University of Exeter
College of Medicine and Health

Developing and evaluating tools to improve the quality of DNA methylation association studies

Dorothea Seiler Vellame

Submitted by Dorothea Seiler Vellame, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Medical Studies, January, 2022.

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Signed: . . . .

Abstract

There is increasing interest in studying DNA methylation in the context of health and disease. A number of technical and analytical considerations are important to take into account when designing and interpreting DNA methylation studies, such as the experimental parameters used when quantifying DNA methylation differences between individuals and how best to account for study confounders, such as cellular composition. This thesis aims to address these issues by first developing a method to assess study power in bisulfite sequencing (BS) studies, second establishing a method for the estimation of error across reference based cellular deconvolution models, and third generating a novel reference based DNA methylation deconvolution model for the brain incorporating data for three neural cell types. In Chapter 2 the impact of bisulfite sequencing depth and sample size on power is investigated. It is shown that study power is not dependent on one specific parameter, but reflects the combination of multiple study-specific variables. Data simulation is utilised to generate an interactive tool for use by the wider research community that can be used to estimate the power of BS studies based on user-defined input variables including sample size and read depth filtering. In Chapter 3 an error metric is established for reference based cellular deconvolution approaches using DNA methylation data, which is validated using datasets derived from both blood and brain tissue. In Chapter 4 the reference based deconvolution model utilised for the deconvolution of brain tissue is refined to include an additional cell type, resulting in a three cell type model. The model was applied to bulk brain DNA methylation samples, showing that the addition of a third cell type improved insight gained from data generated on bulk brain tissue. Overall, this thesis aims to generate tools which can be utilised to better design and interpret DNA methylation studies, all of which have been made publicly available. This thesis also encourages researchers to clearly communicate any DNA methylation quality control decisions made and examine their methodologies to improve the transparency and reproducibility of their findings.

Acknowledgements

I would like to thank my supervisors Dr Eilis Hannon and Professor Jonathan Mill for entrusting me with this opportunity and supporting me through it, I am very grateful for your time, patience and expertise. Eilis, I am always impressed with the depth of your statistical understanding, and deeply appreciate the encouragement you've given me throughout my PhD. Jon, despite being one of busiest people ever you always managed to make yourself available when I needed help and I am forever in awe of how your mind works. I have learned many invaluable things from both of you.

I would like to thank all of the Epigenomics of Complex Disease group members, past and present, that I have been fortunate enough to work and be friends with. Szi Kay and Gemma, thank you for being excellent friends and housemates and listening to my numerous rants, you most definitely helped keep me sane through this seemingly never ending process and I couldn't appreciate you more. Bex, thank you for introducing me to all things nerdy, and for being a great joiner and initiator of fun, it's always a pleasure interacting with you inside and out of work. Marta, thank you for your encouraging words when I was deep in thesis writing, they really helped me. Outside of that you're a lovely and fun person to be around and I'm glad we're friends. Sam, Isabel and Jon D, as well as being wonderful people your enthusiasm for science is contagious, and I really appreciate that you were all incredibly approachable, meaning I could always ask even the most basic questions with no fear of judgement. Joana, thank you for your honesty in all things, your no bullsh*t approach to life is inspirational and I learned a lot from you. Greg and Michael, thank you for being chill guys, you're always fun to chat to. Finally, thank you to all those who generated the data I utilised in this thesis, without all your hard work I would have had nothing to analyse.

I would also like to express my gratitude to my industry supervisors at Eli Lilly, Dr David Collier and Professor Emma Laing. Emma you are one of the most generous and supportive people I have had the luck of working under, your compassion and eagerness to learn was awe inspiring. I am very grateful for the opportunity to have worked with so many brilliant researchers and especially thankful for Neil, Laura, Essie and Ruby for making the work culture so inviting and friendly. I would also like to say a special thank you Layla, who was my partner in crime during my placement. You made those three months some of the most enjoyable of my PhD, from running together to salsa dancing to cooking all our favourite meals together (I'm still obsessed with that aubergine chilli).

A special thanks to Jahcub, your laid back attitude to thesis writing was inspirational and you never fail to make me chuckle (even when you're trying to push me down hills); and to Katherine and Serena (who still have no idea what my research is about), your friendship is invaluable, thanks for the long long calls talking about anything and everything.

I would also like to thank my family. Mum and Phil, you are my rocks and will forever be grateful for your endless support, encouragement, and delicious cooking. Lu and Geo (aka the Sisquatch), I'm so happy to have you both in my life, you're the absolute best. You both made this process much more fun (especially given the pandemic of it all) through virtually working together and I can't thank you enough. Dad and Igor, your interest in science and want to understand all things technical has certainly rubbed off on me so thank you.

Finally, Josh, I don't know that I have the words to express how thankful I am that I met you. I'm honestly not sure how I would have completed this thesis without your unwavering support. My life is a thousand times better with you in it and I couldn't love you more.

Contents

List of Figures

List of Tables

Abbreviations

5caC 5-carboxycytosine

5fC 5-formylcytosine

5hmC 5-hydroxymethylcytosine

5mC 5-methylcytosine

A adenine

AD Alzheimer's disease

ADHD attention-deficit/hyperactivity disorder

ALS amyotrophic lateral sclerosis

ANOVA analysis of variance

ASD autism spectrum disorder

BA Broddman area

Bcell B cell

BDR Brains for Dementia Research

BMI body mass index

BS bisulfite sequencing

C cytosine

CD4T CD4$^+$ T cell

CD8T CD8$^+$ T cell

CEREB cerebellum

CETS cell epigenotype specific

Cetygo CEll TYpe GOodness

CGI CpG island

CP constrained projection

CpG cytosine-guanine dinucleotide

DL deep learning

DLPFC dorsolateral prefrontal cortex

DMP differentially methylated position

DMR differentially methylated region

DNA deoxyribonucleic acid

DNAm DNA methylation

Dnmt DNA methyltransferases

Double- NeuN-/SOX10-

DVM differentially variable methylation

EPIC Infinium Methylation EPIC BeadChip

ESC embryonic stem cell

EWAS epigenome wide association study

FACS fluorescence-activated cell sorting

FANS fluorescence-activated nuclear sorting

FEP first episode psychosis

G guanine

GEO Gene Expression Omnibus

Gran granulocyte

HIP hippocampus

HM27 Human Methylation 27K BeadChip

HM450 Human Methylation 450K BeadChip

IM iterative method

iPSC induced pluripotent stem cell

IRF8+ NeuN-/SOX10-/IRF8+

ISVA independent surrogate variable analysis

LR linear regression

LSM linear slope model

MACS magnetic-activated cell sorting

Mam35 Mammalian Methylation 35k BeadChip

MBD methyl CpG binding domain

MDD major depressive disorder

MM285 Infinium Mouse Methylation 285k BeadChip

Mono monocyte

MR Mendelian randomisation

mRNA messenger RNA

MS multiple sclerosis

NAFLD non-alcoholic fatty liver disease

NCBI National Center for Biotechnology Information

NIH National Institutes of Health

NIHR National Institute for Health Research

NK natural killer

OCC occipital lobe

PBMC peripheral blood mononuclear cell

PC principal component

PCA principal component analysis

PCR polymerase chain reaction

pcw post conception weeks

PD Parkinson's disease

PFC prefrontal cortex

PMI post-mortem interval

PPMI Parkinson's Progression Marker Innitiative

QC quality control

QP quadratic programming

RA rheumatoid arthritis

RBDM reference based deconvolution model

RMSE root mean squared error

RNA ribonucleic acid

RPC robust partial correlation

RRBS reduced representation bisulfite sequencing

RUV removing unwanted variance

SD standard deviation

SNP single nucleotide polymorphism

Sox10+ NeuN-/SOX10+

STR striatum

SVA surrogate variable analysis

SVR support vector regressions

SWAN subset-quantile within array normalization

SZ schizophrenia

T thymine

TBS targeted bisulfite sequencing

TET ten-eleven translocation

THAL thalamus

THC tetrahydrocannabinol

TpG thymine-guanine dinucleotide

Triple- NeuN-/SOX10-/IRF8-

U uracil

US United States

WGBS whole genome bisulfite sequencing

1. Introduction

1.1 DNA and gene expression

All individuals (excluding identical twins) have a unique deoxyribonucleic acid (DNA) sequence, comprising the genetic code from which the functional machinery of every cell is constructed. DNA is comprised of two polynucleotide strands coiled into a double helix, in which genetic information is stored as a sequence of four bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The DNA sequence can be categorized into two classes, coding and non-coding. The coding section contains contiguous strings of bases that code for genes which can be transcribed by enzymatic machinery to produce a single stranded polynucleotide ribonucleic acid (RNA), termed messenger RNA (mRNA). This process is known as gene expression. Ribosomes within each cell translate this mRNA in three base pair codons into individual amino acid components, which produce proteins.

Coding regions directly producing proteins account for as little as 1-3% of the human genome (Dunham et al., 2012). For a long while it was unclear exactly what the function of the rest of the genome was, however, it has since been found that the remaining non-coding regions have important functional implications including acting as regulatory regions for gene expression and influencing the DNA's 3D structure. For example, promoter regions, which are DNA sequences in close proximity to gene-coding regions, act as specific binding sites for transcription factors, allowing for recruitment of an enzymatic protein complex to the transcription start site (Alberts et al., 2002).

