.1093/hmg/ddt531Available: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddt531.

171. Bacalini MG, Boattini A, Gentilini D, Giampieri E, Pirazzini C, Giuliani C, Fontanesi E, Remondini D, Capri M, Del Rio A, Luiselli D, Vitale G, Mari D, Castellani G, Di Blasio AM, Salvioli S, Franceschi C, Garagnani P: **A meta-analysis on age-associated changes in blood DNA methylation: results from an original analysis pipeline for Infinium 450k data.** *Aging* 2015, **7**:97–10910.18632/aging.100718Available: http://www.aging-us.com/article/100718.

172. Zaghlool SB, Al-Shafai M, Al Muftah WA, Kumar P, Falchi M, Suhre K: **Association of DNA methylation with age, gender, and smoking in an Arab population.** *Clinical Epigenetics* 2015, **7**:610.1186/s13148-014-0040-6Available: http://www.ncbi.nlm.nih.gov/pubmed/25663950.

173. Benton MC, Sutherland HG, Macartney-Coxson D, Haupt LM, Lea RA, Griffiths LR: **Methylome-wide association study of whole blood DNA in the Norfolk Island isolate identifies robust loci associated with age.** *Aging* 2017, **9**:753–76810.18632/aging.101187Available: http://www.aging-us.com/article/101187/text.

174. Johnson ND, Wiener HW, Smith AK, Nishitani S, Absher DM, Arnett DK, Aslibekyan S, Conneely KN: **Non-linear patterns in age-related DNA methylation may reflect CD4 + T cell differentiation.** *Epigenetics* 2017, **12**:492–50310.1080/15592294.2017.1314419Available: http://www.ncbi.nlm.nih.gov/pubmed/28387568.

175. Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, Tylavsky FA, Conneely KN: **Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type.** *BMC Genomics* 2014, **15**:14510.1186/1471-2164-15-145Available: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-145.

176. Kuehnen P, Mischke M, Wiegand S, Sers C, Horsthemke B, Lau S, Keil T, Lee YA, Grueters A, Krude H: **An alu element-associated hypermethylation variant of the POMC gene is associated with childhood obesity.** *PLoS Genetics* 2012, **8**:1–1210.1371/journal.pgen.1002543.

177. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y: **Abundant contribution of short tandem repeats to gene expression variation in humans.** *Nature Genetics* 2016, **48**:22–2910.1038/ng.3461Available: http://dx.doi.org/10.1038/ng.3461.

178. Velten L, Haas SF, Raffel S, Blaszkiewicz S, Islam S, Hennig BP, Hirche C, Lutz C, Buss EC, Nowak D, Boch T, Hofmann W, Ho AD, Huber W, Trumpp A, Essers MAG, Steinmetz LM: **Human haematopoietic stem cell lineage commitment is a continuous process.** *Nature Cell Biology* 2017, **19**:271–28110.1038/ncb3493Available: http://www.nature.com/articles/ncb3493.

179. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, Lindsley RC, Mermel CH, Burtt N, Chavez A, Higgins JM, Moltchanov V, Kuo FC, Kluk MJ, Henderson B, Kinnunen L, Koistinen Ha, Ladenvall C, Getz G, Correa A, Banahan BF, Gabriel S, Kathiresan S, Stringham HM, McCarthy MI, Boehnke M, Tuomilehto J, Haiman C, Groop L, Atzmon G, et al.: **Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes.** *New England Journal of Medicine* 2014, **371**:2488–249810.1056/NEJMoa1408617Available: http://www.ncbi.nlm.nih.gov/pubmed/25426837.

180. Watson CJ, Papula AL, Poon GYP, Wong WH, Young AL, Druley TE, Fisher DS, Blundell JR: **The evolutionary dynamics and fitness landscape of clonal hematopoiesis.** *Science* 2020, **367**:1449–145410.1126/science.aay9333Available: https://www.sciencemag.org/lookup/do

i/10.1126/science.aay9333.

181. Kirschner K, Chandra T, Kiselev V, Flores-Santa Cruz D, Macaulay IC, Park HJ, Li J, Kent DG, Kumar R, Pask DC, Hamilton TL, Hemberg M, Reik W, Green AR: **Proliferation Drives Aging-Related Functional Decline in a Subpopulation of the Hematopoietic Stem Cell Compartment**. *Cell Reports* 2017, **19**:1503–151110.1016/j.celrep.2017.04.074Available: https://linkinghub.elsevier.com/retrieve/pii/S2211124717306034.

182. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, Lindsley RC, Mermel CH, Burtt N, Chavez A, Higgins JM, Moltchanov V, Kuo FC, Kluk MJ, Henderson B, Kinnunen L, Koistinen HA, Ladenvall C, Getz G, Correa A, Banahan BF, Gabriel S, Kathiresan S, Stringham HM, McCarthy MI, Boehnke M, Tuomilehto J, Haiman C, Groop L, Atzmon G, et al.: **Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes**. *New England Journal of Medicine* 2014, **371**:2488–249810.1056/NEJMoa1408617Available: http://www.nejm.org/doi/10.1056/NEJMoa1408617.

183. Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, Bauerschlag DO, Jöckel K-H, Erbel R, Mühleisen TW, Zenke M, Brümmendorf TH, Wagner W: **Aging of blood can be tracked by DNA methylation changes at just three CpG sites.** *Genome biology* 2014, **15**:R2410.1186/gb-2014-15-2-r24Available: http://www.ncbi.nlm.nih.gov/pubmed/24490752.

184. Johnson TE: **Recent results: Biomarkers of aging**. *Experimental Gerontology* 2006, **41**:1243–124610.1016/j.exger.2006.09.006Available: https://linkinghub.elsevier.com/retrieve/pii/S0531556506002865.

185. Thompson MJ, VonHoldt B, Horvath S, Pellegrini M: **An epigenetic aging clock for dogs and wolves.** *Aging* 2017, **9**:1055–106810.18632/aging.101211Available: http://www.ncbi.nlm.nih.gov/pubmed/28373601.

186. Stubbs TM, Bonder MJ, Stark A-K, Krueger F, Meyenn F von, Stegle O, Reik W: **Multi-tissue DNA methylation age predictor in mouse**. *Genome Biology* 2017, **18**:6810.1186/s13059-017-1203-5Available: http://www.ncbi.nlm.nih.gov/pubmed/28399939.

187. Lowe R, Danson AF, Rakyan VK, Yildizoglu S, Saldmann F, Viltard M, Friedlander G, Faulkes CG: **DNA methylation clocks as a predictor for ageing and age estimation in naked mole-rats , Heterocephalus glaber**. 2020, **12**:1–1310.18632/aging.102892.

188. Lowe R, Barton C, Jenkins CA, Ernst C, Forman O, Fernandez-Twinn DS, Bock C, Rossiter SJ, Faulkes CG, Ozanne SE, Walter L, Odom DT, Mellersh C, Rakyan VK: **Ageing-associated DNA methylation dynamics are a molecular readout of lifespan variation among mammalian species.** *Genome Biology* 2018, **19**:2210.1186/s13059-018-1397-1Available: http://www.ncbi.nlm.nih.gov/pubmed/29452591.

189. Wang M, Lemos B: **Ribosomal DNA harbors an evolutionarily conserved clock**

of biological aging. *Genome research* 2019, **29**:325–33310.1101/gr.241745.118Available: http://www.ncbi.nlm.nih.gov/pubmed/30765617.

190. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, Gibson J, Henders AK, Redmond P, Cox SR, Pattie A, Corley J, Murphy L, Martin NG, Montgomery GW, Feinberg AP, Fallin MD, Multhaup ML, Jaffe AE, Joehanes R, Schwartz J, Just AC, Lunetta KL, Murabito JM, Starr JM, Horvath S, Baccarelli AA, Levy D, Visscher PM, Wray NR, et al.: **DNA methylation age of blood predicts all-cause mortality in later life.** *Genome biology* 2015, **16**:2510.1186/s13059-015-0584-6Available: http://www.ncbi.nlm.nih.gov/pubmed/25633388.

191. Christiansen L, Lenart A, Tan Q, Vaupel JW, Aviv A, McGue M, Christensen K: **DNA methylation age is associated with mortality in a longitudinal Danish twin study**. *Aging Cell* 2016, **15**:149–15410.1111/acel.12421Available: http://doi.wiley.com/10.1111/acel.12421.

192. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai P-C, Roetker NS, Just AC, Demerath EW, Guan W, Bressler J, Fornage M, Studenski S, Vandiver AR, Moore AZ, Tanaka T, Kiel DP, Liang L, Vokonas P, Schwartz J, Lunetta KL, Murabito JM, Bandinelli S, Hernandez DG, Melzer D, Nalls M, Pilling LC, Price TR, Singleton AB, Gieger C, et al.: **DNA methylation-based measures of biological age: meta-analysis predicting time to death.** *Aging* 2016, **8**:1844–186510.18632/aging.101020Available: http://www.ncbi.nlm.nih.gov/pubmed/27690265.

193. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, Hou L, Baccarelli AA, Stewart JD, Li Y, Whitsel EA, Wilson JG, Reiner AP, Aviv A, Lohman K, Liu Y, Ferrucci L, Horvath S: **An epigenetic biomarker of aging for lifespan and healthspan**. *Aging* 2018, **10**:573–59110.18632/aging.101414Available: http://www.ncbi.nlm.nih.gov/pubmed/29676998.

194. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, Hou L, Baccarelli AA, Li Y, Stewart JD, Whitsel EA, Assimes TL, Ferrucci L, Horvath S: **DNA methylation GrimAge strongly predicts lifespan and healthspan**. *Aging* 2019, **11**:303–32710.18632/aging.101684Available: http://www.aging-us.com/article/101684/text.

195. Halloran AMO, Hever A, Ni C, Ake C, Horvath S: **Association of 4 epigenetic clocks with measures of functional health, cognition, and all- cause mortality in The Irish Longitudinal Study on Ageing (TILDA)**. 2020, :1–3510.1101/2020.04.27.063164Available: https://www.biorxiv.org/content/10.1101/2020.04.27.063164v1.

196. Belsky DW, Caspi A, Arseneault L, Baccarelli A, Corcoran DL, Gao X, Hannon E, Harrington HL, Rasmussen LJH, Houts R, Huffman K, Kraus WE, Kwon D, Mill J, Pieper CF, Prinz JA, Poulton R, Schwartz J, Sugden K, Vokonas P, Williams BS, Moffitt TE: **Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm**. *eLife* 2020, **9**:1–5610.7554/eLife.54870Available:

https://elifesciences.org/articles/54870.

197. Fahy GM, Brooke RT, Watson JP, Good Z, Vasanawala SS, Maecker H, Leipold MD, Lin DTS, Kobor MS, Horvath S: **Reversal of epigenetic aging and immunosenescent trends in humans**. *Aging Cell* 2019, :1–1210.1111/acel.13028.

198. Lu AT, Seeboth A, Tsai PC, Sun D, Quach A, Reiner AP, Kooperberg C, Ferrucci L, Hou L, Baccarelli AA, Li Y, Harris SE, Corley J, Taylor A, Deary IJ, Stewart JD, Whitsel EA, Assimes TL, Chen W, Li S, Mangino M, Bell JT, Wilson JG, Aviv A, Marioni RE, Raj K, Horvath S: **DNA methylation-based estimator of telomere length**. *Aging* 2019, **11**:5895–592310.18632/aging.102173.

199. Naylor K, Li G, Vallejo AN, Lee W-W, Koetz K, Bryl E, Witkowski J, Fulbright J, Weyand CM, Goronzy JJ: **The Influence of Age on T Cell Generation and TCR Diversity**. *The Journal of Immunology* 2005, **174**:7446–745210.4049/jimmunol.174.11.7446Available: http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.174.11.7446.

200. MARCUS R, BUTTERFIELD G, HOLLOWAY L, GILLILAND L, BAYLINK DJ, HINTZ RL, SHERMAN BM: **Effects of Short Term Administration of Recombinant Human Growth Hormone to Elderly People***. *The Journal of Clinical Endocrinology & Metabolism* 1990, **70**:519–52710.1210/jcem-70-2-519Available: https://academic.oup.com/jcem/article-lookup/doi/10.1210/jcem-70-2-519.

201. Gensous N, Franceschi C, Santoro A, Milazzo M, Garagnani P, Bacalini MG: **The Impact of Caloric Restriction on the Epigenetic Signatures of Aging**. *International Journal of Molecular Sciences* 2019, **20**:202210.3390/ijms20082022Available: https://www.mdpi.com/1422-0067/20/8/2022.