All mitotic cells in an organism contain an identical DNA sequence, yet their functionality and morphology are highly variable. This variation arises from a divergent set of expressed genes and the functional proteins formed from these genes (Zhong et al., 2018). Regulation of gene expression, therefore, cannot be due to the base genetic sequence of non-coding regions alone; gene expression is regulated by epigenetics.

1.2 Epigenetics

The concept of epigenetic mechanisms was first described in 1957 by Conrad Waddington (Waddington, 1957). Waddington theorised the existence of a mechanism by which cell fate divergence and differentiation could ensue, which would result in a diverse range of cell types within an organism, despite each cell being derived from an identical genetic sequence. More recently, the term epigenetics, the Greek origin of which translates to *above the genome*, has been broadly defined as a set of processes that influence gene expression which are not driven by changes in the DNA sequence (Waterland, 2006; Gan et al., 2007). Several types of epigenetic marks exist and their patterns across the genome are, by definition, cell type specific, in order to orchestrate the specific gene expression profiles that determine cell identify. These marks are mitotically heritable, meaning that the patterns will be passed from one cell to it's daughter cells during mitosis to maintain cell identity.

1.2.1 The 3D structure of DNA

If laid flat, each human cell contains approximately 2 metres of DNA, and as such DNA has to be stored efficiently to fit into a cell. Given that each cell type is defined by the expression of a subset of genes, access to all genes will never be necessary. Therefore, to compact the DNA while still allowing access to gene coding regions for cell type specific gene expression, it is organised into a 3D complex consisting of the DNA and protein. The DNA double helix is coiled in 145-147 basepair lengths around eight histone proteins to form nucleosome complexes, which are subsequently further packaged into chromatin and finally chromosomes.

Chromatin state is broadly defined by how condensed it is; euchromatin refers to regions of chromatin that is less densely compact and is associated with active gene expression, allowing access to transcription factors and transcriptional machinery. In contrast, highly compact heterochromatin is associated with silenced gene expression. The state of chromatin is changeable, with constitutive heterochromatin referring to regions that are dynamic and can be converted to a euchromatin state (Strachan and

Read, 2018). Epigenetic modifications regulate the chromatin state; two of which, histone modifications and DNA modifications, are described in the following Sections.

1.2.2 DNA modifications

A major class of epigenetic marks are DNA modifications. The base DNA sequence can undergo the covalent addition of specific chemical groups, altering the functionality of the base DNA sequence. DNA modifications have been observed at both A and C bases, although modifications at C have been better characterised as they occur more frequently in the mammalian genome (Kumar, Chinnusamy and Mohapatra, 2018).

The most well-studied of these C modifications is DNA methylation, often denoted as 5-methylcytosine (5mC), which is the addition of a methyl group to the fifth carbon position of a C. This modification is stable but reversible (Wu and Zhang, 2017) and has been considered a fifth nucleotide base, both due to its prevalence and its important functional role (Lister and Ecker, 2009).

Other C modifications exist; the 5-hydroxymethylcytosine (5hmC) modification is produced as a stable intermediate step of methylation (Münzel et al., 2010), through the oxidisation of 5mC by the ten-eleven translocation (TET) protein (**Figure 1.1**). It has been found to play a key role in regulating gene expression and chromatin structure (Mellén et al., 2012; He et al., 2021) and its abundance has been shown to correlate negatively with cell proliferation (Bachman et al., 2014). 5hmC is highly tissue specific, with the highest occurrence found in brain tissues (Globisch et al., 2010; Lunnon et al., 2016), and has been shown to change dynamically across human development (Spiers et al., 2017; Numata et al., 2012). The TET protein can further oxidise 5hmC into 5-formylcytosine (5fC) and 5-carboxycytosine (5caC), the latter of which can be decarboxylated to return to an unmodified C (**Figure 1.1**). These intermediates may have their own functional implications (Wang et al., 2015; Raiber et al., 2018; Klungland and Robertson, 2017), however, due to the rarity of said modifications, as well as challenges in experimental methodology (Plongthongkum, Diep and Zhang, 2014), their full functionality is as of yet unknown.

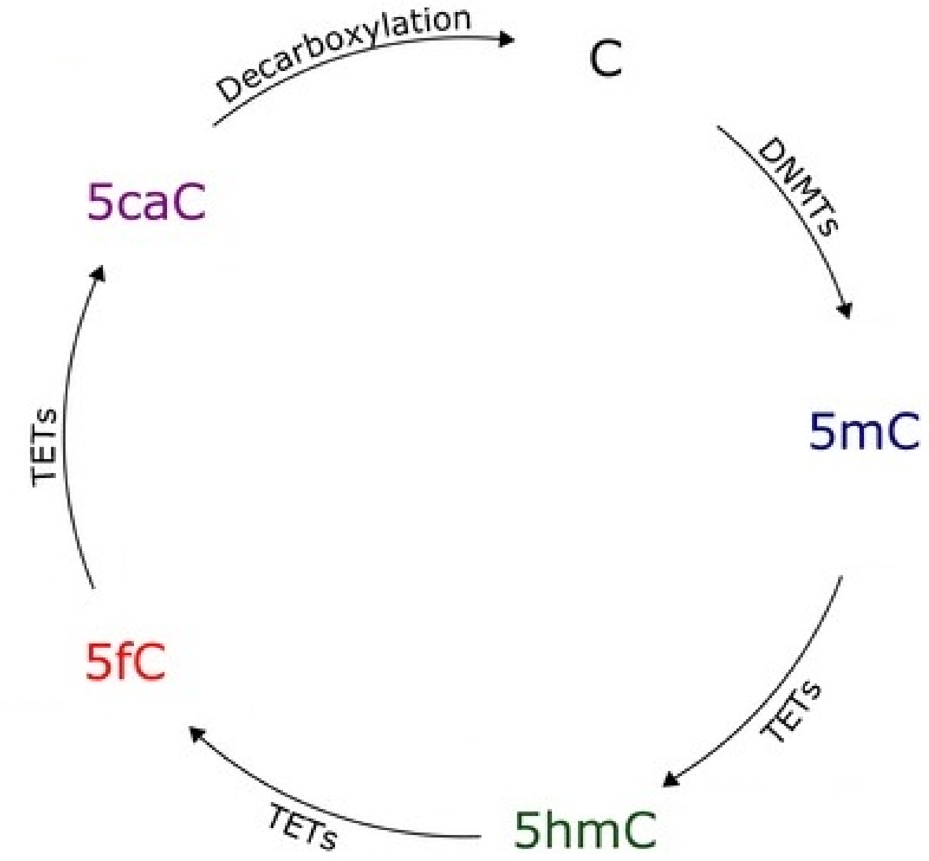


Figure 1.1: **A diagram of the cycle of cytosine modifications, adapted from Roubroeks et al., 2017**. Unmethylated cytosine (C) is methylated via the DNMT protein family to form methylated C (5mC). The TET family converts 5mC into hydroxymethylated C (5hmC), 5-formylcytosine (5fC) and finally 5-carboxycytosine (5caC), which, through decarboxylation, can return to an unmethylated C.

1.2.2.1 Cytosine context

DNA methylation primarily occurs at C next to a G, also known as a cytosine-guanine dinucleotide (CpG), which will be symmetrical on both DNA strands, as C and G nucleotides are complimentary. Mammalian genomes are globally CpG-depleted, with roughly 28 million CpGs in the human genome (Smith and Meissner, 2013), which is less than expected if their distribution was random. This is widely accepted to be the result of spontaneous deamination of 5mC nucleotides to form thymine-guanine dinucleotide (TpG) sequences (Simmen, 2008). About 60–80% of the CpG sites in the mammalian genome are methylated (Smith and Meissner, 2013; Illingworth et al., 2010). Broadly, areas of open chromatin and active transcription are unmethylated (Stadler et al., 2011), implying DNA methylation is associated with repressed gene expression.

1.2.2.1.1 CpG islands, shores, and shelves

CpG islands (CGIs) are loosely defined as areas of high CpG prevalence (relative to the rest of the genome) that are $\sim$1kb in length (Cooper, Taggart and Bird, 1983; Bird et al., 1985; Gardiner-Garden and Frommer, 1987). Less than 10% of CpGs occur in CGIs, and they are largely devoid of methylation (Deaton and Bird, 2011; Illingworth et al., 2010). The region between 0-2kb from a CGI are defined as shores, and 2-4kb are defined as shelves. Unlike CGIs, shores and shelves are more likely to be methylated, which, in these regions, is highly tissue specific and correlates with gene expression (Deaton and Bird, 2011).

CGIs overlap with 60-70% of promotor regions in human genes (Bird et al., 1985; Larsen et al., 1992). The lack of DNA methylation in these regions suggests that there will be an increased rate of transcription at the gene of the promoter, although this is not always the case (Jones, 2012). DNA methylation has been shown to affect the binding of transcription factors to their specific motifs (Yin et al., 2017), primarily with an inhibitive effect in CGIs containing these motifs. Additionally, methyl CpG binding domain (MBD)-containing transcriptional repressors can recognise CpG sequences, recruiting further proteins such as histone deacetylases, altering chromatin compaction and leading to gene silencing (Nan et al., 1998). However this effect is not universal and is influenced by the interplay of other epigenetic mechanisms. For example, methylation has been positively associated with increased transcription in certain genes, by disturbing the binding of protein complexes maintaining H3K27me3, a histone modification that generally represses transcription (Li et al., 2018).

1.2.2.1.2 Non-CpG methylation

The majority of DNA methylation occurs at CpG sites, however, there is also evidence of methylation occurring at non-CpG sites (Varley et al., 2013), such as CpH sites, where the H represents a non-G base, either A, C or T. Unlike CpGs, C methylation occurring at CpH sites will not be symmetrical across both DNA strands, as only guanine is complementary to C. CpH sites are enriched in low CpG density regions (Guo et al., 2013). Their methylation occurs only in certain cell types, such as embryonic and pluripotent stem cells, neurons and glia (Jang et al., 2017; Deaton and Bird, 2011), and are generally negatively correlated with gene expression (Guo et al., 2013). Even in these cell types, methylation at CpH sites is uncommon and its full functionality is not yet fully understood (Patil, Ward and Hesson, 2014).

1.2.3 Histone modifications

Histone proteins (H2A, H2B, H3 and H4, see **Figure 1.2**), around which DNA is coiled, have long N-terminal tails (i.e. sequences of amino acids that extend outwards from the core nucleosome complex) which can undergo reversible covalent modification at particular residues (Strahl and Allis, 2000). Functionally, the projection of the N-terminal tail from the nucleosome complex means that these sections of the histone protein make contact with other histones as well as specific binding sites of proteins involved with chromatin state and transcription. The post-translational modifications include phosphorylation, ubiquitination, acetylation and methylation and can occur at a number of amino acids along the tail (see **Figure 1.3** for the most common modification sites). The resulting number of possible combinations of histone modifications is enormous, with over 100 distinct histone modifications having been described (Zentner and Henikoff, 2013). The nomenclature for naming these modifications refers to the histone protein

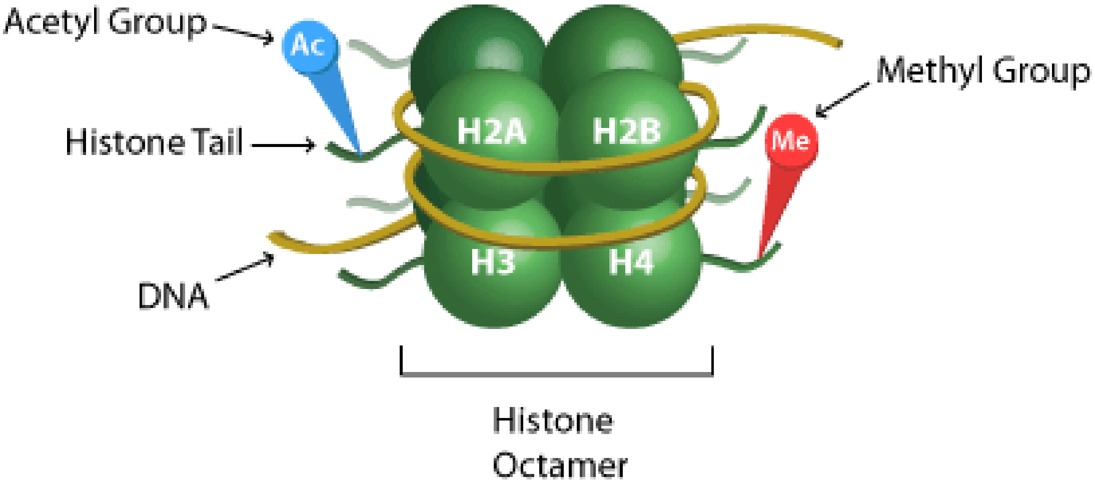


Figure 1.2: **A diagram showing a histone octamer, taken from CUSABIO, n.d.** The histone proteins H2A, H2B, H3 and H4 are shown, each of which have their own histone tail.

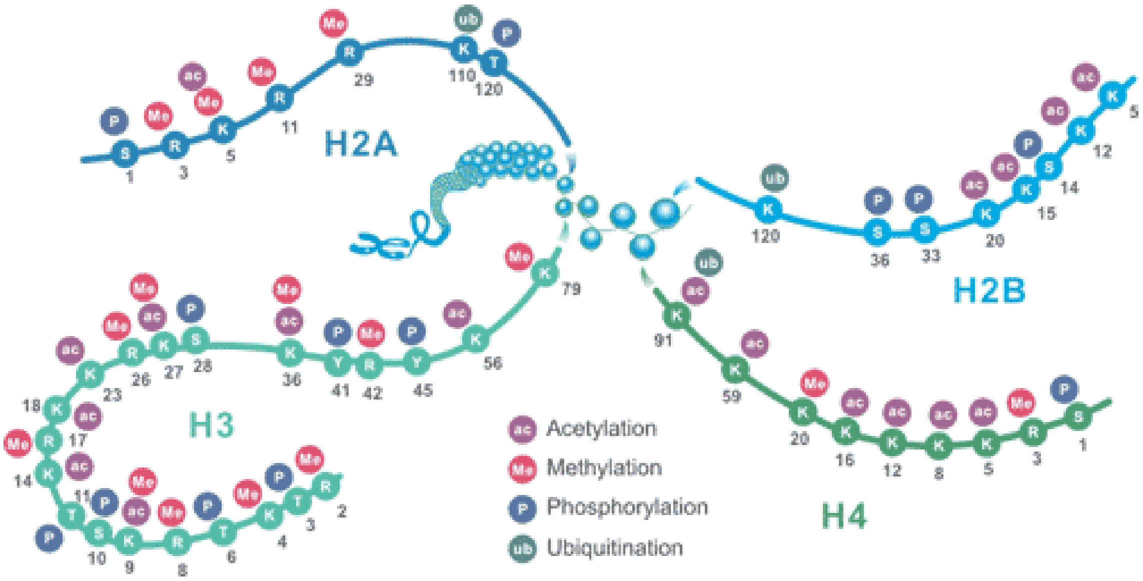


Figure 1.3: **A diagram showing common histone modification sites along each histone tail, taken from EpiGentek, n.d.** Each histone tail will be commonly modified at different positions along the tail. Acetylation is shown in purple, methylation is shown in pink, phosphorylation is shown in blue, and ubiquitination is shown in green.

type it occurs to, the amino acid residue at which it is added and the type of covalent modification. For example, a modification on the H3 histone protein at lysine residue 4 (K4) that adds three methyl groups (me3) is referred to as H3K4me3. The diversity of these modifications reflects the diversity of the roles they play in determining hetero and euchromatin state, DNA accessibility, and subsequently, differential gene expression. As with many epigenetic modifications, most histone modifications are not specific to one particular function and show a diverse range of functional associations based on the context in which they occur, in particular in their interaction with other histone modifications (Ernst et al., 2011).

For example, there has been considerable research interest in histone acetylation, which, over 50 years ago, was observed to be associated with highly transcribed genes (Allfrey, Faulkner and Mirsky, 1964) and more recently with the modification H3K27ac being identified as a marker of active gene enhancers (Creyghton et al., 2010), as apposed to non-active "poised" enhancers, identified by the presence of H3K4me1 alone. H3K27ac is the most common histone modification, and acts by reducing the strength of the charge dependent interactions between DNA and histone, and histone and histone, increasing DNA accessibility (Zentner and Henikoff, 2013). Neither histone modifications or DNA methylation function in solitude and have been shown to both influence and be influenced by each other (Cedar and Bergman, 2009; Fu, Bonora and Pellegrini, 2020).

1.2.4 The role of epigenetics in cell type differentiation

Cellular development is a meticulously orchestrated transition in which pluripotent cells differentiate into specialised cell types with distinctive functions and genes expressed. A simplified version of this differentiation into distinct cell lineages is summarised in **Figure 1.4A**, which utilises Waddington's 'epigenetic landscape' concept (Waddington, 1957). Here, as cells move down the landscape, they gain specificity and become more differentiated. This lineage specification requires cell type specific gene expression (Zaidi et al., 2011), the mediation of which is regulated by epigenetics, including DNA methylation.

DNA methylation patterns are established and maintained by the DNA methyl-

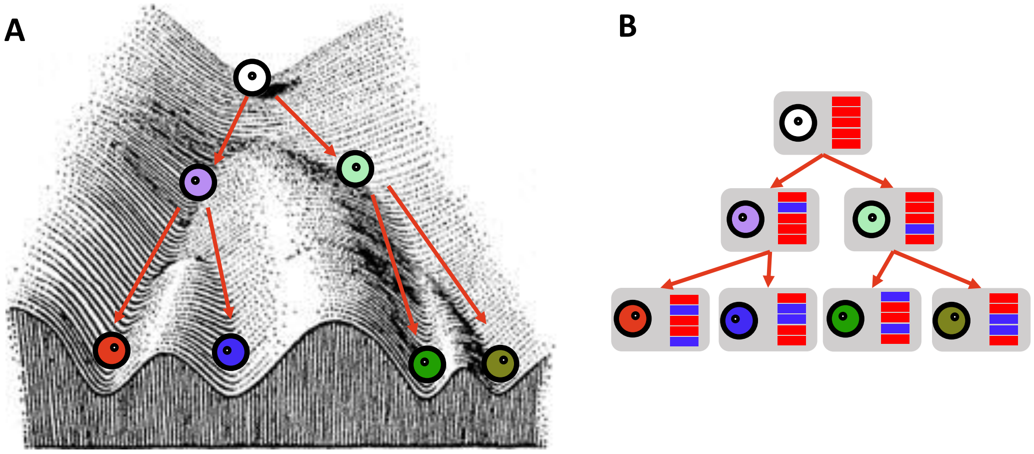


Figure 1.4: **A diagram demonstrating cell type specific demethylation as cell type differentiation progresses.** A) Cells can be seen moving along Conrad Waddington's 'epigenetic landscape' (Waddington, 1957), becoming increasingly differentiated as they move downwards. B) A simplification of each cell type DNA methylation profile can be seen next to the cell, with the same five DNA methylation sites depicted at each cell. Red represents the presence of DNA methylation and blue the absence of DNA methylation. As differentiation occurs, demethylation occurs at cell type specific genomic locations.

transferases (Dnmt) protein family; Dnmt3a and Dnmt3b establish new DNA methylation patterns, and Dnmt1 maintains DNA methylation status and replicates DNA methylation patterns during mitosis (Bhattacharya et al., 1999; Moore, Le and Fan, 2013). This mitotic heritability by daughter cells allows DNA methylation to act as cell type memory (Zaidi et al., 2011). To that end, DNA methylation plays an essential role in mammalian development (Li, Bestor and Jaenisch, 1992) and is fundamental for the establishment and maintenance of cellular identity (Suelves et al., 2016).