202. Wang T, Tsui B, Kreisberg JF, Robertson NA, Gross AM, Yu MK, Carter H, Brown-Borg HM, Adams PD, Ideker T: **Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment**. *Genome Biology* 2017, **18**:5710.1186/s13059-017-1186-2Available: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1186-2.

203. Horvath S, Singh K, Raj K, Khairnar S, Sanghavi A, Shrivastava A, Zoller JA, Li CZ, Herenu CB, Canatelli-Mallat M, Lehmann M, Woods LCS, Martinez AG, Wang T, Chiavellini P, Levine AJ, Chen H, Goya RG, Katcher HL: **Reversing age: dual species measurement of epigenetic age with a single clock**. *bioRxiv* 2020, :2020.05.07.08291710.1101/2020.05.07.082917Available: https://www.biorxiv.org/content/10.1101/2020.05.07.082917v1.

204. Kulkarni AS, Gubbi S, Barzilai N: **Benefits of Metformin in Attenuating the Hallmarks of Aging**. *Cell Metabolism* 2020, **32**:15–3010.1016/j.cmet.2020.04.001Available: https:

//doi.org/10.1016/j.cmet.2020.04.001.

205. TAME trial website: 2020, Available: https://www.afar.org/tame-trial.

206. Justice JN, Ferrucci L, Newman AB, Aroda VR, Bahnson JL, Divers J, Espeland MA, Marcovina S, Pollak MN, Kritchevsky SB, Barzilai N, Kuchel GA: **A framework for selection of blood-based biomarkers for geroscience-guided clinical trials: report from the TAME Biomarkers Workgroup**. *GeroScience* 2018, **40**:419–43610.1007/s11357-018-0042-yAvailable: http://link.springer.com/10.1007/s11357-018-0042-y.

207. Harvey NC, Javaid K, Bishop N, Kennedy S, Papageorghiou AT, Fraser R, Gandhi SV, Schoenmakers I, Prentice A, Cooper C: **MAVIDOS Maternal Vitamin D Osteoporosis Study: study protocol for a randomized controlled trial. The MAVIDOS Study Group**. *Trials* 2012, **13**:1310.1186/1745-6215-13-13Available: http://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-13-13.

208. Cooper C, Harvey NC, Bishop NJ, Kennedy S, Papageorghiou AT, Schoenmakers I, Fraser R, Gandhi SV, Carr A, D'Angelo S, Crozier SR, Moon RJ, Arden NK, Dennison EM, Godfrey KM, Inskip HM, Prentice A, Mughal MZ, Eastell R, Reid DM, Javaid MK: **Maternal gestational vitamin D supplementation and offspring bone health (MAVIDOS): a multicentre, double-blind, randomised placebo-controlled trial**. *The Lancet Diabetes & Endocrinology* 2016, **4**:393–40210.1016/S2213-8587(16)00044-9Available: http://dx.doi.org/10.1016/S2213-8587(16)00044-9.

209. Bell CG, Xia Y, Yuan W, Gao F, Ward K, Roos L, Mangino M, Hysi PG, Bell J, Wang J, Spector TD: **Novel regional age-associated DNA methylation changes within human common disease-associated loci**. *Genome Biology* 2016, **17**:19310.1186/s13059-016-1051-8Available: http://www.ncbi.nlm.nih.gov/pubmed/27663977.

210. Moran S, Arribas C, Esteller M: **Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences.** *Epigenomics* 2016, **8**:389–9910.2217/epi.15.114Available: http://www.ncbi.nlm.nih.gov/pubmed/26673039.

211. Gunderson KL: **Decoding Randomly Ordered DNA Arrays**. *Genome Research* 2004, **14**:870–87710.1101/gr.2255804Available: http://www.genome.org/cgi/doi/10.1101/gr.2255804.

212. Michael KL, Taylor LC, Schultz SL, Walt DR: **Randomly Ordered Addressable High-Density Optical Sensor Arrays**. *Analytical Chemistry* 1998, **70**:1242–124810.1021/ac971343rAvailable: https://pubs.acs.org/doi/10.1021/ac971343r.

213. Clark C, Palta P, Joyce CJ, Scott C, Grundberg E, Deloukas P, Palotie A, Coffey AJ: **A Comparison of the Whole Genome Approach of MeDIP-Seq to the Targeted Approach of the Infinium HumanMethylation450 BeadChip for Methylome Profiling**. *PLoS ONE* 2012, **7**:e5023310.1371/journal.pone.0050233Available: https://dx.plos.org/10.1371/

journal.pone.0050233.

214. Smith ML, Baggerly KA, Bengtsson H, Ritchie ME, Hansen KD: **illuminaio: An open source IDAT parsing tool for Illumina microarrays**. *F1000Research* 2013, **2**:1–810.12688/f1000research.2-264.v1.

215. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F: **Evaluation of the Infinium Methylation 450K technology**. *Epigenomics* 2011, **3**:771–78410.2217/epi.11.105Available: https://www.futuremedicine.com/doi/10.2217/epi.11.105.

216. Fortin J-P, Triche TJ, Hansen KD: **Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi**. *Bioinformatics* 2016, :btw69110.1093/bioinformatics/btw691Available: http://biorxiv.org/lookup/doi/10.1101/065490.

217. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S: **A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data**. *Bioinformatics* 2013, **29**:189–19610.1093/bioinformatics/bts680Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts680.

218. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data**. *Nature Reviews Genetics* 2010, **11**:733–73910.1038/nrg2825Available: http://www.nature.com/articles/nrg2825.

219. Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, Greenwood CMT, Hansen KD: **Functional normalization of 450k methylation array data improves replication in large cancer studies**. *Genome Biology* 2014, **15**:50310.1186/s13059-014-0503-2Available: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0503-2.

220. Ehrich M, Zoll S, Sur S, Boom D van den: **A new method for accurate assessment of DNA quality after bisulfite treatment.** *Nucleic acids research* 2007, **35**:e2910.1093/nar/gkl1134Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl1134%20http://www.ncbi.nlm.nih.gov/pubmed/17259213%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1865059.

221. Illumina: *Infinium HD Assay Methylation Protocol Guide*. Illumina; 2015 Available: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry%7B/_%7Ddocumentation/infinium%7B/_%7Dassays/infinium%7B/_%7Dhd%7B/_%7Dmethylation/infinium-hd-methylation-guide-15019519-01.pdf.

222. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, Van Djik S, Muhlhausler B, Stirzaker C, Clark SJ: **Critical evaluation of the Illumina Methy-**

lationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome biology* 2016, **17**:20810.1186/s13059-016-1066-1Available: http://www.ncbi.nlm.nih.gov/pubmed/27717381.

223. Zhou W, Laird PW, Shen H: **Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes**. *Nucleic Acids Research* 2016, **45**:gkw96710.1093/nar/gkw967Available: http://www.ncbi.nlm.nih.gov/pubmed/27924034.

224. Birney E, Smith GD, Greally JM: **Epigenome-wide Association Studies and the Interpretation of Disease -Omics**. *PLOS Genetics* 2016, **12**:e100610510.1371/journal.pgen.1006105Available: http://www.ncbi.nlm.nih.gov/pubmed/27336614.

225. Lappalainen T, Greally JM: **Associating cellular epigenetic models with human phenotypes**. *Nature Reviews Genetics* 2017, **18**:441–45110.1038/nrg.2017.32Available: http://www.nature.com/doifinder/10.1038/nrg.2017.32.

226. Andrews SV, Ladd-Acosta C, Feinberg AP, Hansen KD, Fallin MD: **"Gap hunting" to characterize clustered probe signals in Illumina methylation array data**. *Epigenetics & Chromatin* 2016, **9**:5610.1186/s13072-016-0107-zAvailable: http://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-016-0107-z.

227. Taiwo O, Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D, Beck S, Butcher LM: **Methylome analysis using MeDIP-seq with low DNA concentrations**. *Nature Protocols* 2012, **7**:617–63610.1038/nprot.2012.012Available: http://www.nature.com/articles/nprot.2012.012.

228. Chavez L, Jozefczuk J, Grimm C, Dietrich J, Timmermann B, Lehrach H, Herwig R, Adjaye J: **Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage**. *Genome Research* 2010, **20**:1441–145010.1101/gr.110114.110Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.110114.110.

229. Andrews S: **FastQC**. 2010, Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

230. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078–207910.1093/bioinformatics/btp352Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352.

231. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**:841–84210.1093/bioinformatics/btq033Available: http://bedtools.readthedocs.io/en/latest/index.html.

232. Moayyeri A, Hammond CJ, Valdes AM, Spector TD: **Cohort Profile: TwinsUK and Healthy Ageing Twin Study**. *International Journal of Epidemiology* 2013, **42**:76–8510.1093/ije/dyr207Available: http://www.ncbi.nlm.nih.gov/pubmed/22253318.

233. Korbie D, Lin E, Wall D, Nair SS, Stirzaker C, Clark SJ, Trau M: **Multiplex bisulfite PCR resequencing of clinical FFPE DNA**. *Clinical Epigenetics* 2015, **7**:2810.1186/s13148-015-0067-3Available: https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-015-0067-3.

234. Tran H, Porter J, Sun M, Xie H, Zhang L: **Objective and Comprehensive Evaluation of Bisulfite Short Read Mapping Tools**. *Advances in Bioinformatics* 2014, **2014**:1–1110.1155/2014/472045Available: http://www.hindawi.com/journals/abi/2014/472045/.

235. Böttcher R, Amberg R, Ruzius FP, Guryev V, Verhaegh WFJ, Beyerlein P, Zaag PJ van der: **Using a priori knowledge to align sequencing reads to their exact genomic position**. *Nucleic Acids Research* 2012, **40**:e125–e12510.1093/nar/gks393Available: https://academic.oup.com/nar/article/40/16/e125/1026881.

236. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications**. *Bioinformatics* 2011, **27**:1571–157210.1093/bioinformatics/btr167Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr167.

237. Li L-C, Dahiya R: **MethPrimer: designing primers for methylation PCRs.** *Bioinformatics (Oxford, England)* 2002, **18**:1427–31Available: http://www.ncbi.nlm.nih.gov/pubmed/12424112.

238. Staa TP van, Dennison EM, Leufkens HGM, Cooper C: **Epidemiology of fractures in England and Wales**. *Bone* 2001, **29**:517–52210.1016/S8756-3282(01)00614-7Available: https://linkinghub.elsevier.com/retrieve/pii/S8756328201006147.

239. Streubel PN, Ricci WM, Wong A, Gardner MJ: **Mortality After Distal Femur Fractures in Elderly Patients**. *Clinical Orthopaedics and Related Research* 2011, **469**:1188–119610.1007/s11999-010-1530-2Available: http://link.springer.com/10.1007/s11999-010-1530-2.

240. Thompson E, Greenspan S: *National Osteoporosis Foundation Annual Report (2018)*. National Osteoporosis Foundation; 2018 Available: https://cdn.nof.org/wp-content/uploads/2018%7B/_%7DNOF%7B/_%7DAnnual%7B/_%7Dreport%7B/_%7DFINAL.pdf.

241. Hernandez CJ, Beaupré GS, Carter DR: **A theoretical analysis of the relative influences of peak BMD, age-related bone loss and menopause on the development of osteoporosis**. *Osteoporosis International* 2003, **14**:843–84710.1007/s00198-003-1454-8.

242. Nguyen TV, Center JR, Eisman JA: **Femoral Neck Bone Loss Predicts Fracture Risk Independent of Baseline BMD**. *Journal of Bone and Mineral Research* 2005, **20**:1195–120110.1359/JBMR.050215Available: http://doi.wiley.com/10.1359/JBMR.050215.

243. Finkelstein JS, Brockwell SE, Mehta V, Greendale GA, Sowers MR, Ettinger B, Lo JC, Johnston JM, Cauley JA, Danielson ME, Neer RM: **Bone Mineral Density Changes during the Menopause Transition in a Multiethnic Cohort of Women**. *The Journal of Clinical Endocrinology & Metabolism* 2008, **93**:861–86810.1210/jc.2007-1876Available: https://academic.oup.com/jcem/article-lookup/doi/10.1210/jc.2007-1876.

244. Hui SL, Slemenda CW, Johnston CC: **The contribution of bone loss to postmenopausal osteoporosis**. *Osteoporosis International* 1990, **1**:30–3410.1007/BF01880413.

245. Harvey N, Dennison E, Cooper C: **Osteoporosis: A Lifecourse Approach**. *Journal of Bone and Mineral Research* 2014, **29**:1917–192510.1002/jbmr.2286Available: http://doi.wiley.com/10.1002/jbmr.2286.

246. Bikle DD: **Vitamin D and Bone**. *Current Osteoporosis Reports* 2012, **10**:151–15910.1007/s11914-012-0098-zAvailable: http://link.springer.com/10.1007/s11914-012-0098-z.