In vitro studies of embryonic stem cell (ESC) lineages allow for the assessment of lineage specific DNA methylation profiles; Xie et al., 2013 found that the majority of gene promoters driving early lineage specific gene expression were CpG rich and unmethylated. By contrast, the promotors of genes differentially expressed in later stages of development contained fewer CpGs and, in lineages not expressing said genes, DNA methylation mediated gene silencing was implicated (Xie et al., 2013). In general, the loss of plasticity and narrowing of cellular identity amongst ESCs was stably maintained by DNA methylation, with the loss of DNA methylation occurring in a more lineage specific manner than the gain (Suelves et al., 2016).

Cell type specific DNA methylation patterns have been observed across the genome, comparing DNA methylation quantified across purified cell types (resulting in a proportion or percentage of DNA methylation across cells in the purified sample at each DNA methylation site characterised, see Section 1.3.4 for details on DNA methylation quantification). The resulting DNA methylation differences between tissue and cell types are of large magnitude and occur at many sites across the genome, as found by Hannon et al., 2021b. The authors compared the DNA methylation profiles of nasal, buccal, whole blood, and purified blood (monocytes, granulocytes, CD4+ T cells, CD8+ T cells, and B cells) samples at over ~850,000 DNA methylation sites, finding significant differences between at least two sample types at 77.9% of sites. Site specific differential DNA methylation was often upwards of 60%, as seen in **Figure 1.5** (Hannon et al., 2021b). Their data also highlighted that methylomic differences between cell types are hierarchical, with cell types of the same lineage sharing a larger proportion of DNA methylation. This is visualised in **Figure 1.4B**, in which cell types resulting from the purple lineage have more similar DNA methylation profiles to each other than those differentiated from the green lineage. Large differences have also been observed across cells within other tissue types, such as prefrontal cortex (PFC) tissue (Guintivano, Aryee and Kaminsky, 2013), and within saliva (Middleton et al., 2020). Between cell type variation in DNA methylation dramatically exceeds that of within cell type variation between individuals (Ziller et al., 2013; Byun et al., 2009). Inter-individual variation in DNA methylation is described further in Section 1.3.8.

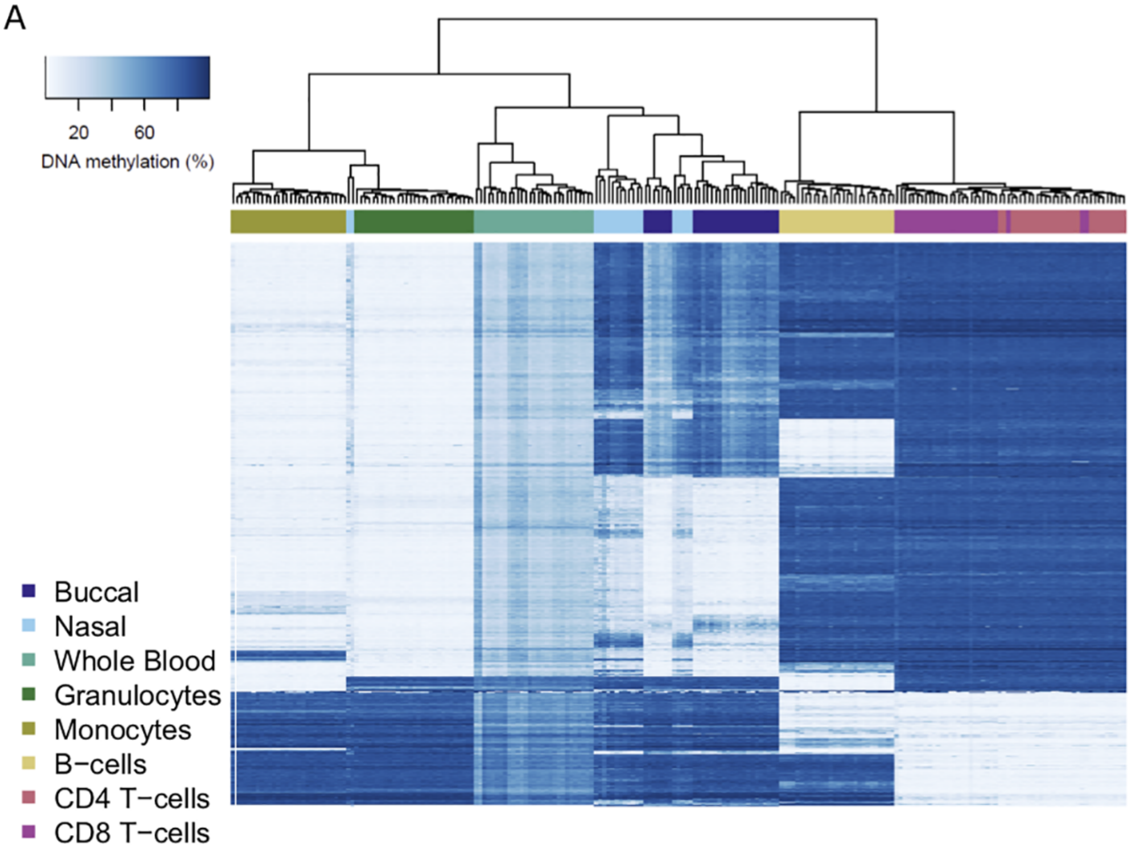


Figure 1.5: **A heatmap demonstrating differential methylation across nasal, buccal, whole blood, and purified blood (monocytes, granulocytes, CD4+ T cells, CD8+ T cells, and B cells) samples from Hannon et al., 2021b.** A heatmap of DNA methylation values across purified cell types and peripheral tissues for the top 1000 most variable DNA methylation sites (ranked by standard deviation). Each row depicts data for an individual DNA methylation site, and each column depicts data from one individual sample. The order of rows and columns was determined by hierarchical clustering to group together similar profiles of DNA methylation. Low levels of DNA methylation are represented by white boxes and high levels of DNA methylation represented by blue boxes. The colored bars across the top of the columns depict different sample types.

1.3 Epigenome wide association studies of DNA methylation

DNA methylation is influenced by many factors other than cell identity, and as such there has been an increased interest in the study of DNA methylation over the past 50 years (**Figure 1.6**), especially in the context of health and disease. Epigenome wide association studies (EWAS), primarily carried out in humans, have identified genome wide DNA methylation differences across many complex disorders, traits and exposures, including but by no means limited to Alzheimer's disease (AD) (Lunnon et al., 2014; De Jager et al., 2014; Pishva et al., 2020), cancer (Koch et al., 2018), autism spectrum disorder (ASD) (Wong et al., 2019), suicide completion (Policicchio et al., 2020b), attention-deficit/hyperactivity disorder (ADHD) (Walton et al., 2016), schizophrenia (SZ) (Hannon et al., 2016), rheumatoid arthritis (RA) (Liu et al., 2013), body mass index (BMI) and adiposity (Wahl et al., 2016), high blood pressure (Kazmi et al., 2020), smoking (Bollepalli et al., 2019; Lee et al., 2016; Zeilinger et al., 2013), and diabetes (Stefan et al., 2014; Davegårdh et al., 2018).

There are many decisions a researcher must make to conduct an EWAS, with many tools available to ease the process and improve quality control (QC) and analysis pipelines, however, naturally, no analysis is without limitations and as such, methodologies for analysis are constantly evolving. The focus of this thesis is to develop tools that can be used within the QC pipeline to address current challenges within EWAS with the aim of improving general reproducibility of findings (see Section 1.5). In the following Sections, the utility of and necessity for these tools is contextualised within the EWAS pipeline.

1.3.1 General considerations - sample size and statistical power

Statistical power is defined as the probability that the null hypothesis is correctly rejected. In the context of DNA methylation, power refers to the probability of identifying DNA methylation sites or regions at which there is a true difference in DNA methylation between groups. Observing true differences in DNA methylation associated with an outcome can be challenging due to the small magnitude of effect (see Section 1.3), as

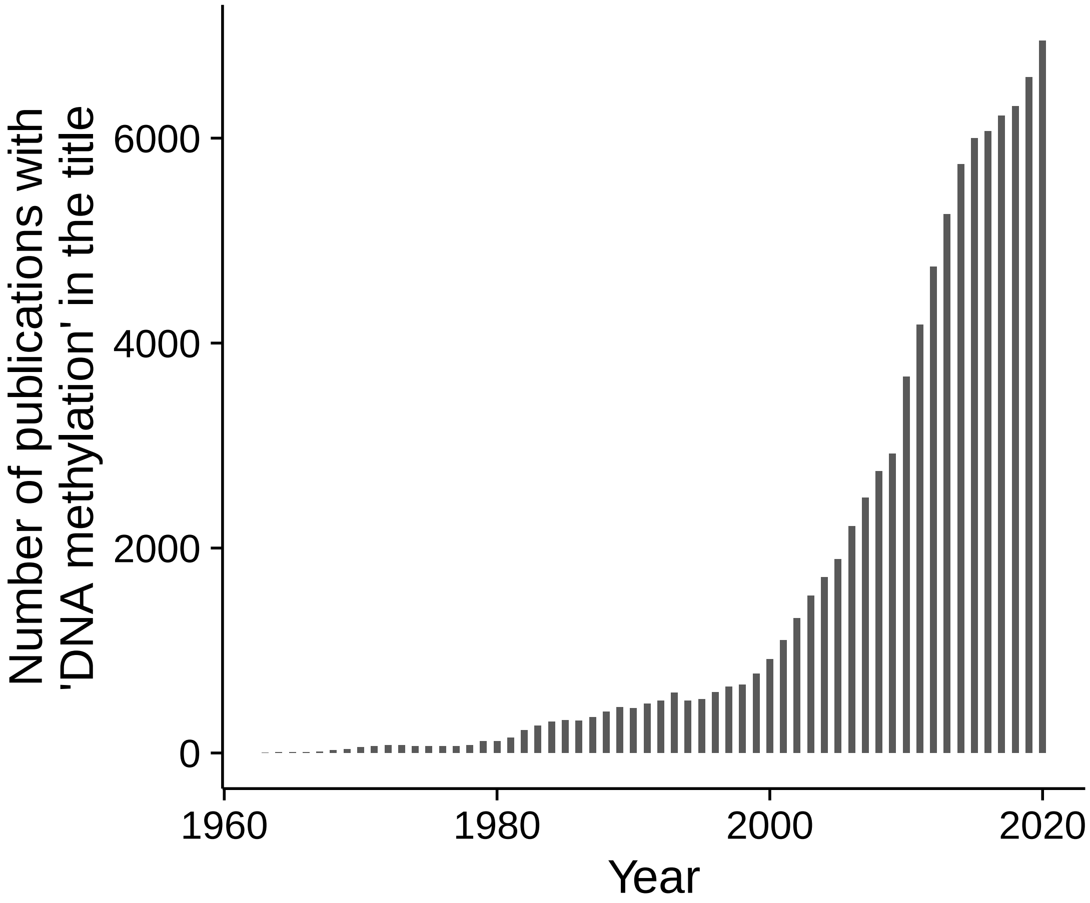


Figure 1.6: **A summary of the number of publications on DNA methylation has increased over the last 60 years.** The figure plots the number of publications with key phrase "DNA methylation" in the title, stratified by the year of publication. Data from pubmed.ncbi.nlm.nih.gov/

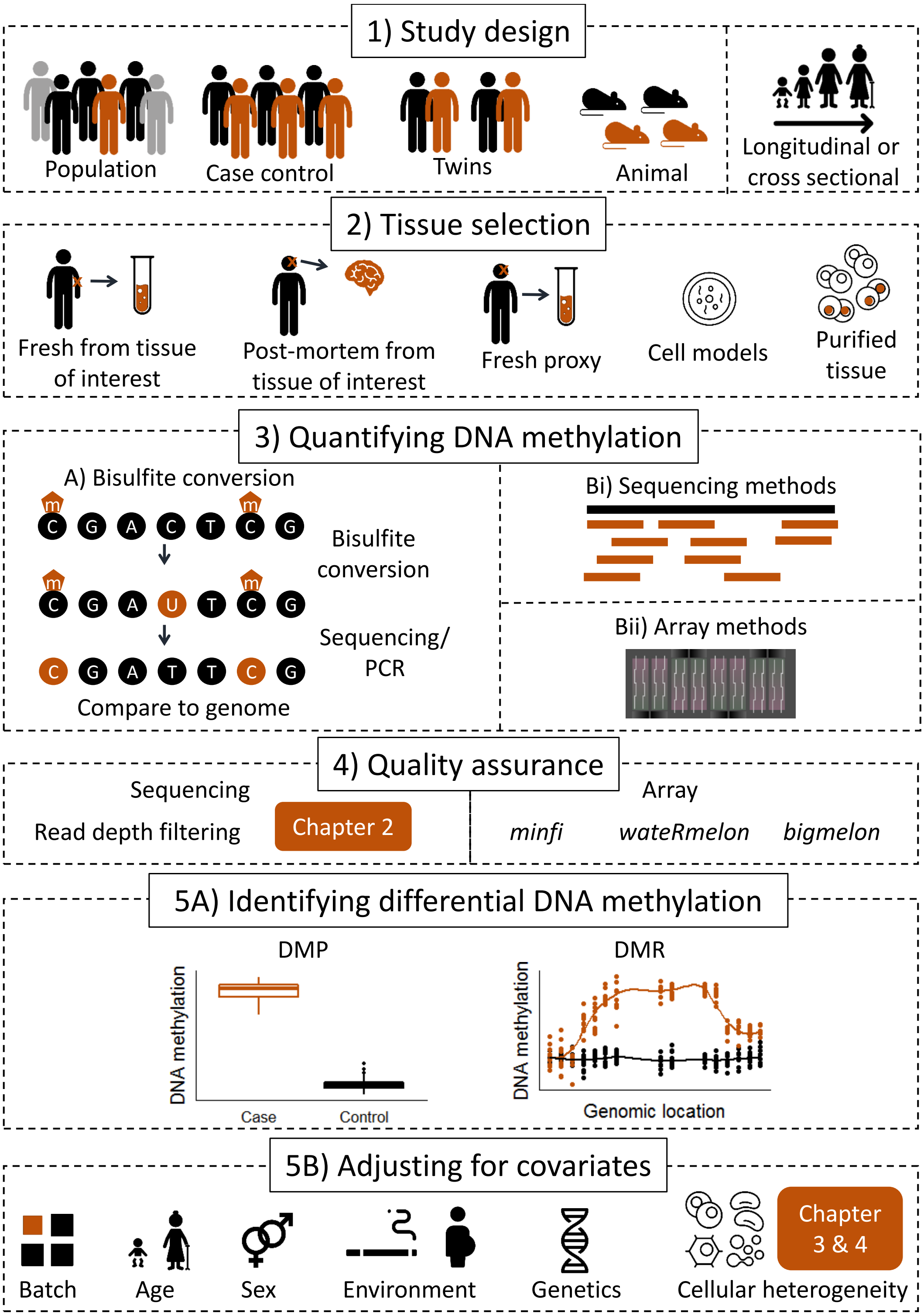


Figure 1.7: **An overview of the main steps in an epigenome wide association study of DNA methylation.** 1) study design; 2) tissue selection; 3) quantifying DNA methylation, the methods shown first require bisulfite conversion, followed either by sequencing or profiling using array based techniques; 4) quality assurance, which depends on 3. The somewhat arbitrary QC of sequencing methods is addressed in Chapter 2; 5A) identifying differential DNA methylation, which can include identifying differentially methylated positions (DMPs) and/or differentially methylated regions (DMRs); 5B) adjusting for covariates, i.e. features that coincide with differential DNA methylation that is unrelated to the disorder of interest, including cellular heterogeneity of the tissue sampled. Chapter 3 and 4 improve aspects of cellular composition prediction.

well as the high number of DNA methylation sites compared, and as such, a large number of samples are often required. Without high enough power, true results may not be uncovered and as such the reproducibility between studies will be lower. The exact size of the dataset will be dependent on a number of factors, each of which can decrease power:

- **Multiple testing burden** – In an EWAS, to identify differential DNA methylation, a statistical test is performed across all DNA methylation sites profiled. If only one site were being assessed, it would usually be considered significant if the statistical test p value was less than 0.05. However, since the number of sites being tested is far greater than one (the human genome contains approximately 28 million CpGs, although most DNA methylation quantification methods profile somewhere in the 100,000's, see Section 1.3.4), using a p value threshold that does not account for the multiple testing would lead to a high proportion of Type I error, i.e. false positive results, due to random chance. To that end, a multiple testing adjusted significance threshold should be applied to correct for the number of independent tests, making the p value threshold more stringent and minimising the proportion of DNA methylation sites detected as significant by chance. This means more samples are required to detect robust differences between groups.
- **Small magnitude of DNA methylation differences between groups or across phenotypes** – Power is also lower when the DNA methylation difference between groups or across phenotypes of interest is small. This is the case in most complex disease cases, where DNA methylation differences are generally lower than 5% (Lunnon et al., 2014; Wong et al., 2019; Hannon et al., 2016). To robustly detect small differences, more samples will be required.
- **High variability across individuals** – Variability in DNA methylation across individuals due to genetics or lifestyle factors (see Section 1.3.8 for details of covariates), as well as stochastic variation, can further complicate analysis and reduce power; representing this variability across the cohort requires a large sample size. Furthermore, a study requiring a large number of covariates will need more samples to enable the accurate estimation of their coefficients. Aspects of this variability will be reduced in twin, longitudinal and animal studies, where genetics

and environment can be more carefully controlled.