247. Christakos S, Dhawan P, Porta A, Mady LJ, Seth T: **Vitamin D and intestinal calcium absorption**. *Molecular and Cellular Endocrinology* 2011, **347**:25–2910.1016/j.mce.2011.05.038Available: https://linkinghub.elsevier.com/retrieve/pii/S0303720711002930.

248. Fukumoto S: **Phosphate metabolism and vitamin D**. *BoneKEy Reports* 2014, **3**:1–510.1038/bonekey.2013.231Available: http://dx.doi.org/10.1038/bonekey.2013.231.

249. Baird J, Kurshid MA, Kim M, Harvey N, Dennison E, Cooper C: **Does birthweight predict bone mass in adulthood? A systematic review and meta-analysis**. *Osteoporosis International* 2011, **22**:1323–133410.1007/s00198-010-1344-9.

250. Cooper C, Eriksson JG, Forsén T, Osmond C, Tuomilehto J, Barker DJP: **Maternal height, childhood growth and risk of hip fracture in later life: A longitudinal study**. *Osteoporosis International* 2001, **12**:623–62910.1007/s001980170061.

251. Inskip HM, Godfrey KM, Robinson SM, Law CM, Jp D: **Europe PMC Funders Group Cohort Profile : The Southampton Women ' s Survey**. *International Journal of Epidemiology* 2006, **35**:42–4810.1093/ije/dyi202.

252. Harvey NC, Mahon PA, Robinson SM, Nisbet CE, Javaid MK, Crozier SR, Inskip HM, Godfrey KM, Arden NK, Dennison EM, Cooper C, Taylor P, Greenaway LJ, Hanson M, Barker DJP, Law CM: **Different indices of fetal growth predict bone size and volumetric density at 4 years of age**. *Journal of Bone and Mineral Research* 2010, **25**:920–92710.1359/jbmr.091022.

253. Harvey NC, Cole ZA, Crozier SR, Ntani G, Mahon PA, Robinson SM, Inskip HM, Godfrey KM, Dennison EM, Cooper C: **Fetal and infant growth predict hip geometry at 6 y old: Findings from the Southampton Women's Survey**. *Pediatric Research* 2013, **74**:450–45610.1038/pr.2013.119.

254. Mahon P, Harvey N, Crozier S, Inskip H, Robinson S, Arden N, Swaminathan R, Cooper C, Godfrey K: **Low maternal vitamin D status and fetal bone development: Cohort study**. *Journal of Bone and Mineral Research* 2010, **25**:14–1910.1359/jbmr.090701Available: http://doi.wiley.com/10.1359/jbmr.090701.

255. Viljakainen HT, Saarnio E, Hytinantti T, Miettinen M, Surcel H, Mäkitie O, Andersson S, Laitinen K, Lamberg-Allardt C: **Maternal Vitamin D Status Determines Bone Variables in the Newborn**. *The Journal of Clinical Endocrinology & Metabolism* 2010, **95**:1749–175710.1210/jc.2009-1391Available: https://academic.oup.com/jcem/article-lookup/doi/10.1210/jc.2009-1391.

256. Viljakainen HT, Korhonen T, Hytinantti T, Laitinen EKA, Andersson S, Mäkitie O, Lamberg-Allardt C: **Maternal vitamin D status affects bone growth in early childhood—a prospective cohort study**. *Osteoporosis International* 2011, **22**:883–89110.1007/s00198-010-1499-4Available: http://link.springer.com/10.1007/s00198-010-1499-4.

257. Zhu K, Whitehouse AJ, Hart PH, Kusel M, Mountain J, Lye S, Pennell C, Walsh JP: **Maternal Vitamin D Status During Pregnancy and Bone Mass in Offspring at 20 Years of Age: A Prospective Cohort Study**. *Journal of Bone and Mineral Research* 2014, **29**:1088–109510.1002/jbmr.2138Available: http://doi.wiley.com/10.1002/jbmr.2138.

258. Kip SN, Strehler EE: **Vitamin D 3 upregulates plasma membrane Ca 2+ -ATPase expression and potentiates apico-basal Ca 2+ flux in MDCK cells**. *American Journal of Physiology-Renal Physiology* 2004, **286**:F363–F36910.1152/ajprenal.00076.2003Available: http://www.ncbi.nlm.nih.gov/pubmed/14583431.

259. Burdge GC, Slater-Jefferies J, Torrens C, Phillips ES, Hanson MA, Lillycrop KA: **Dietary protein restriction of pregnant rats in the F 0 generation induces altered methylation of hepatic gene promoters in the adult male offspring in the F 1 and F 2 generations**. *British Journal of Nutrition* 2007, **97**:435–43910.1017/S0007114507352392Available: https://www.cambridge.org/core/product/identifier/S0007114507352392/type/journal%7B/_%7Darticle.

260. Lillycrop KA, Phillips ES, Torrens C, Hanson MA, Jackson AA, Burdge GC: **Feeding pregnant rats a protein-restricted diet persistently alters the methylation of specific cytosines in the hepatic PPARα promoter of the offspring**. *British Journal of Nutrition* 2008, **100**:278–28210.1017/S0007114507894438Available: https://www.cambridge.org/core/product/identifier/S0007114507894438/type/journal%7B/_%7Darticle.

261. Burdge GC, Hanson MA, Slater-Jefferies JL, Lillycrop KA: **Epigenetic regulation of transcription: a mechanism for inducing variations in phenotype (fetal programming) by differences in nutrition during early life?** *British Journal of Nutrition* 2007, **97**:1036–104610.1017/S0007114507682920Available: https://www.cambridge.org/core/product/identifier/S0007114507682920/type/journal%7B/_%7Darticle.

262. Holroyd C, Harvey N, Dennison E, Cooper C: **Epigenetic influences in the developmental origins of osteoporosis**. *Osteoporosis International* 2012, **23**:401–41010.1007/s00198-011-1671-5.

263. Harvey NC, Sheppard A, Godfrey KM, McLean C, Garratt E, Ntani G, Davies L, Murray R, Inskip HM, Gluckman PD, Hanson MA, Lillycrop KA, Cooper C: **Childhood Bone Mineral Content Is Associated With Methylation Status of the RXRA Promoter at Birth**. *Journal of Bone and Mineral Research* 2014, **29**:600–60710.1002/jbmr.2056Available: http://doi.wiley.com/10.1002/jbmr.2056.

264. Curtis EM, Murray R, Titcombe P, Cook E, Clarke-Harris R, Costello P, Garratt E, Holbrook JD, Barton S, Inskip H, Godfrey KM, Bell CG, Cooper C, Lillycrop KA, Harvey NC: **Perinatal DNA Methylation at CDKN2A Is Associated With Offspring Bone Mass: Findings From the Southampton Women's Survey**. *Journal of Bone and Mineral Research* 2017, **32**:2030–204010.1002/jbmr.3153Available: http://doi.wiley.com/10.1002/jbmr.3153.

265. Richards EJ: **Inherited epigenetic variation–revisiting soft inheritance.** *Nature reviews. Genetics* 2006, **7**:395–40110.1038/nrg1834Available: http://www.ncbi.nlm.nih.gov/pubmed/16534512.

266. Holland ML, Lowe R, Caton PW, Gemma C, Carbajosa G, Danson AF, Carpenter AAM, Loche E, Ozanne SE, Rakyan VK: **Early-life nutrition modulates the epigenetic state of specific rDNA genetic variants in mice**. *Science* 2016, **353**:495–49810.1126/science.aaf7040Available: http://www.sciencemag.org/lookup/doi/10.1126/science.aaf7040.

267. Dogan MV, Beach SRH, Philibert RA: **Genetically contextual effects of smoking on genome wide DNA methylation**. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2017, **174**:595–60710.1002/ajmg.b.32565Available: http://doi.wiley.com/10.1002/ajmg.b.32565.

268. Dai L, Mehta A, Mordukhovich I, Just AC, Shen J, Hou L, Koutrakis P, Sparrow D, Vokonas PS, Baccarelli AA, Schwartz JD: **Differential DNA methylation and PM 2.5 species in a 450K epigenome-wide association study**. *Epigenetics* 2017, **12**:139–14810.1080/15592294.2016.1271853Available: https://www.tandfonline.com/doi/full/10.1080/15592294.2016.1271853.

269. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, Tsai P-C, Ried JS, Zhang W, Yang Y, Tan S, Fiorito G, Franke L, Guarrera S, Kasela S, Kriebel J, Richmond RC, Adamo M, Afzal U, Ala-Korpela M, Albetti B, Ammerpohl O, Apperley JF, Beekman M, Bertazzi PA, Black SL, Blancher C, Bonder M-J, Brosch M, Carstensen-Kirberg M, et al.: **Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity.** *Nature* 2017, **541**:81–8610.1038/nature20784Available: http://www.ncbi.nlm.nih.gov/pubmed/

28002404.

270. Min J, Hemani G, Smith GD, Relton CL, Suderman M: **Meffil: efficient normalisation and analysis of very large DNA methylation samples.** *bioRxiv* 2017, **44**:12596310.1101/125963Available: https://www.biorxiv.org/content/early/2017/04/27/125963.

271. Jaffe AE, Irizarry RA: **Accounting for cellular heterogeneity is critical in epigenome-wide association studies.** *Genome biology* 2014, **15**:R3110.1186/gb-2014-15-2-r31Available: http://www.ncbi.nlm.nih.gov/pubmed/24495553.

272. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT: **DNA methylation arrays as surrogate measures of cell mixture distribution.** *BMC Bioinformatics* 2012, **13**:8610.1186/1471-2105-13-86Available: http://www.biomedcentral.com/1471-2105/13/86.

273. Cardenas A, Allard C, Doyon M, Houseman EA, Bakulski KM, Perron P, Bouchard L, Hivert M-F: **Validation of a DNA methylation reference panel for the estimation of nucleated cells types in cord blood.** *Epigenetics* 2016, **11**:773–77910.1080/15592294.2016.1233091Available: http://www.ncbi.nlm.nih.gov/pubmed/27668573.

274. Goede OM de, Razzaghian HR, Price EM, Jones MJ, Kobor MS, Robinson WP, Lavoie PM: **Nucleated red blood cells impact DNA methylation and expression analyses of cord blood hematopoietic cells.** *Clinical Epigenetics* 2015, **7**:9510.1186/s13148-015-0129-6Available: http://dx.doi.org/10.1186/s13148-015-0129-6.

275. Bakulski KM, Feinberg JI, Andrews SV, Yang J, Brown S, L. McKenney S, Witter F, Walston J, Feinberg AP, Fallin MD: **DNA methylation of cord blood cell types: Applications for mixed cell birth studies.** *Epigenetics* 2016, **11**:354–36210.1080/15592294.2016.1161875Available: http://dx.doi.org/10.1080/15592294.2016.1161875.

276. Gervin K, Page CM, Aass HCD, Jansen MA, Fjeldstad HE, Andreassen BK, Duijts L, Meurs JB van, Zelm MC van, Jaddoe VW, Nordeng H, Knudsen GP, Magnus P, Nystad W, Staff AC, Felix JF, Lyle R: **Cell type specific DNA methylation in cord blood: A 450K-reference data set and cell count-based validation of estimated cell type composition.** *Epigenetics* 2016, **11**:690–69810.1080/15592294.2016.1214782Available: http://dx.doi.org/10.1080/15592294.2016.1214782.

277. Leek JT, Storey JD: **Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis.** *PLoS Genetics* 2007, **3**:e16110.1371/journal.pgen.0030161Available: https://dx.plos.org/10.1371/journal.pgen.0030161.

278. Teschendorff AE, Zhuang J, Widschwendter M: **Independent surrogate variable anal-**

ysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* 2011, **27**:1496–150510.1093/bioinformatics/btr171Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr171.

279. McGregor K, Bernatsky S, Colmegna I, Hudson M, Pastinen T, Labbe A, Greenwood CMT: **An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies**. *Genome Biology* 2016, **17**:8410.1186/s13059-016-0935-yAvailable: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0935-y.

280. Teschendorff AE, Zheng SC: **Cell-type deconvolution in epigenome-wide association studies: a review and recommendations**. *Epigenomics* 2017, **9**:757–76810.2217/epi-2016-0153Available: https://www.futuremedicine.com/doi/10.2217/epi-2016-0153.

281. Lin X, Tan JYL, Teh AL, Lim IY, Liew SJ, MacIsaac JL, Chong YS, Gluckman PD, Kobor MS, Cheong CY, Karnani N: **Cell type-specific DNA methylation in neonatal cord tissue and cord blood: a 850K-reference panel and comparison of cell types**. *Epigenetics* 2018, **13**:941–95810.1080/15592294.2018.1522929Available: https://doi.org/10.1080/15592294.2018.1522929.