With the aim of maximising the DNA methylation sites investigated, one way to increase power in the population of interest is to increase the study sample size. However, the ability to do so will depend on the resources available, also on accessibility of the disease and tissue to be profiled, and specifics of the study.

1.3.2 Study design

When selecting a study population for a DNA methylation association study, the main considerations will be the study aims and the resources available. Several common cohort types exist, which are described below.

1.3.2.1 Population cohorts

Population cohorts aim to profile DNA methylation across a representative sample of the wider population and can be used to assess DNA methylation across various phenotypes. In general, these cohorts have larger sample sizes in order to capture inter-individual variation. They can be leveraged to investigate the relationship with DNA methylation across a range of phenotypes or traits, usually utilising medical records and questionnaires. However, if interest lies in a rarer phenotypes, such as disease with lower prevalence across the general population, there may not be sufficient statistical power to investigate them due to the low phenotype specific sample size. Furthermore population cohorts typically lack robust characterisation of particular disease measures, unlike cohorts designed to enrich for affected cases. Relying on questionnaires or clinical records may mean that disease status has not been measured as precisely as when biological measures are utilised, and cohorts may lack information on sub-phenotypic measures in complex disease, which may confound results.

1.3.2.2 Enriching for affected cases

Complex disorders may have low prevalence across the general population. As such, disease specific study design can be used in which recruitment targets those with said disease of interest. To ascertain DNA methylation differences between the disease or phenotype of interest and the general population, a control group is utilised. Ideally, this control group is matched for other variables, such as age, sex, and other potential drivers of differential DNA methylation outside of the disorder as these can confound findings (see Section 1.3.8). Therefore, a strength of the case control study design is the potential for tighter control of confounding variables.

A limitation of this study design selection method is that the resulting comparison of cases and controls is often a comparison of extreme cases (i.e. high disease severity and no disease) and so will not necessarily be informative in the investigation of disease progression, although this will depend on the disease of interest.

1.3.2.3 Twin studies

Twins can generally be categorised into two groups: monozygotic, in which twins are genetically identical, and dizygotic, in which twins are genetically as similar as siblings.

There are two main utilities to studying DNA methylation through twins: firstly, they can allow researchers to quantify the heritability of DNA methylation (Dongen et al., 2021). Secondly, disease discordant twins (i.e. where only one twin has the disease of interest) can be utilised to identify differential DNA methylation across disease while minimising some confounders (see Section 1.3.8 for an overview of common confounders in EWAS). When comparing disease discordant monozygotic twins, any genetic effects on DNA methylation patterns will be consistent, twins will be matched for age (assuming samples are taken at the same time), early environment will be shared (i.e. *in utero* maternal effects (Bell and Spector, 2012)) and later environment is more likely to be similar (Castillo-Fernandez, Spector and Bell, 2014). This will potentially increase the statistical power to identify DNA methylation between groups.

The main challenge to twin studies is in acquiring samples, especially for disease discordance, due to rarity of both twins, and potential rarity and heritability of disease, which can limit the maximum sample size, subsequently reducing study power. Furthermore, it is not always possible to ensure that disease discordance is complete, as it might be, for example, that disease onset will occur in the non-diseased twin at a later date, which may still coincide with differential DNA methylation.

1.3.2.4 Animal models

Most population based, enriched case, and twin studies are carried out in human, however, human studies have their limitations. Animal models are often utilised when trying to understand the environmental effects of DNA methylation, as the environments of animals can be carefully controlled in a way that is unethical to apply to humans. Animals can also be bred so as to control genetics, meaning that the genetic effect on DNA methylation will be more consistent between animals.

Animal models can utilise many different species, from *C. elegans* to non-human primates, with one of the most commonly used being rodent models. Phenotypic proxies for disease in rodents can be researched either using transgenic models which have been genetically modified to emulate disease or by inducing disease effects experimentally. Rodents have the advantage of having a shorter life span than humans making age, a major risk factor for many disorders, easier (and cheaper) to study.

One unanswered question is how translatable animal research is to humans, given DNA methylation difference observed. For example, Zhou et al., 2017 found that only 11-37% of tissue specific DNA methylation was conserved between rodents and humans, largely due to sequence conservation (Zhou et al., 2017). The importance of the disparity between species will depend on the biological question that a study poses. Researchers utilising these models must be careful not to over-interpret their findings to human disease (Neff, 2019).

1.3.2.5 Longitudinal vs cross sectional studies

Longitudinal studies are those with repeated measures of DNA methylation from the same individual across time. They can allow for a comparison within an individual of DNA methylation changes that co-occur with disease progression, exposure, or some other feature. By investigating individual trajectories, inter-individual variation can be controlled for as factors, such as genetics and many lifestyle factors, will remain consistent, minimising confounding.

Longitudinal studies can allow for the study of the relationship between DNA methylation and disease progression across individuals, which cannot be done using cross sectional sampling. When designing a longitudinal study of disease, who or what time points to include takes careful consideration. The criteria will depend on the specific aim, for example, if the interest in early disease stage with the aim of identifying biomarkers to predict onset, then a cohort of high risk individuals that could be followed until conversion to disease status may be preferable. A second consideration is how long to follow individuals for, which may depend on the resources available to the study and expected change being investigated.

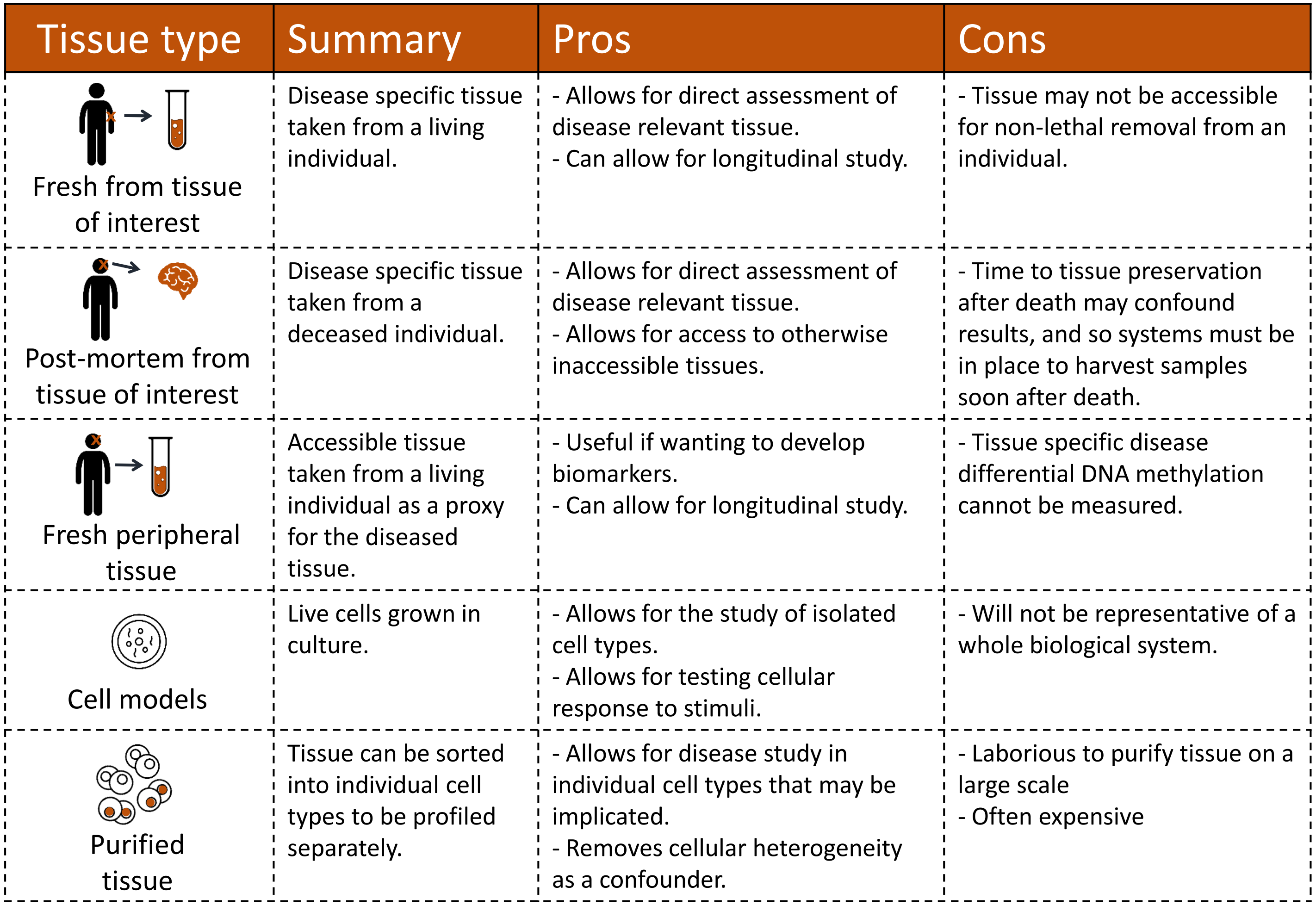
A limitation of longitudinal data is that it commonly has missing data points, as retention rates are never 100%. When investigating disease, attrition resulting from death or drop out of those with the worst severity may lead to survival bias in the cohort.

A further limitation is that the tissue in which DNA methylation is profiled for longitudinal studies must be samplable from living people and as such may not be the primary tissue of interest for the disease (described in 1.3.3).

In contrast, cross sectional studies are those in which only one measure of DNA methylation is obtained per sample at only one time point. Such studies require fewer resources than their repeated measure counterparts. Since a second time point is not required they can be carried out in post-mortem samples. However, they only allow for the investigation of a specific state and cannot as easily be utilised to investigate disease progression, as may be confounded by individual differences.

1.3.3 Tissues or cell types used

As outlined in Section 1.2.4, DNA methylation patterns are cell type specific and therefore the tissue and as such, the theory that phenotypic differences are mirrored between tissues is unlikely to be valid (Hannon et al., 2015a). Therefore, the tissue or cell types used in EWAS will impact the results uncovered, as between tissue variability will generally exceed within tissue phenotypic differences. The tissue used will primarily be dependent on the specific study aims and biological question being posed, as well as tissue availability. Five common options have been summarised in **Figure 1.8**.



Tissue type	Summary	Pros	Cons
Fresh from tissue of interest	Disease specific tissue taken from a living individual.	- Allows for direct assessment of disease relevant tissue. - Can allow for longitudinal study.	- Tissue may not be accessible for non-lethal removal from an individual.
Post-mortem from tissue of interest	Disease specific tissue taken from a deceased individual.	- Allows for direct assessment of disease relevant tissue. - Allows for access to otherwise inaccessible tissues.	- Time to tissue preservation after death may confound results, and so systems must be in place to harvest samples soon after death.
Fresh peripheral tissue	Accessible tissue taken from a living individual as a proxy for the diseased tissue.	- Useful if wanting to develop biomarkers. - Can allow for longitudinal study.	- Tissue specific disease differential DNA methylation cannot be measured.
Cell models	Live cells grown in culture.	- Allows for the study of isolated cell types. - Allows for testing cellular response to stimuli.	- Will not be representative of a whole biological system.
Purified tissue	Tissue can be sorted into individual cell types to be profiled separately.	- Allows for disease study in individual cell types that may be implicated. - Removes cellular heterogeneity as a confounder.	- Laborious to purify tissue on a large scale - Often expensive

Figure 1.8: **A summary of tissue types available for use in EWAS studies and their pros and cons.**

1.3.3.1 Fresh from tissue of interest

To investigate the underlying biology of a disorder, the primary tissue type implicated in disease progression is preferable (where it is known). This could be biopsied tissue of interest, as commonly used for studies of diet and metabolism which utilise adipose tissue (Anguita-Ruiz et al., 2021; Crujeiras et al., 2016; Guénard et al., 2017). Tissue taken from living individuals has the advantage of being utilisable for a longitudinal study design (subject to ethics) as it can allow for repeated sampling over time. It also allows for greater flexibility in making new phenotypic assessments that cannot be made post-mortem but may be of interest.

1.3.3.2 Post-mortem from tissue of interest

Depending on the disorder, the tissue type most impacted may not be accessible or ethical to sample from living patients. If investigating disease impact, an alternative would be using post-mortem tissue, as commonly used in studies of brain disorders (Smith et al., 2021; Policicchio et al., 2020b). DNA methylation has been shown to be generally stable in post-mortem samples, although may be dependent on post-mortem interval, that is, the time between death and tissue preservation (Rhein et al., 2015), as well as the method of preservation used. Post-mortem brain samples are often also scarce for specific disorders, such as autism, schizophrenia, and bipolar disorder and do not even exist for more common conditions like anxiety, as samples can be challenging to obtain in large numbers (Bakulski et al., 2016b). There may also be limited phenotypic information on post-mortem samples, which cannot always be collected after death (i.e. questionnaires).

1.3.3.3 Fresh peripheral tissue

A commonly employed alternative to post-mortem tissue is utilising an accessible peripheral tissue, such as blood, that can be taken from live samples, again allowing for longitudinal study. Not all peripheral tissues will exhibit the specific changes that occur across disease in the tissues most affected (Hannon et al., 2015a) and so careful consideration is needed when choosing which tissue to use. Differential DNA methylation uncovered in said tissues can lead to the development of biomarkers and predictive tools for disease classification or progression, as such tools will only have utility if they can be applied across live patients. Although peripheral tissues may not relate directly to the pathogenic processes of disease, they may capture secondary effects in diseased individuals such as medication status or peripheral side effects.

1.3.3.4 Cell types

The proportion of DNA methylation in bulk tissue at any one DNA methylation site is the proportion of methylated cytosines, which will be a binary value per DNA strand, and so have three possible proportions across all cells profiled (0%, 50% and 100% due to the two copies of each chromosome). It might be the case that only a subset of cell types within a tissue are affected by the disease of interest and therefore, ideally, those cell types would be investigated in isolation. Furthermore, cell type specific DNA methylation differences are typically larger magnitude than those associated with disease (see Sections 1.2.4 and 1.3.1). Therefore, a small difference in cell type composition between samples or groups of samples being compared in EWAS may result in false results. To that end, the next two sample types investigate individual cell types only, rather than tissue.

1.3.3.4.1 Cell models

Cell models may be utilised in research, although are less common in EWAS with large sample size. Specific cell lines that resemble a distinct cell type or precursor can be cultured, allowing for the measurement of differential methylation as the result of some stimuli (more details on neuronal cell lines can be found in Section 4.1.1.4.4). Cell lines will be genetically identical, and as such can have similar utility to twin or animal studies. However, a limitation to the use of cell models is that they do not represent cells as a part of a wider system and so will not be able to capture the interactions that occur between cells that may alter DNA methylation.

1.3.3.4.2 Purified tissue

Purified cells, isolated from bulk tissue from affected and non-affected individuals using methods such as fluorescence-activated nuclear sorting (FANS) (detailed in Section

4.1.1.1), enable cell type specific investigations of diseased tissue and have high utility in investigating disease-associated epigenetic variation in a cell type specific manner. This is important as many disorders have been shown to have cell type specific signatures, such as SZ(Chen et al., 2015) and AD (Gasparoni et al., 2018). These methods also reduce the issue of cellular heterogeneity, in which observed differential DNA methylation is driven by differences in cell type composition between samples, rather than by a disorder (see Section 1.3.8). Methods for sorting individual cell types from bulk tissue are increasingly prevalent for tissues such as brain (Policicchio et al., 2020a; Matevossian and Akbarian, 2008), however, they are not always feasible for all tissue types, be it due to prohibitive cost or a lack of effective methodologies for the tissue of interest.

1.3.4 Quantifying DNA methylation across the genome

In order to perform an EWAS, a high throughput, accurate method to quantify DNA methylation across the genome in multiple samples is required. At a single strand of DNA, DNA methylation is binary where methyl groups are either bonded to a cytosine or not. When quantifying the DNA methylation status at a site across multiple cells at a time (e.g. in a tissue), the resulting value will be the average DNA methylation at that site across cells, represented as a proportion or percentage.

Many of the methods to quantify DNA methylation, start with a bisulfite conversion step (as seen in **Figure 1.7 3A**), in which sodium bisulfite treatment of DNA converts non-methylated C into uracil (U), which will be converted to thymine during PCR (Wang, Gehrke and Ehrlich, 1980; El-Maarri, 2003; Hayatsu, 2008) and methylated Cs remain unchanged. Large scale EWAS (i.e. those including many hundreds of individuals) rely on one of two types of subsequent approaches for DNA methylation quantification: bisulfite sequencing (BS) methods, and microarray based methods.

1.3.4.1 Bisulfite sequencing methods

Bisulfite sequencing was first developed by Frommer et al., 1992, the first stage of which is bisulfite conversion. Bisulfite conversion results in the deamination of unmethylated Cs, while methylated Cs remain unchanged. Sequencing libraries are generated from the converted single stranded DNA using multiple primers and a modified polymerase, producing individual sequences of DNA which are uniquely tagged and barcoded. These amplified libraries then undergo next generation sequencing, producing data for each individual DNA molecule, defined as reads, that are subsequently aligned to the genome using computational methods, to determine where they originate from (Frommer et al., 1992). The most comprehensive method of BS is whole genome bisulfite sequencing (WGBS), which can theoretically be used to interrogate all $\sim$28 million CpGs in the human genome (Cokus et al., 2008; Lister et al., 2009; Laurent et al., 2010; Urich et al., 2015) and is often considered the gold standard of DNA methylation quantification (Plongthongkum, Diep and Zhang, 2014).

Many DNA methylation sites across the genome are consistent in their methylation status across the population (with the notable exception of cancer tissues (Cao et al., 2020)). For example, Ziller et al., 2013 found that 70–80% of CpGs in the human genome were stably methylated across 30 diverse cell and tissue types profiled using WGBS (Ziller et al., 2013). As such, the broad genomic coverage provided by WGBS may be inefficient if one aims to capture differential DNA methylation within a tissue. Deep sequencing, in which many reads are used to average DNA methylation across DNA methylation sites, which would be required for sensitive DNA methylation estimates, can be cost prohibitive for larger EWAS studies (Ziller et al., 2015). To that end, Gu et al., 2011 developed an alternative method, reduced representation bisulfite sequencing (RRBS) in which the methylation-insensitive enzyme, MspI, is utilised to enrich CpG-rich regions of the genome. These are of interest due to their presence in CGIs overlap with promoter regions (Bird et al., 1985; Larsen et al., 1992) and subsequent association with gene expression. The method typically captures 85-90% of CGI (Meissner et al., 2008; Smith et al., 2009), reducing sequencing of the largely methylated non-CGI CpGs. This may result in a less complete understanding of the epigenetic mechanisms in action, as lower density CpG regions have been suggested to play an important regulatory role (Skinner and Guerrero-Bosagna, 2014), but the balance between understanding and cost will depend on the study aims.