282. Wang S: **Method to detect differentially methylated loci with case-control designs using Illumina arrays**. *Genetic Epidemiology* 2011, **35**:686–69410.1002/gepi.20619Available: http://doi.wiley.com/10.1002/gepi.20619.

283. Tsai PC, Bell JT: **Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation**. *International Journal of Epidemiology* 2015, **44**:1429–144110.1093/ije/dyv041.

284. Graw S, Henn R, Thompson JA, Koestler DC: **PwrEWAS: A user-friendly tool for comprehensive power estimation for epigenome wide association studies (EWAS)**. *BMC Bioinformatics* 2019, **20**:1–1110.1186/s12859-019-2804-7.

285. Cohen J: *Statisticsl Power Analysis for the Behavioural Sciences*. 2nd ed. New York: Lawrence Erlbaum Associates; 1988.

286. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, Reese SE, Markunas CA, Richmond RC, Xu C-J, Küpers LK, Oh SS, Hoyo C, Gruzieva O, Söderhäll C, Salas LA, Baïz N, Zhang H, Lepeule J, Ruiz C, Ligthart S, Wang T, Taylor JA, Duijts L, Sharp GC, Jankipersadsing SA, Nilsen RM, Vaez A, Fallin MD, Hu D, et al.: **DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis**. *The American Journal of Human Genetics* 2016, **98**:680–69610.1016/j.ajhg.2016.02.019Available: http://www.ncbi.nlm.nih.gov/pubmed/27040690.

287. Breton CV, Marsit CJ, Faustman E, Nadeau K, Goodrich JM, Dolinoy DC, Herbstman J, Holland N, LaSalle JM, Schmidt R, Yousefi P, Perera F, Joubert BR, Wiemels J, Taylor M,

Yang IV, Chen R, Hew KM, Freeland DMH, Miller R, Murphy SK: **Small-Magnitude Effect Sizes in Epigenetic End Points are Important in Children's Environmental Health Studies: The Children's Environmental Health and Disease Prevention Research Center's Epigenetics Working Group**. *Environmental Health Perspectives* 2017, **125**:511–52610.1289/EHP595Available: https://ehp.niehs.nih.gov/doi/10.1289/EHP595.

288. Partridge L, Deelen J, Slagboom PE: **Facing up to the global challenges of ageing**. *Nature* 2018, **561**:45–5610.1038/s41586-018-0457-8Available: http://www.nature.com/articles/s41586-018-0457-8.

289. Campisi J, Kapahi P, Lithgow GJ, Melov S, Newman JC, Verdin E: **From discoveries in ageing research to therapeutics for healthy ageing**. *Nature* 2019, **571**:183–19210.1038/s41586-019-1365-2Available: http://www.nature.com/articles/s41586-019-1365-2.

290. Chuong EB, Elde NC, Feschotte C: **Regulatory activities of transposable elements: from conflicts to benefits**. *Nature Reviews Genetics* 2017, **18**:71–8610.1038/nrg.2016.139Available: http://www.nature.com/articles/nrg.2016.139.

291. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, Esteller M: **Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome**. *Epigenetics* 2011, **6**:692–70210.4161/epi.6.6.16196Available: http://www.tandfonline.com/doi/abs/10.4161/epi.6.6.16196.

292. Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, Felton S, Matsuyama M, Lowe D, Kabacik S, Wilson JG, Reiner AP, Maierhofer A, Flunkert J, Aviv A, Hou L, Baccarelli AA, Li Y, Stewart JD, Whitsel EA, Ferrucci L, Matsuyama S, Raj K: **Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies**. *Aging* 2018, **10**:1758–177510.18632/aging.101508Available: http://www.aging-us.com/article/101508/text.

293. Horvath S, Raj K: **DNA methylation-based biomarkers and the epigenetic clock theory of ageing**. *Nature Reviews Genetics* 2018, **19**:371–38410.1038/s41576-018-0004-3Available: http://www.nature.com/articles/s41576-018-0004-3.

294. Field AE, Robertson NA, Wang T, Havas A, Ideker T, Adams PD: **DNA Methylation Clocks in Aging: Categories, Causes, and Consequences**. *Molecular Cell* 2018, **71**:882–89510.1016/j.molcel.2018.08.008Available: https://doi.org/10.1016/j.molcel.2018.08.008.

295. Bhattacharyya S, Varshney U: **Evolution of initiator tRNAs and selection of methionine as the initiating amino acid**. *RNA Biology* 2016, **13**:810–81910.1080/15476286.2016.1195943Available: http://dx.doi.org/10.1080/15476286.2016.1195943.

296. Kamhi E, Raitskin O, Sperling R, Sperling J: **A potential role for initiator-tRNA**

in pre-mRNA splicing regulation. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**:11319–1132410.1073/pnas.0911561107.

297. Birch J, Clarke CJ, Campbell AD, Campbell K, Mitchell L, Liko D, Kalna G, Strathdee D, Sansom OJ, Neilson M, Blyth K, Norman JC: **The initiator methionine tRNA drives cell migration and invasion leading to increased metastatic potential in melanoma**. *Biology Open* 2016, **5**:1371–137910.1242/bio.019075.

298. Rideout EJ, Marshall L, Grewal SS: **Drosophila RNA polymerase III repressor Maf1 controls body size and developmental timing by modulating tRNAiMet synthesis and systemic insulin signaling**. *Proceedings of the National Academy of Sciences* 2012, **109**:1139–114410.1073/pnas.1113311109Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1113311109.

299. Kolitz SE, Lorsch JR: **Eukaryotic initiator tRNA: Finely tuned and ready for action**. *FEBS Letters* 2010, **584**:396–40410.1016/j.febslet.2009.11.047Available: http://doi.wiley.com/10.1016/j.febslet.2009.11.047.

300. Pavon-Eternod M, Gomes S, Rosner MR, Pan T: **Overexpression of initiator methionine tRNA leads to global reprogramming of tRNA expression and increased proliferation in human epithelial cells**. *RNA* 2013, **19**:461–46610.1261/rna.037507.112Available: http://www.ncbi.nlm.nih.gov/pubmed/23431330.

301. Eigen M, Lindemann B, Tietze M, Winkler-Oswatitsch R, Dress A, Haeseler A von: **How old is the genetic code? Statistical geometry of tRNA provides an answer**. *Science* 1989, **244**:673–67910.1126/science.2497522Available: http://www.sciencemag.org/cgi/doi/10.1126/science.2497522.

302. Tavernarakis N: **Ageing and the regulation of protein synthesis: a balancing act?** *Trends in Cell Biology* 2008, **18**:228–23510.1016/j.tcb.2008.02.004Available: https://linkinghub.elsevier.com/retrieve/pii/S0962892408000792.

303. Parisien M, Wang X, Pan T: **Diversity of human tRNA genes from the 1000-genomes project**. *RNA Biology* 2013, **10**:1853–186710.4161/rna.27361Available: http://www.tandfonline.com/doi/full/10.4161/rna.27361.

304. Chan PP, Lowe TM: **GtRNAdb: a database of transfer RNA genes detected in genomic sequence.** *Nucleic acids research* 2009, **37**:D93–710.1093/nar/gkn787Available: http://www.ncbi.nlm.nih.gov/pubmed/18984615.

305. Torres AG: **Enjoy the Silence: Nearly Half of Human tRNA Genes Are Silent**. *Bioinformatics and Biology Insights* 2019, **13**:11779322198684510.1177/1177932219868454Available: http://journals.sagepub.com/doi/10.1177/1177932219868454.

306. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J: *Molecular Cell*

Biology, 4th edition. 4th ed. New York: W. H. Freeman; 2000 Available: https://www.ncbi.nlm.nih.gov/books/NBK21475/.

307. Schramm L: **Recruitment of RNA polymerase III to its target promoters**. *Genes & Development* 2002, **16**:2593–262010.1101/gad.1018902Available: http://www.genesdev.org/cgi/doi/10.1101/gad.1018902.

308. Canella D, Praz V, Reina JH, Cousin P, Hernandez N: **Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells.** *Genome research* 2010, **20**:710–2110.1101/gr.101337.109Available: http://www.ncbi.nlm.nih.gov/pubmed/20413673.

309. Dieci G, Sentenac A: **Facilitated Recycling Pathway for RNA Polymerase III**. *Cell* 1996, **84**:245–25210.1016/S0092-8674(00)80979-4Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867400809794.

310. Murawski M, Szczesniak B, Zoladek T, Hopper AK, Martin NC, Boguta M: **maf1 mutation alters the subcellular localization of the Mod5 protein in yeast.** *Acta biochimica Polonica* 1994, **41**:441–8Available: http://www.ncbi.nlm.nih.gov/pubmed/7732762.

311. Pluta K, Lefebvre O, Martin NC, Smagowicz WJ, Stanford DR, Ellis SR, Hopper AK, Sentenac A, Boguta M: **Maf1p, a Negative Effector of RNA Polymerase III in Saccharomyces cerevisiae**. *Molecular and Cellular Biology* 2001, **21**:5031–504010.1128/MCB.21.15.5031-5040.2001Available: http://www.ncbi.nlm.nih.gov/pubmed/11438659.

312. Vorländer MK, Baudin F, Moir RD, Wetzel R, Hagen WJH, Willis IM, Müller CW: **Structural basis for RNA polymerase III transcription repression by Maf1**. *Nature Structural & Molecular Biology* 2020, 10.1038/s41594-020-0383-yAvailable: http://www.nature.com/articles/s41594-020-0383-y.

313. Mange F, Praz V, Migliavacca E, Willis IM, Schütz F, Hernandez N: **Diurnal regulation of RNA polymerase III transcription is under the control of both the feeding–fasting response and the circadian clock**. *Genome Research* 2017, **27**:973–98410.1101/gr.217521.116Available: http://genome.cshlp.org/lookup/doi/10.1101/gr.217521.116.

314. Kennedy BK, Lamming DW: **The Mechanistic Target of Rapamycin: The Grand ConducTOR of Metabolism and Aging**. *Cell Metabolism* 2016, **23**:990–100310.1016/j.cmet.2016.05.009Available: http://dx.doi.org/10.1016/j.cmet.2016.05.009.

315. Crighton D: **p53 represses RNA polymerase III transcription by targeting TBP and inhibiting promoter occupancy by TFIIIB**. *The EMBO Journal* 2003, **22**:2810–282010.1093/emboj/cdg265Available: http://emboj.embopress.org/cgi/doi/10.1093/emboj/cd

g265.

316. Sutcliffe JE, Brown TRP, Allison SJ, Scott PH, White RJ: **Retinoblastoma Protein Disrupts Interactions Required for RNA Polymerase III Transcription**. *Molecular and Cellular Biology* 2000, **20**:9192–920210.1128/MCB.20.24.9192-9202.2000Available: http://mcb.asm.org/cgi/doi/10.1128/MCB.20.24.9192-9202.2000.

317. Gomez-Roman N, Grandori C, Eisenman RN, White RJ: **Direct activation of RNA polymerase III transcription by c-Myc**. *Nature* 2003, **421**:290–29410.1038/nature01327Available: http://www.nature.com/articles/nature01327.

318. Krishnan P, Ghosh S, Wang B, Heyns M, Li D, Mackey JR, Kovalchuk O, Damaraju S: **Genome-wide profiling of transfer RNAs and their role as novel prognostic markers for breast cancer**. *Scientific Reports* 2016, **6**:3284310.1038/srep32843Available: http://dx.doi.org/10.1038/srep32843.

319. Huang S, Sun B, Xiong Z, Shu Y, Zhou H, Zhang W, Xiong J, Li Q: **The dysregulation of tRNAs and tRNA derivatives in cancer**. *Journal of Experimental & Clinical Cancer Research* 2018, **37**:10110.1186/s13046-018-0745-zAvailable: https://jeccr.biomedcentral.com/articles/10.1186/s13046-018-0745-z.

320. Besser D, Götz F, Schulze-Forster K, Wagner H, Kröger H, Simon D: **DNA methylation inhibits transcription by RNA polymerase III of a tRNA gene, but not of a 5S rRNA gene.** *FEBS letters* 1990, **269**:358–6210.1016/0014-5793(90)81193-RAvailable: http://www.ncbi.nlm.nih.gov/pubmed/2401361.

321. Varshney D, Vavrova-Anderson J, Oler AJ, Cowling VH, Cairns BR, White RJ: **SINE transcription by RNA polymerase III is suppressed by histone methylation but not by DNA methylation.** *Nature communications* 2015, **6**:656910.1038/ncomms7569Available: http://www.ncbi.nlm.nih.gov/pubmed/25798578.

322. Gerber A, Ito K, Chu C-S, Roeder RG: **Gene-Specific Control of tRNA Expression by RNA Polymerase II**. *Molecular Cell* 2020, **78**:765–778.e710.1016/j.molcel.2020.03.023Available: https://doi.org/10.1016/j.molcel.2020.03.023.

323. Gingold H, Dahan O, Pilpel Y: **Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome.** *Nucleic acids research* 2012, **40**:10053–6310.1093/nar/gks772Available: http://www.ncbi.nlm.nih.gov/pubmed/22941644.

324. Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, Simonyan V, Kimchi-Sarfaty C: **A new and updated resource for codon usage tables**. *BMC Bioinformatics* 2017, **18**:39110.1186/s12859-017-1793-7Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1793-7.

325. Dittmar KA, Goodenbour JM, Pan T: **Tissue-specific differences in human transfer**

RNA expression. *PLoS genetics* 2006, **2**:e22110.1371/journal.pgen.0020221Available: http://www.ncbi.nlm.nih.gov/pubmed/17194224.

326. Sagi D, Rak R, Gingold H, Adir I, Maayan G, Dahan O, Broday L, Pilpel Y, Rechavi O: **Tissue- and Time-Specific Expression of Otherwise Identical tRNA Genes**. *PLOS Genetics* 2016, **12**:e100626410.1371/journal.pgen.1006264Available: http://www.ncbi.nlm.nih.gov/pubmed/27560950.

327. Kirchner S, Cai Z, Rauscher R, Kastelic N, Anding M, Czech A, Kleizen B, Ostedgaard LS, Braakman I, Sheppard DN, Ignatova Z: **Alteration of protein function by a silent polymorphism linked to tRNA abundance**. *PLOS Biology* 2017, **15**:e200077910.1371/journal.pbio.2000779Available: http://dx.plos.org/10.1371/journal.pbio.2000779.

328. Ishimura R, Nagy G, Dotu I, Zhou H, Yang X-L, Schimmel P, Senju S, Nishimura Y, Chuang JH, Ackerman SL: **RNA function. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration.** *Science (New York, N.Y.)* 2014, **345**:455–910.1126/science.1249749Available: http://www.ncbi.nlm.nih.gov/pubmed/25061210.

329. Pliatsika V, Loher P, Magee R, Telonis AG, Londin E, Shigematsu M, Kirino Y, Rigoutsos I: **MINTbase v2.0: a comprehensive database for tRNA-derived fragments that includes nuclear and mitochondrial fragments from all The Cancer Genome Atlas projects**. *Nucleic Acids Research* 2018, **46**:D152–D15910.1093/nar/gkx1075Available: http://academic.oup.com/nar/article/46/D1/D152/4653530.

330. Lee YS, Shibata Y, Malhotra A, Dutta A: **A novel class of small RNAs: tRNA-derived RNA fragments (tRFs).** *Genes & development* 2009, **23**:2639–4910.1101/gad.1837609Available: http://www.ncbi.nlm.nih.gov/pubmed/19933153.

331. Torres AG, Reina O, Stephan-Otto Attolini C, Ribas de Pouplana L: **Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments**. *Proceedings of the National Academy of Sciences* 2019, **116**:20182112010.1073/pnas.1821120116Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1821120116.

332. Li S, Xu Z, Sheng J: **tRNA-Derived Small RNA: A Novel Regulatory Small Non-Coding RNA**. *Genes* 2018, **9**:24610.3390/genes9050246Available: http://www.mdpi.com/2073-4425/9/5/246.

333. Xu W-L, Yang Y, Wang Y-D, Qu L-H, Zheng L-L: **Computational Approaches to tRNA-Derived Small RNAs**. *Non-Coding RNA* 2017, **3**:210.3390/ncrna3010002Available: http://www.mdpi.com/2311-553X/3/1/2.

334. Keam SP, Young PE, McCorkindale AL, Dang THY, Clancy JL, Humphreys DT, Preiss T, Hutvagner G, Martin DIK, Cropley JE, Suter CM: **The human Piwi protein**

Hiwi2 associates with tRNA-derived piRNAs in somatic cells. *Nucleic Acids Research* 2014, **42**:8984–899510.1093/nar/gku620Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku620.

335. Honda S, Kawamura T, Loher P, Morichika K, Rigoutsos I, Kirino Y: **The biogenesis pathway of tRNA-derived piRNAs in Bombyx germ cells.** *Nucleic Acids Research* 2017, **45**:9108–912010.1093/nar/gkx537Available: http://academic.oup.com/nar/article/45/15/9108/3883741/The-biogenesis-pathway-of-tRNAderived-piRNAs-in.

336. Tosar JP, Rovira C, Cayota A: **Non-coding RNA fragments account for the majority of annotated piRNAs expressed in somatic non-gonadal tissues.** *Communications Biology* 2018, **1**:210.1038/s42003-017-0001-7Available: http://www.nature.com/articles/s42003-017-0001-7.

337. Loher P, Telonis AG, Rigoutsos I: **MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data.** *Scientific reports* 2017, **7**:4118410.1038/srep41184Available: http://www.ncbi.nlm.nih.gov/pubmed/28220888.

338. Kim HK, Fuchs G, Wang S, Wei W, Zhang Y, Park H, Roy-Chaudhuri B, Li P, Xu J, Chu K, Zhang F, Chua M-S, So S, Zhang QC, Sarnow P, Kay MA: **A transfer-RNA-derived small RNA regulates ribosome biogenesis.** *Nature* 2017, **552**:57–6210.1038/nature25005Available: http://www.nature.com/articles/nature25005.

339. Martinez G, Choudury SG, Slotkin RK: **tRNA-derived small RNAs target transposable element transcripts.** *Nucleic Acids Research* 2017, **45**:5142–515210.1093/nar/gkx103Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx103.

340. Schimmel P: **The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis.** *Nature reviews. Molecular cell biology* 2018, **19**:45–510.1038/nrm.2017.77Available: http://www.ncbi.nlm.nih.gov/pubmed/28875994.

341. Cristodero M, Polacek N: **The multifaceted regulatory potential of tRNA-derived fragments.** *Non-coding RNA Investigation* 2017, **1**:7–710.21037/ncri.2017.08.07Available: http://ncri.amegroups.com/article/view/3820/4459.

342. Pace DA: **Fixed metabolic costs for highly variable rates of protein synthesis in sea urchin embryos and larvae.** *Journal of Experimental Biology* 2006, **209**:158–17010.1242/jeb.01962Available: http://jeb.biologists.org/cgi/doi/10.1242/jeb.01962.

343. Nwanaji-Enwerem JC, Weisskopf MG, Baccarelli AA: **Multi-tissue DNA methylation age: Molecular relationships and perspectives for advancing biomarker utility.** *Ageing Research Reviews* 2018, **45**:15–2310.1016/j.arr.2018.04.005Available: https://doi.org/10.101

6/j.arr.2018.04.005.

344. Hansen M, Taubert S, Crawford D, Libina N, Lee S-J, Kenyon C: **Lifespan extension by conditions that inhibit translation in Caenorhabditis elegans.** *Aging Cell* 2007, **6**:95–11010.1111/j.1474-9726.2006.00267.xAvailable: http://doi.wiley.com/10.1111/j.1474-9726.2006.00267.x.

345. Filer D, Thompson MA, Takhaveev V, Dobson AJ, Kotronaki I, Green JWM, Heinemann M, Tullet JMA, Alic N: **RNA polymerase III limits longevity downstream of TORC1.** *Nature* 2017, **552**:263–26710.1038/nature25007Available: http://www.nature.com/articles/nature25007.

346. Dhahbi JM, Spindler SR, Atamna H, Yamakawa A, Boffelli D, Mote P, Martin DI: **5' tRNA halves are present as abundant complexes in serum, concentrated in blood cells, and modulated by aging and calorie restriction.** *BMC Genomics* 2013, **14**:29810.1186/1471-2164-14-298Available: http://www.ncbi.nlm.nih.gov/pubmed/23638709.

347. Bhargava P: **Epigenetic regulation of transcription by RNA polymerase III.** *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 2013, **1829**:1015–102510.1016/j.bbagrm.2013.05.005Available: https://linkinghub.elsevier.com/retrieve/pii/S1874939913001004.

348. Bhargava P: **Epigenetic regulation of transcription by RNA polymerase III.** *Biochimica et biophysica acta* 2013, **1829**:1015–2510.1016/j.bbagrm.2013.05.005Available: http://www.ncbi.nlm.nih.gov/pubmed/23732820.

349. Wilusz JE: **Controlling translation via modulation of tRNA levels.** *Wiley Interdisciplinary Reviews: RNA* 2015, **6**:453–47010.1002/wrna.1287Available: http://doi.wiley.com/10.1002/wrna.1287.

350. Cozen AE, Quartley E, Holmes AD, Hrabeta-Robinson E, Phizicky EM, Lowe TM: **ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments.** *Nature Methods* 2015, **12**:879–88410.1038/nmeth.3508Available: http://www.nature.com/articles/nmeth.3508.

351. Shigematsu M, Honda S, Loher P, Telonis AG, Rigoutsos I, Kirino Y: **YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs.** *Nucleic acids research* 2017, **45**:e7010.1093/nar/gkx005Available: http://www.ncbi.nlm.nih.gov/pubmed/28108659.

352. Gogakos T, Brown M, Garzia A, Meyer C, Hafner M, Tuschl T: **Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP.** *Cell Reports* 2017, **20**:1463–147510.1016/j.celrep.2017.07.029Available: http://dx.doi.org/10.1016/j.celrep.2017.07.029.

353. Raab JR, Chiu J, Zhu J, Katzman S, Kurukuti S, Wade PA, Haussler D, Kamakaka RT: **Human tRNA genes function as chromatin insulators.** *The EMBO journal* 2012, **31**:330–5010.1038/emboj.2011.406Available: http://www.ncbi.nlm.nih.gov/pubmed/22085927.

354. Van Bortle K, Phanstiel DH, Snyder MP: **Topological organization and dynamic regulation of human tRNA genes during macrophage differentiation.** *Genome Biology* 2017, **18**:18010.1186/s13059-017-1310-3Available: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1310-3.

355. Noma K, Cam HP, Maraia RJ, Grewal SIS: **A role for TFIIIC transcription factor complex in genome organization.** *Cell* 2006, **125**:859–7210.1016/j.cell.2006.04.028Available: http://www.ncbi.nlm.nih.gov/pubmed/16751097.

356. Syddall H, Aihie Sayer A, Dennison E, Martin H, Barker D, Cooper C: **Cohort Profile: The Hertfordshire Cohort Study.** *International Journal of Epidemiology* 2005, **34**:1234–124210.1093/ije/dyi127Available: http://academic.oup.com/ije/article/34/6/1234/707357/Cohort-Profile-The-Hertfordshire-Cohort-Study.

357. Amemiya HM, Kundaje A, Boyle AP: **The ENCODE Blacklist: Identification of Problematic Regions of the Genome.** *Scientific Reports* 2019, **9**:935410.1038/s41598-019-45839-zAvailable: http://www.nature.com/articles/s41598-019-45839-z.

358. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P: **Fast Computation and Applications of Genome Mappability.** *PLoS ONE* 2012, **7**:e3037710.1371/journal.pone.0030377Available: https://dx.plos.org/10.1371/journal.pone.0030377.

359. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473**:43–4910.1038/nature09906Available: http://www.nature.com/articles/nature09906.

360. Bell CG, Wilson GA, Butcher LM, Roos C, Walter L, Beck S: **Human-specific CpG "beacons" identify loci associated with human-specific traits and disease.** *Epigenetics* 2012, **7**:1188–9910.4161/epi.22127Available: http://www.ncbi.nlm.nih.gov/pubmed/22968434.

361. North BV, Curtis D, Sham PC: **A Note on the Calculation of Empirical P Values from Monte Carlo Procedures.** *The American Journal of Human Genetics* 2003, **72**:498–49910.1086/346173Available: https://linkinghub.elsevier.com/retrieve/pii/S0002929707605618.

362. Ewels P, Magnusson M, Lundin S, Käller M: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics (Oxford, England)* 2016, **32**:3047–810.1093/bioinformatics/btw354Available: http://www.ncbi.nlm.nih.gov/pubmed/27312411.

363. Krueger F: **Trim Galore**. 2015, Available: https://www.bioinformatics.babraham.ac.uk/projects/trim%7B/_%7Dgalore/.

364. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads**. *EMBnet.journal* 2011, **17**:1010.14806/ej.17.1.200Available: http://journal.embnet.org/index.php/embnetjournal/article/view/200.

365. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**:357–35910.1038/nmeth.1923Available: http://www.ncbi.nlm.nih.gov/pubmed/22388286.

366. Müller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, Bock C: **RnBeads 2.0: comprehensive analysis of DNA methylation data**. *Genome Biology* 2019, **20**:5510.1186/s13059-019-1664-9Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1664-9.

367. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Research* 2015, **43**:e47–e4710.1093/nar/gkv007Available: http://www.ncbi.nlm.nih.gov/pubmed/25605792.

368. Meuleman W: **Epilogos**. 2019, Available: https://epilogos.altius.org/%20https://github.com/Altius/epilogos.

369. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, Söderhäll C, Scheynius A, Kere J: **Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility**. *PLoS ONE* 2012, **7**:e4136110.1371/journal.pone.0041361Available: https://dx.plos.org/10.1371/journal.pone.0041361.

370. Sean D, Meltzer PS: **GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor**. *Bioinformatics* 2007, **23**:1846–184710.1093/bioinformatics/btm254.

371. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM: **Toward a Shared Vision for Cancer Genomic Data**. *New England Journal of Medicine* 2016, **375**:1109–111210.1056/NEJMp1607591Available: http://www.nejm.org/doi/10.1056/NEJMp1607591.

372. Yang Z, Wong A, Kuh D, Paul DS, Rakyan VK, Leslie RD, Zheng SC, Widschwendter M, Beck S, Teschendorff AE: **Correlation of an epigenetic mitotic clock with cancer risk**. *Genome Biology* 2016, **17**:20510.1186/s13059-016-1064-3Available: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1064-3.

373. Nazor KL, Altun G, Lynch C, Tran H, Harness JV, Slavin I, Garitaonandia I, Müller F-J, Wang Y-C, Boscolo FS, Fakunle E, Dumevska B, Lee S, Park HS, Olee T, D'Lima

DD, Semechkin R, Parast MM, Galat V, Laslett AL, Schmidt U, Keirstead HS, Loring JF, Laurent LC: **Recurrent Variations in DNA Methylation in Human Pluripotent Stem Cells and Their Differentiated Derivatives**. *Cell Stem Cell* 2012, **10**:620–63410.1016/j.stem.2012.02.013Available: http://www.ncbi.nlm.nih.gov/pubmed/10449618.

374. Juzenas S, Venkatesh G, Hübenthal M, Hoeppner MP, Du ZG, Paulsen M, Rosenstiel P, Senger P, Hofmann-Apitius M, Keller A, Kupcinskas L, Franke A, Hemmrich-Stanisak G: **A comprehensive, cell specific microRNA catalogue of human peripheral blood**. *Nucleic Acids Research* 2017, **45**:9290–93010.1093/nar/gkx706Available: http://academic.oup.com/nar/article/45/16/9290/4080663.

375. Petkovich DA, Podolskiy DI, Lobanov AV, Lee S-G, Miller RA, Gladyshev VN: **Using DNA Methylation Profiling to Evaluate Biological Age and Longevity Interventions**. *Cell Metabolism* 2017, **25**:954–960.e610.1016/j.cmet.2017.03.016Available: https://linkinghub.elsevier.com/retrieve/pii/S1550413117301687.

376. Lowe TM, Chan PP: **tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes**. *Nucleic Acids Research* 2016, **44**:W54–W5710.1093/nar/gkw413Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw413.

377. Geiger H, Haan G de, Florian MC: **The ageing haematopoietic stem cell compartment**. *Nature Reviews Immunology* 2013, **13**:376–38910.1038/nri3433Available: http://dx.doi.org/10.1038/nri3433.

378. Schmitt BM, Rudolph KLM, Karagianni P, Fonseca NA, White RJ, Talianidis I, Odom DT, Marioni JC, Kutter C: **High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA–tRNA interface**. *Genome Research* 2014, **24**:1797–18071O.1101/gr.176784.114Available: http://www.ncbi.nlm.nih.gov/pubmed/25122613.

379. Gu Z, Eils R, Schlesner M: **Complex heatmaps reveal patterns and correlations in multidimensional genomic data**. *Bioinformatics* 2016, **32**:2847–28491O.1093/bioinformatics/btw313Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw313.

380. Thornlow BP, Armstrong J, Holmes AD, Howard JM, Corbett-Detig RB, Lowe TM: **Predicting transfer RNA gene activity from sequence and genome context**. *Genome Research* 2020, **30**:85–9410.1101/gr.256164.119.

381. Ehrlich M: **DNA hypermethylation in disease: mechanisms and clinical relevance.** *Epigenetics* 2019, **14**:1141–116310.1080/15592294.2019.1638701Available: http://www.ncbi.nlm.nih.gov/pubmed/31284823.

382. Xu Z, Taylor JA: **Genome-wide age-related DNA methylation changes in blood**

and other tissues relate to histone modification, expression and cancer. *Carcinogenesis* 2014, **35**:356–36410.1093/carcin/bgt391Available: https://academic.oup.com/carcin/article-lookup/doi/10.1093/carcin/bgt391.

383. 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 and Chromatin* 2018, **11**:1–1110.1186/s13072-018-0191-3Available: https://doi.org/10.1186/s13072-018-0191-3.

384. Zhu T, Zheng SC, Paul DS, Horvath S, Teschendorff AE: **Cell and tissue type independent age-associated DNA methylation changes are not rare but common.** *Aging* 2018, **10**:3541–35510.18632/aging.101666Available: http://www.ncbi.nlm.nih.gov/pubmed/30482885.

385. Goodenbour JM, Pan T: **Diversity of tRNA genes in eukaryotes.** *Nucleic acids research* 2006, **34**:6137–4610.1093/nar/gkl725Available: http://www.ncbi.nlm.nih.gov/pubmed/17088292.

386. Geslain R, Pan T: **Functional Analysis of Human tRNA Isodecoders**. *Journal of Molecular Biology* 2010, **396**:821–83110.1016/j.jmb.2009.12.018Available: https://linkinghub.elsevier.com/retrieve/pii/S002228360901523X.

387. Li S, Shi X, Chen M, Xu N, Sun D, Bai R, Chen H, Ding K, Sheng J, Xu Z: **Angiogenin promotes colorectal cancer metastasis via tiRNA production**. *International Journal of Cancer* 2019, :ijc.3224510.1002/ijc.32245Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.32245.

388. Plotkin JB, Robins H, Levine AJ: **Tissue-specific codon usage and the expression of human genes.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**:12588–9110.1073/pnas.0404957101Available: http://www.ncbi.nlm.nih.gov/pubmed/15314228.

389. Schmitt BM, Rudolph KLM, Karagianni P, Fonseca NA, White RJ, Talianidis I, Odom DT, Marioni JC, Kutter C: **High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface.** *Genome research* 2014, **24**:1797–80710.1101/gr.176784.114Available: http://www.ncbi.nlm.nih.gov/pubmed/25122613.

390. Powell JR, Moriyama EN: **Evolution of codon usage bias in Drosophila**. *Proceedings of the National Academy of Sciences* 1997, **94**:7784–779010.1073/pnas.94.15.7784Available: http://www.pnas.org/cgi/doi/10.1073/pnas.94.15.7784.

391. Rudolph KLM, Schmitt BM, Villar D, White RJ, Marioni JC, Kutter C, Odom DT: **Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States.** *PLoS genetics* 2016, **12**:e100602410.1371/journal.pgen.1006024Available: http://www.

ncbi.nlm.nih.gov/pubmed/27166679.

392. Gu T, Lin X, Cullen SM, Luo M, Jeong M, Estecio M, Shen J, Hardikar S, Sun D, Su J, Rux D, Guzman A, Lee M, Qi LS, Chen J-J, Kyba M, Huang Y, Chen T, Li W, Goodell MA: **DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells**. *Genome Biology* 2018, **19**:8810.1186/s13059-018-1464-7Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1464-7.

393. Wang Y, Brady KS, Caiello BP, Ackerson SM, Stewart JA: **Human CST suppresses origin licensing and promotes AND-1/Ctf4 chromatin association**. *Life Science Alliance* 2019, **2**:e20180027010.26508/lsa.201800270Available: http://www.life-science-alliance.org/lookup/doi/10.26508/lsa.201800270.

394. Sargolzaeiaval F, Zhang J, Schleit J, Lessel D, Kubisch C, Precioso DR, Sillence D, Hisama FM, Dorschner M, Martin GM, Oshima J: **CTC1 mutations in a Brazilian family with progeroid features and recurrent bone fractures**. *Molecular Genetics & Genomic Medicine* 2018, **6**:1148–115610.1002/mgg3.495Available: http://doi.wiley.com/10.1002/mgg3.495.

395. Thornlow BP, Hough J, Roger JM, Gong H, Lowe TM, Corbett-Detig RB: **Transfer RNA genes experience exceptionally elevated mutation rates**. *Proceedings of the National Academy of Sciences* 2018, **115**:8996–900110.1073/pnas.1801240115Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1801240115.

396. Iben JR, Maraia RJ: **tRNA gene copy number variation in humans**. *Gene* 2014, **536**:376–38410.1016/j.gene.2013.11.049Available: https://linkinghub.elsevier.com/retrieve/pii/S0378111913015758.

397. Darrow EM, Chadwick BP: **A novel tRNA variable number tandem repeat at human chromosome 1q23.3 is implicated as a boundary element based on conservation of a CTCF motif in mouse**. *Nucleic Acids Research* 2014, **42**:6421–643510.1093/nar/gku280Available: http://academic.oup.com/nar/article/42/10/6421/2435926/A-novel-tRNA-variable-number-tandem-repeat-at.

398. Müller CA, Nieduszynski CA: **DNA replication timing influences gene expression level**. *The Journal of Cell Biology* 2017, **216**:1907–191410.1083/jcb.201701061Available: http://www.jcb.org/lookup/doi/10.1083/jcb.201701061.

399. Du Q, Bert SA, Armstrong NJ, Caldon CE, Song JZ, Nair SS, Gould CM, Luu P-L, Peters T, Khoury A, Qu W, Zotenko E, Stirzaker C, Clark SJ: **Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer**. *Nature Communications* 2019, **10**:41610.1038/s41467-019-08302-1Available: http://dx.doi.org/10.1038/s41467-019-08302-1.

400. Cruickshanks HA, McBryan T, Nelson DM, VanderKraats ND, Shah PP, Tuyn J van,

Singh Rai T, Brock C, Donahue G, Dunican DS, Drotar ME, Meehan RR, Edwards JR, Berger SL, Adams PD: **Senescent cells harbour features of the cancer epigenome**. *Nature Cell Biology* 2013, **15**:1495–150610.1038/ncb2879Available: http://www.ncbi.nlm.nih.gov/pubmed/24270890.

401. Sun L, Yu R, Dang W: **Chromatin Architectural Changes during Cellular Senescence and Aging**. *Genes* 2018, **9**:21110.3390/genes9040211Available: http://www.mdpi.com/2073-4425/9/4/211.

402. Moskowitz DM, Zhang DW, Hu B, Le Saux S, Yanes RE, Ye Z, Buenrostro JD, Weyand CM, Greenleaf WJ, Goronzy JJ: **Epigenomics of human CD8 T cell differentiation and aging**. *Science Immunology* 2017, **2**:eaag019210.1126/sciimmunol.aag0192Available: https://immunology.sciencemag.org/lookup/doi/10.1126/sciimmunol.aag0192.

403. Hashimoto K, Kouno T, Ikawa T, Hayatsu N, Miyajima Y, Yabukami H, Terooatea T, Sasaki T, Suzuki T, Valentine M, Pascarella G, Okazaki Y, Suzuki H, Shin JW, Minoda A, Taniuchi I, Okano H, Arai Y, Hirose N, Carninci P: **Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians**. *Proceedings of the National Academy of Sciences of the United States of America* 2019, **116**:24242–2425110.1073/pnas.1907883116.

404. Deniz Ö, Frost JM, Branco MR: **Regulation of transposable elements by DNA modifications**. *Nature Reviews Genetics* 2019, **20**:417–43110.1038/s41576-019-0106-6.

405. Kazazian HH, Moran JV: **Mobile DNA in health and disease**. *New England Journal of Medicine* 2017, **377**:361–37010.1056/NEJMra1510092.

406. Gregory TR: **Synergy between sequence and size in large-scale genomics**. *Nature Reviews Genetics* 2005, **6**:699–70810.1038/nrg1674.

407. Deaton AM, Bird A: **CpG islands and the regulation of transcription**. *Genes & development* 2011, **25**:1010–2210.1101/gad.2037511Available: http://www.ncbi.nlm.nih.gov/pubmed/21576262.

408. Hellman A, Chess A: **Gene body-specific methylation on the active X chromosome**. *Science* 2007, **315**:1141–114310.1126/science.1136352.

409. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S: **CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing**. *Nature* 2011, **479**:74–7910.1038/nature10442Available: http://dx.doi.org/10.1038/nature10442.

410. Jordà M, Díez-Villanueva A, Mallona I, Martín B, Lois S, Barrera V, Esteller M, Vavouri T, Peinado MA: **The epigenetic landscape of Alu repeats delineates the structural and functional genomic architecture of colon cancer cells**. *Genome Research* 2017, **27**:118–13210.1101/gr.207522.116Available: http://www.ncbi.nlm.nih.gov/pubmed/27999094.

411. Bollati V, Schwartz J, Wright R, Litonjua A, Tarantini L, Suh H, Sparrow D, Vokonas P, Baccarelli A: **Decline in genomic DNA methylation through aging in a cohort of elderly subjects**. *Mechanisms of Ageing and Development* 2009, **130**:234–23910.1016/j.mad.2008.12.003Available: https://linkinghub.elsevier.com/retrieve/pii/S0047637408002820.

412. Dewannieux M, Esnault C, Heidmann T: **LINE-mediated retrotransposition of marked Alu sequences**. *Nature Genetics* 2003, **35**:41–4810.1038/ng1223Available: http://www.nature.com/articles/ng1223.

413. Deininger P: **Alu elements: know the SINEs**. *Genome Biology* 2011, **12**:23610.1186/gb-2011-12-12-236Available: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-12-236.

414. Ullu E, Tschudi C: **Alu sequences are processed 7SL RNA genes**. *Nature* 1984, **312**:171–17210.1038/312171a0Available: http://www.nature.com/articles/312171a0.

415. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution**. *Nature Reviews Genetics* 2009, **10**:691–70310.1038/nrg2640Available: http://www.nature.com/articles/nrg2640.

416. Quentin Y: **Origin of the Alu family:** *Nucleic Acids Research* 1992, **20**:3397–3401.

417. Chen L-L, Carmichael GG: **Gene regulation by SINES and inosines: biological consequences of A-to-I editing of Alu element inverted repeats**. *Cell Cycle* 2008, **7**:3294–33010.4161/cc.7.21.6927Available: http://www.tandfonline.com/doi/abs/10.4161/cc.7.21.6927.

418. Batzer MA, Deininger PL: **Alu repeats and human genomic diversity**. *Nature Reviews Genetics* 2002, **3**:370–37910.1038/nrg798Available: http://www.nature.com/articles/nrg798.

419. Ade C, Roy-Engel AM, Deininger PL: **Alu elements: an intrinsic source of human genome instability**. *Current Opinion in Virology* 2013, **3**:639–64510.1016/j.coviro.2013.09.002Available: https://linkinghub.elsevier.com/retrieve/pii/S1879625713001533.

420. Belancio VP, Hedges DJ, Deininger P: **Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health**. *Genome Research* 2008, **18**:343–35810.1101/gr.5558208Available: http://www.genome.org/cgi/doi/10.1101/gr.5558208.

421. Chen L-L, Yang L: **ALU ternative Regulation for Gene Expression**. *Trends in Cell Biology* 2017, **27**:480–49010.1016/j.tcb.2017.01.002Available: https://linkinghub.elsevier.com/retrieve/pii/S0962892417300028.

422. Ferrari R, de Llobet Cucalon LI, Di Vona C, Le Dilly F, Vidal E, Lioutas A, Oliete JQ, Jochem L, Cutts E, Dieci G, Vannini A, Teichmann M, Luna S de la, Beato M: **TFIIIC Binding to Alu Elements Controls Gene Expression via Chromatin Looping and**

Histone Acetylation. *Molecular Cell* 2019, :1–1310.1016/j.molcel.2019.10.020Available: https://linkinghub.elsevier.com/retrieve/pii/S1097276519307981.

423. Polak P, Domany E: **Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes.** *BMC genomics* 2006, **7**:13310.1186/1471-2164-7-133Available: http://www.ncbi.nlm.nih.gov/pubmed/16740159.

424. Su M, Han D, Boyd-Kirkup J, Yu X, Han J-DJ: **Evolution of Alu Elements toward Enhancers**. *Cell Reports* 2014, **7**:376–38510.1016/j.celrep.2014.03.011Available: http://dx.doi.org/10.1016/j.celrep.2014.03.011.

425. Blattler A, Yao L, Witt H, Guo Y, Nicolet CM, Berman BP, Farnham PJ: **Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes**. *Genome Biology* 2014, **15**:46910.1186/s13059-014-0469-0Available: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0469-0.

426. Ward MC, Wilson MD, Barbosa-Morais NL, Schmidt D, Stark R, Pan Q, Schwalie PC, Menon S, Lukk M, Watt S, Thybert D, Kutter C, Kirschner K, Flicek P, Blencowe BJ, Odom DT: **Latent Regulatory Potential of Human-Specific Repetitive Elements**. *Molecular Cell* 2013, **49**:262–27210.1016/j.molcel.2012.11.013Available: http://dx.doi.org/10.1016/j.molcel.2012.11.013.

427. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, Gascard P, Sigaroudinia M, Tlsty TD, Kadlecek T, Weiss A, O'Geen H, Farnham PJ, Madden PAF, Mungall AJ, Tam A, Kamoh B, Cho S, Moore R, Hirst M, Marra MA, Costello JF, Wang T: **DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape**. *Nature Genetics* 2013, **45**:836–84110.1038/ng.2649Available: http://www.ncbi.nlm.nih.gov/pubmed/23708189.

428. Zhang X-O, Gingeras TR, Weng Z: **Genome-wide analysis of polymerase III–transcribed Alu elements suggests cell-type–specific enhancer function**. *Genome Research* 2019, **29**:1402–141410.1101/gr.249789.119Available: http://genome.cshlp.org/lookup/doi/10.1101/gr.249789.119.

429. Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, Zhang D, Li D, Xing X, Kim S, O'Donnell D, Gordon JI, Wang T: **Transposable elements drive widespread expression of oncogenes in human cancers**. *Nature Genetics* 2019, **51**:611–61710.1038/s41588-019-0373-3Available: http://dx.doi.org/10.1038/s41588-019-0373-3.

430. Rajendiran S, Gibbs LD, Van Treuren T, Klinkebiel DL, Vishwanatha JK: **MIEN1 is tightly regulated by SINE Alu methylation in its promoter**. *Oncotarget* 2016, **7**:65307–6531910.18632/oncotarget.11675Available: http://www.oncotarget.com/fulltext/11675.

431. Jintaridth P, Tungtrongchitr R, Preutthipan S, Mutirangura A: **Hypomethylation**

of Alu Elements in Post-Menopausal Women with Osteoporosis. *PLoS ONE* 2013, **8**:e7038610.1371/journal.pone.0070386Available: https://dx.plos.org/10.1371/journal.pone.0070386.

432. Lu S, Niu Z, Chen Y, Tu Q, Zhang Y, Chen W, Tong W, Zhang Z: **Repetitive Element DNA Methylation is Associated with Menopausal Age**. *Aging and Disease* 2018, **9**:43510.14336/AD.2017.0810Available: http://www.aginganddisease.org/EN/10.14336/AD.2017.0810.

433. Payer LM, Steranka JP, Yang WR, Kryatova M, Medabalimi S, Ardeljan D, Liu C, Boeke JD, Avramopoulos D, Burns KH: **Structural variants caused by Alu insertions are associated with risks for many human diseases**. *Proceedings of the National Academy of Sciences* 2017, **114**:E3984–E399210.1073/pnas.1704117114Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1704117114.

434. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, Christensen BC, Gladyshev VN, Heijmans BT, Horvath S, Ideker T, Issa J-PJ, Kelsey KT, Marioni RE, Reik W, Relton CL, Schalkwyk LC, Teschendorff AE, Wagner W, Zhang K, Rakyan VK: **DNA methylation aging clocks: challenges and recommendations.** *Genome biology* 2019, **20**:24910.1186/s13059-019-1824-yAvailable: http://www.ncbi.nlm.nih.gov/pubmed/31767039.

435. Mozhui K, Pandey AK: **Conserved effect of aging on DNA methylation and association with EZH2 polycomb protein in mice and humans**. *Mechanisms of Ageing and Development* 2017, **162**:27–3710.1016/j.mad.2017.02.006Available: http://dx.doi.org/10.1016/j.mad.2017.02.006.

436. Zhang Y, Hapala J, Brenner H, Wagner W: **Individual CpG sites that are associated with age and life expectancy become hypomethylated upon aging**. *Clinical Epigenetics* 2017, **9**:910.1186/s13148-017-0315-9Available: http://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-017-0315-9.

437. Smit A, Hubley R, Green: **Repeat Masker**. 2009, Available: http://www.repeatmasker.org.

438. Privé F, Aschard H, Blum MGB: **Efficient Implementation of Penalized Regression for Genetic Risk Prediction**. *Genetics* 2019, **212**:65–7410.1534/genetics.119.302019Available: http://www.genetics.org/lookup/doi/10.1534/genetics.119.302019%20https://figshare.com/articles/code/7178750%20https://gsajournals.figshare.com/articles/Supplemental%7B/_%7DMaterial%7B/_%7Dfor%7B/_%7DPriv%7B/_%7DAschard%7B/_%7Dand%7B/_%7DBlum%7B/_%7D2019/7851470.

439. Fairley S, Lowy-Gallego E, Perry E, Flicek P: **The International Genome Sample Resource (IGSR) collection of open human genomic variation resources**. *Nucleic*

Acids Research 2020, **48**:D941–D94710.1093/nar/gkz836Available: https://academic.oup.com/nar/article/48/D1/D941/5580898.

440. Payer LM, Steranka JP, Ardeljan D, Walker J, Fitzgerald KC, Calabresi PA, Cooper TA, Burns KH: **Alu insertion variants alter mRNA splicing**. *Nucleic Acids Research* 2019, **47**:421–43110.1093/nar/gky1086.

441. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, Tibshirani RJ: **Strong rules for discarding predictors in lasso-type problems**. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 2012, **74**:245–26610.1111/j.1467-9868.2011.01004.xAvailable: http://arxiv.org/abs/1011.2234.

442. Snir S, Farrell C, Pellegrini M: **Human epigenetic ageing is logarithmic with time across the entire lifespan**. *Epigenetics* 2019, **0**:1–1510.1080/15592294.2019.1623634Available: https://www.tandfonline.com/doi/full/10.1080/15592294.2019.1623634.

443. Hellen EH, Brookfield JF: **Alu elements in primates are preferentially lost from areas of high GC content**. *PeerJ* 2013, **1**:e7810.7717/peerj.78Available: https://peerj.com/articles/78.

444. Meyer DH: **A transcriptome based aging clock near the theoretical limit of accuracy**. 2020, 10.1101/2020.05.29.123430Available: https://doi.org/10.1101/2020.05.29.123430.

445. Hysi PG, Young TL, Mackey DA, Andrew T, Fernández-Medarde A, Solouki AM, Hewitt AW, Macgregor S, Vingerling JR, Li Y, Ikram MK, Fai LY, Sham PC, Manyes L, Porteros A, Lopes MC, Carbonaro F, Fahy SJ, Martin NG, Duijn CM van, Spector TD, Rahi JS, Santos E, Klaver CCW, Hammond CJ: **A genome-wide association study for myopia and refractive error identifies a susceptibility locus at 15q25**. *Nature Genetics* 2010, **42**:902–90510.1038/ng.664Available: http://www.nature.com/articles/ng.664.

446. Loh P, Palamara PF, Price AL: **Fast and accurate long-range phasing in a UK Biobank cohort**. *Nature Genetics* 2016, **48**:811–81610.1038/ng.3571Available: http://www.nature.com/articles/ng.3571.

447. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh P-R, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C: **Next-generation genotype imputation service and methods**. *Nature Genetics* 2016, **48**:1284–128710.1038/ng.3656Available: http://www.nature.com/articles/ng.3656.

448. McCarthy S: **A reference panel of 64,976 haplotypes for genotype imputation**. *Nature Genetics* 2016, **48**:1279–128310.1038/ng.3643Available: http://www.nature.com/articles/ng.3643.

449. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association**

studies. *Nature Genetics* 2012, **44**:821–82410.1038/ng.2310Available: http://www.nature.com/articles/ng.2310.

450. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ: **LocusZoom: regional visualization of genome-wide association scan results**. *Bioinformatics* 2010, **26**:2336–233710.1093/bioinformatics/btq419Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq419.

451. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, Paul DS, Freitag D, Burgess S, Danesh J, Young R, Butterworth AS: **PhenoScanner: a database of human genotype–phenotype associations**. *Bioinformatics* 2016, **32**:3207–320910.1093/bioinformatics/btw373Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw373.

452. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, Paul DS, Freitag D, Burgess S, Danesh J, Young R, Butterworth AS: **PhenoScanner: a database of human genotype-phenotype associations.** *Bioinformatics (Oxford, England)* 2016, **32**:3207–320910.1093/bioinformatics/btw373Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5048068.

453. Gagliano Taliun SA, VandeHaar P, Boughton AP, Welch RP, Taliun D, Schmidt EM, Zhou W, Nielsen JB, Willer CJ, Lee S, Fritsche LG, Boehnke M, Abecasis GR: **Exploring and visualizing large-scale genetic associations by using PheWeb**. *Nature Genetics* 2020, **52**:550–5510.1038/s41588-020-0622-5.

454. Cancer Research UK: **Gallbladder cancer incidence statistics**. 2020, Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/gallbladder-cancer/incidence%7B/#%7Dheading-One.Accessed 17 September 2020.

455. Löhler J, Cebulla M, Shehata-Dieler W, Volkenstein S, Völter C, Walther LE: **Hearing impairment in old age**. *Deutsches Aerzteblatt Online* 2019, **116**:301–310.3238/arztebl.2019.0301Available: https://www.aerzteblatt.de/10.3238/arztebl.2019.0301.

456. Hileeto D, Fadare O, Martel M, Zheng W: **Age dependent association of endometrial polyps with increased risk of cancer involvement.** *World journal of surgical oncology* 2005, **3**:810.1186/1477-7819-3-8Available: http://www.ncbi.nlm.nih.gov/pubmed/15703068.

457. Abhishekh HA, Nisarga P, Kisan R, Meghana A, Chandran S, Trichur Raju, Sathyaprabha TN: **Influence of age and gender on autonomic regulation of heart**. *Journal of Clinical Monitoring and Computing* 2013, **27**:259–26410.1007/s10877-012-9424-3Available: http://link.springer.com/10.1007/s10877-012-9424-3.

458. Nikitin NP, Witte KKA, Ingle L, Clark AL, Farnsworth TA, Cleland JGF: **Longitudinal myocardial dysfunction in healthy older subjects as a manifestation of cardiac ageing**. *Age and Ageing* 2005, **34**:343–34910.1093/ageing/afi043Available: http://academic.oup.com/age

ing/article/34/4/343/10217/Longitudinal-myocardial-dysfunction-in-healthy.

459. Lu AT, Hannon E, Levine ME, Crimmins EM, Lunnon K, Mill J, Geschwind DH, Horvath S: **Genetic architecture of epigenetic and neuronal ageing rates in human brain regions**. *Nature Communications* 2017, **8**:1535310.1038/ncomms15353Available: http://dx.doi.org/10.1038/ncomms15353.

460. Gibson J, Russ TC, Clarke T-K, Howard DM, Hillary RF, Evans KL, Walker RM, Bermingham ML, Morris SW, Campbell A, Hayward C, Murray AD, Porteous DJ, Horvath S, Lu AT, McIntosh AM, Whalley HC, Marioni RE: **A meta-analysis of genome-wide association studies of epigenetic age acceleration**. *PLOS Genetics* 2019, **15**:e100810410.1371/journal.pgen.1008104Available: http://www.ncbi.nlm.nih.gov/pubmed/31738745.

461. Mccartney DL, Min JL, Richmond RC, Lu AT, Maria K, Davies G, Broer L, Guo X, Jeong A, Kasela S, Katrinli S, Kuo P, Matias- PR, Mishra PP, Nygaard M, Palviainen T, Soerensen M, Sun D, Tsai P, Matthijs D, Xu Z, Yao J, Zhao W, Correa A, Boerwinkle E, Durda P, Elliott HR, Gieger C, Genetics T, Consortium M, et al.: **Genome-wide association studies identify 137 loci for DNA methylation biomarkers of ageing**. *bioRxiv* 2020, :1–5010.1101/2020.06.29.133702Available: https://doi.org/10.1101/2020.06.29.133702.

462. Deursen JM van: **Senolytic therapies for healthy longevity**. *Science* 2019, **364**:636–63710.1126/science.aaw1299Available: https://www.sciencemag.org/lookup/doi/10.1126/science.aaw1299.

463. Gjaltema RAF, Rots MG: **Advances of epigenetic editing**. *Current Opinion in Chemical Biology* 2020, **57**:75–8110.1016/j.cbpa.2020.04.020Available: https://doi.org/10.1016/j.cbpa.2020.04.020.

464. Ni P, Huang N, Zhang Z, Wang D-P, Liang F, Miao Y, Xiao C-L, Luo F, Wang J: **DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning**. *Bioinformatics* 2019, **35**:4586–459510.1093/bioinformatics/btz276Available: https://academic.oup.com/bioinformatics/article/35/22/4586/5474907.

465. Zhu P, Guo H, Ren Y, Hou Y, Dong J, Li R, Lian Y, Fan X, Hu B, Gao Y, Wang X, Wei Y, Liu P, Yan J, Ren X, Yuan P, Yuan Y, Yan Z, Wen L, Yan L, Qiao J, Tang F: **Single-cell DNA methylome sequencing of human preimplantation embryos**. *Nature Genetics* 2018, **50**:12–1910.1038/s41588-017-0007-6Available: http://dx.doi.org/10.1038/s41588-017-0007-6.

466. Boyle EA, Li YI, Pritchard JK: **An Expanded View of Complex Traits: From Polygenic to Omnigenic**. *Cell* 2017, **169**:1177–118610.1016/j.cell.2017.05.038Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867417306293.

467. Langfelder P, Horvath S: **WGCNA: An R package for weighted correlation network**

analysis. *BMC Bioinformatics* 2008, **9**10.1186/1471-2105-9-559.

468. Blackledge NP, Thomson JP, Skene PJ: **CpG Island Chromatin Is Shaped by Recruitment of ZF-CxxC Proteins**. *Cold Spring Harbor Perspectives in Biology* 2013, **5**:a018648–a01864810.1101/cshperspect.a018648Available: http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a018648.

469. Sijacic P, Holder DH, Bajic M, Deal RB: **Methyl-CpG-binding domain 9 (MBD9) is required for H2A.Z incorporation into chromatin at a subset of H2A.Z-enriched regions in the Arabidopsis genome**. *PLOS Genetics* 2019, **15**:e100832610.1371/journal.pgen.1008326Available: https://dx.plos.org/10.1371/journal.pgen.1008326.

470. Hamdani O, Dhillon N, Hsieh T-HS, Fujita T, Ocampo J, Kirkland JG, Lawrimore J, Kobayashi TJ, Friedman B, Fulton D, Wu KY, Chereji RV, Oki M, Bloom K, Clark DJ, Rando OJ, Kamakaka RT: **tRNA Genes Affect Chromosome Structure and Function via Local Effects.** *Molecular and cellular biology* 2019, **39**:1–2610.1128/MCB.00432-18Available: http://www.ncbi.nlm.nih.gov/pubmed/30718362.

471. Rounge TB, Umu SU, Keller A, Meese E, Ursin G, Tretli S, Lyle R, Langseth H: **Circulating small non-coding RNAs associated with age, sex, smoking, body mass and physical activity**. *Scientific Reports* 2018, **8**:1–1310.1038/s41598-018-35974-4Available: http://dx.doi.org/10.1038/s41598-018-35974-4.

472. Tellier M, Maudlin I, Murphy S: **Transcription and splicing: A two-way street**. *Wiley Interdisciplinary Reviews: RNA* 2020, **11**10.1002/wrna.1593.

473. Wieben ED, Aleff RA, Basu S, Sarangi V, Bowman B, McLaughlin IJ, Mills JR, Butz ML, Highsmith EW, Ida CM, Ekholm JM, Baratz KH, Fautsch MP: **Amplification-free long-read sequencing of TCF4 expanded trinucleotide repeats in Fuchs Endothelial Corneal Dystrophy**. *PLOS ONE* 2019, **14**:e021944610.1371/journal.pone.0219446Available: https://dx.plos.org/10.1371/journal.pone.0219446.

474. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW: **Direct detection of DNA methylation during single-molecule, real-time sequencing**. *Nature Methods* 2010, **7**:461–46510.1038/nmeth.1459Available: http://www.nature.com/articles/nmeth.1459.

475. Price AL, Zaitlen NA, Reich D, Patterson N: **New approaches to population stratification in genome-wide association studies**. *Nature Reviews Genetics* 2010, **11**:459–46310.1038/nrg2813Available: http://www.nature.com/articles/nrg2813.

476. Jjingo D, Conley AB, Wang J, Mariño-Ramírez L, Lunyak VV, Jordan I: **Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human**

gene expression. *Mobile DNA* 2014, **5**:1410.1186/1759-8753-5-14Available: http://mobilednajournal.biomedcentral.com/articles/10.1186/1759-8753-5-14.

477. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen C-A, Lin S, Lin Y, Qiu Y, Xie W, Yue F, Hariharan M, Ray P, Kuan S, Edsall L, Yang H, Chi NC, Zhang MQ, Ecker JR, Ren B: **Integrative analysis of haplotype-resolved epigenomes across human tissues.** *Nature* 2015, **518**:350–35410.1038/nature14217Available: http://www.ncbi.nlm.nih.gov/pubmed/25693566.

478. Billon P, Côté J: **Precise deposition of histone H2A.Z in chromatin for genome expression and maintenance**. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 2012, **1819**:290–30210.1016/j.bbagrm.2011.10.004Available: http://dx.doi.org/10.1016/j.bbagrm.2011.10.004.

479. Tudge SJ, Watson RA, Brede M: **Game theoretic treatments for the differentiation of functional roles in the transition to multicellularity**. *Journal of Theoretical Biology* 2016, **395**:161–17310.1016/j.jtbi.2016.01.041Available: http://dx.doi.org/10.1016/j.jtbi.2016.01.041.

Appendices

6.5 *BioRχiv* manuscript: The Genomic Loci of Specific Human tRNA Genes Exhibit Ageing-Related DNA Hypermethylation

RichardJ.Acton[1,2,3], *WeiYuan*[4,5], *FeiGao*[6], *YudongXia*[6], *EmmaBourne*[7], *EvaWozniak*[7], *JordanaBell*[4], *KarenLillycrop*[3], *JunWang*[6], *ElaineDennison*[2], *NicholasHarvey*[2], *CharlesA.Mein*[7], *TimD.Spector*[4], *PirroG.Hysi*[4], *CyrusCooper*[2], *ChristopherG.Bell*[1]

1. William Harvey Research Institute, Barts & The London School of Medicine and Dentistry, Charterhouse Square, Queen Mary University of London, London, U.K.
2. MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, U.K.
3. Human Development and Health, Institute of Developmental Sciences, University of Southampton, Southampton, U.K.
4. Department of Twin Research & Genetic Epidemiology, St Thomas Hospital, King's College London, London, U.K.
5. Institute of Cancer Research, Sutton, U.K.
6. BGI-Shenzhen, Shenzhen, China
7. Barts & The London Genome Centre, Blizard Institute, Barts & The London School of Medicine and Dentistry, Queen Mary University of London, London, U.K.

the pre-print is included as a pdf file with the supplementary data.

Figure 6.1: https://doi.org/10.1101/870352

bioRχiv [1]

Under Consideration at Nature Communications

6.6 Specific tRNA annotations used in the tRNA analyses

Supplementary_Files/SF5_hg19-tRNAs.bed

6.7 tRNA targeted bisulfite sequencing primers

Supplementary_Files/SF6_Targeted_BiS-seq_Primers.txt

6.8 Additional tRNA Age modelling results

Full age model results for all tRNA genes and all model fitted with different covariates and subsets of the MeDIP-seq Data.

Supplementary_Files/SF1_MeDIPseqtRNACombinedModelSummariesXL.xlsx

6.9 tRNA mouse models

Supplementary_Files/SF4_mouse_tRNA_age_models.xlsx

6.10 tRNA proportional enrichment fisher's test count data

Supplementary_Files/SF11_prop_enrich_fishers.xlsx

6.11 TCGA sample used in tRNA analysis

Supplementary_Files/SF7_TCGA_samples_used.tsv

6.12 Twins UK 450k array tRNA age models

Supplementary_Files/SF3_TwinsUK_450k_blood_tRNA_Age_Models.xlsx

6.13 MAVIDOS RXRA and CDKN2A probe models

Supplementary_Files/SF12_vitD_bone_meth_model_results.xlsx

6.14 Age acceleration GWAS SNPs

model 1 Supplementary_Files/SF8_suggestivem1full.tsv

model 2 Supplementary_Files/SF9_suggestivem2full.tsv