1.3.4.1.1 Mapping of bisulfite sequencing data

For the reads of bisulfite converted DNA to be analysed they first need to be mapped to the genome, in which the read location is found by comparison to the full DNA sequence, also known as a reference genome. Genomic alignment is made more computationally challenging in bisulfite sequencing compared to standard genomic sequencing, as both unmethylated C and T are both represented as T. To overcome this, the reads are compared to both an unmethylated and methylated genome, which adds to the computational load. Alignment will be more challenging still in repetitive regions.

Bespoke tools such as Bismark (Krueger and Andrews, 2011) and BS-seq (Huang, Huang and Chen, 2018) have been developed to handle these types of data (Krueger et al., 2012), however, in a recent study comparing RRBS and WGBS, alignment using the standard Bismark pipeline was still only ~75% or less (Beck, Ben Maamar and Skinner, 2021), and so lack of sequencing complexity remains a major issue for this approach (Laird, 2010) which will persist without longer reads.

Once reads are aligned to the genome, the read depth of each DNA methylation site can be quantified, which is defined as the number of reads that contain that specific DNA methylation site. From an individual read, which originates from a single fragment of DNA, a cytosine will be either methylated or unmethylated. Where there are multiple reads overlapping the same site, the proportion of DNA methylation can be calculated across the reads (**Figure 1.9**).

While the goal of WGBS is uniform coverage, some areas of the genome will be covered more deeply than others, due to the underlying genomic structure and somewhat stochastic nature of sequencing. Read depth will be higher at certain sites if an enrichment step has been carried out, e.g. in RRBS. As a result, read depth is not consistent across sites, but will generally follow a count distribution.

Sufficient read depth is essential for any sequencing study but is especially important in DNA methylation quantification; low read depth will reduce the sensitivity to accurately detect the proportion of DNA methylation at a site, as, for example, a site with only 2 reads can only detect DNA methylation in steps of 50%, at 0%, 50% and 100%. This has implications for subsequent analysis and means that the data requires careful QC prior to analysis, described in Section 1.3.5.1.

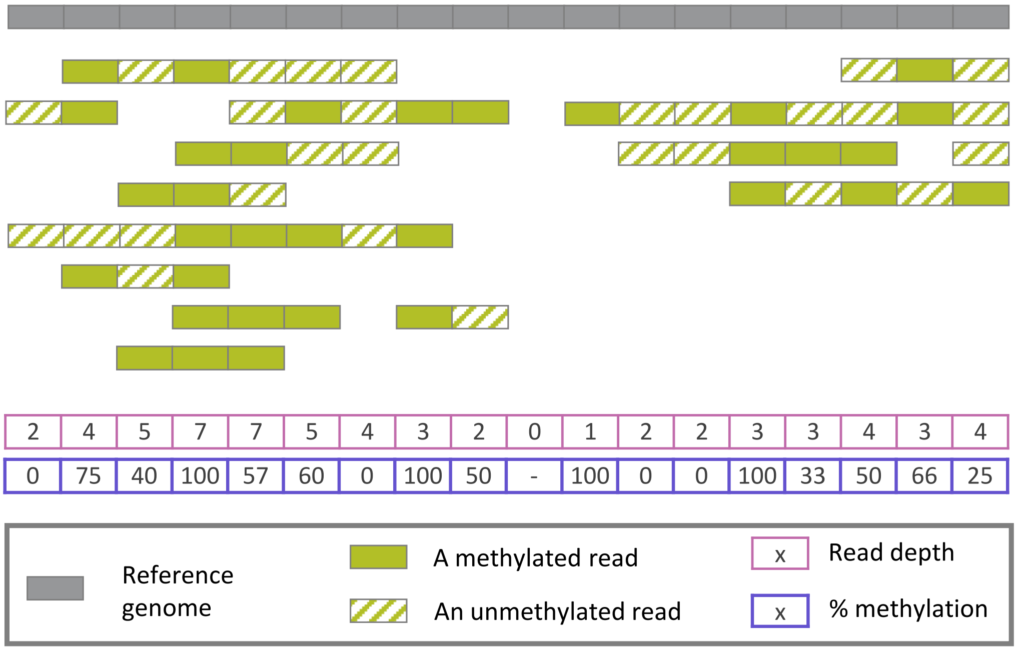
The resulting data contains a percentage or proportion of DNA methylation and the read depth per DNA methylation site. Across the genome, the distribution of DNA methylation is bimodal, with peaks at 100% and 0% methylated, and the distribution of read depths is a negative binomial distribution, with a high peak at the lowest read depth and a long tail to higher read depths.

1.3.4.2 Microarray based methods

Microarray based technologies, henceforth referred to as 'arrays', are assays that quantify DNA methylation at a pre-selected subset of DNA methylation sites. They are often utilised to compare DNA methylation levels between experimental groups at specific sites in humans but until recently have not been available for quantifying DNAm in other species. The arrays most frequently used for human DNA methylation studies are Illumina arrays (Illumina, n.d.[a]), of which there have been three iterations, with each respective version profiling a larger number of DNA methylation sites: Human Methylation 27K BeadChip (HM27) featuring 25,578 DNA methylation sites, Human Methylation 450K BeadChip (HM450) featuring 485,577 DNA methylation sites, and Infinium Methylation EPIC BeadChip (EPIC) featuring 866,836 DNA methylation sites (Pidsley et al., 2013). The maximum proportion of the genome covered by array platform (i.e. the EPIC) is 3.1% of CpGs (863,904 out of $\sim$28 million total across the human genome) (Zhou, Laird and Shen, 2016; Pidsley et al., 2016).

To profile DNA methylation at specific positions in the genome, complementary sequences of 50 bases, defined as probes, are utilised that each target DNA methylation site to be profiled (most of which but not all are CpGs). Each probe is identifiable by an additional 23 unique bases. Once the complementary bisulfite converted DNA has hybridised to the probe, a fluorescent marker is incorporated to estimate C/T conversion, that is, to differentiate between methylated and unmethylated sites (Bibikova et al., 2009; Pidsley et al., 2016).

There are two probe types that capture the methylation status of an individual site included on arrays manufactured by Illumina: Type I probes, which consist of two



2	4	5	7	7	5	4	3	2	0	1	2	2	3	3	4	3	4
0	75	40	100	57	60	0	100	50	-	100	0	0	100	33	50	66	25

Figure 1.9: **A diagram demonstrating DNA methylation quantification from bisulfite sequencing data.** The grey rectangles represent cytosines within the reference genome, at which the proportion of DNA methylation will vary. Each read sequenced (shown in green) will have undergone bisulfite conversion and as such can quantify DNA methylation status, with filled representing a methylated cytosine and dashed being unmethylated. The read view represents alignment to the genome, with some areas having higher read coverage than others. The total read depth for each nucleotide is shown in the pink box underneath, and the proportion of DNA methylation calculated by averaging across the reads is shown in the blue box.

separate probe sequences, one each for the methylated and unmethylated sequence to hybridise to, and Type II probes, which can characterise methylation using only one probe sequence to distinguish between methylated and unmethylated CpGs. This makes Type II probes more spatially efficient, however, they have reduced accuracy in CpG dense areas and therefore it is not always possible to design a type II probe for a specific position. Research suggests that Type II probes are also generally less accurate and reproducible than their Type I counterparts (Pidsley et al., 2016).

Both probe types utilise red and green fluorescence to quantify DNA methylation (**Figure 1.10**). Fluorescence is measured across the array using a laser, resulting in a signal intensity of red and green. For Type I probes, the methylated intensities (M) will be the intensity at a complementary methylated probe, and unmethylated intensity (U) will be the intensity at a complementary unmethylated probe. For Type II probes, M is intensity of green signal and U is the intensity of red signal. The ratio of M to M + U is the proportion of methylation at any one site, routinely referred to as β values, the formula for which is $\beta = M/(M + U + \alpha)$, meaning that β is bound between 0 and 1, with $\beta = 0$ representing an unmethylated DNA methylation site across cells profiled and $\beta = 1$ representing a fully methylated DNA methylation site. The addition of α, commonly set to 100, stabilises the β values in the scenario where both M and U are low (although this is largely a non-issue, as low intensity sites are typically excluded from analysis in QC). As in sequencing data, the resulting distribution of βs is bimodal, with peaks near 0 and 1. The probe types have different beta distributions (**Figure 1.11**) and as a result, require separate QC (see Section 1.3.5.2) (Pidsley et al., 2013; Dedeurwaerder et al., 2014).

1.3.5 Quality control of DNA methylation data

Given that the generation of DNA methylation data is highly complex and is influenced by various experimental factors, it is important that data is stringently filtered prior to analysis, as poor quality samples will introduce unwanted variation which will in turn reduce study power. The QC process is dependent on the method used to profile DNA methylation, as the resulting data structures differ. In this Section, the steps and methods

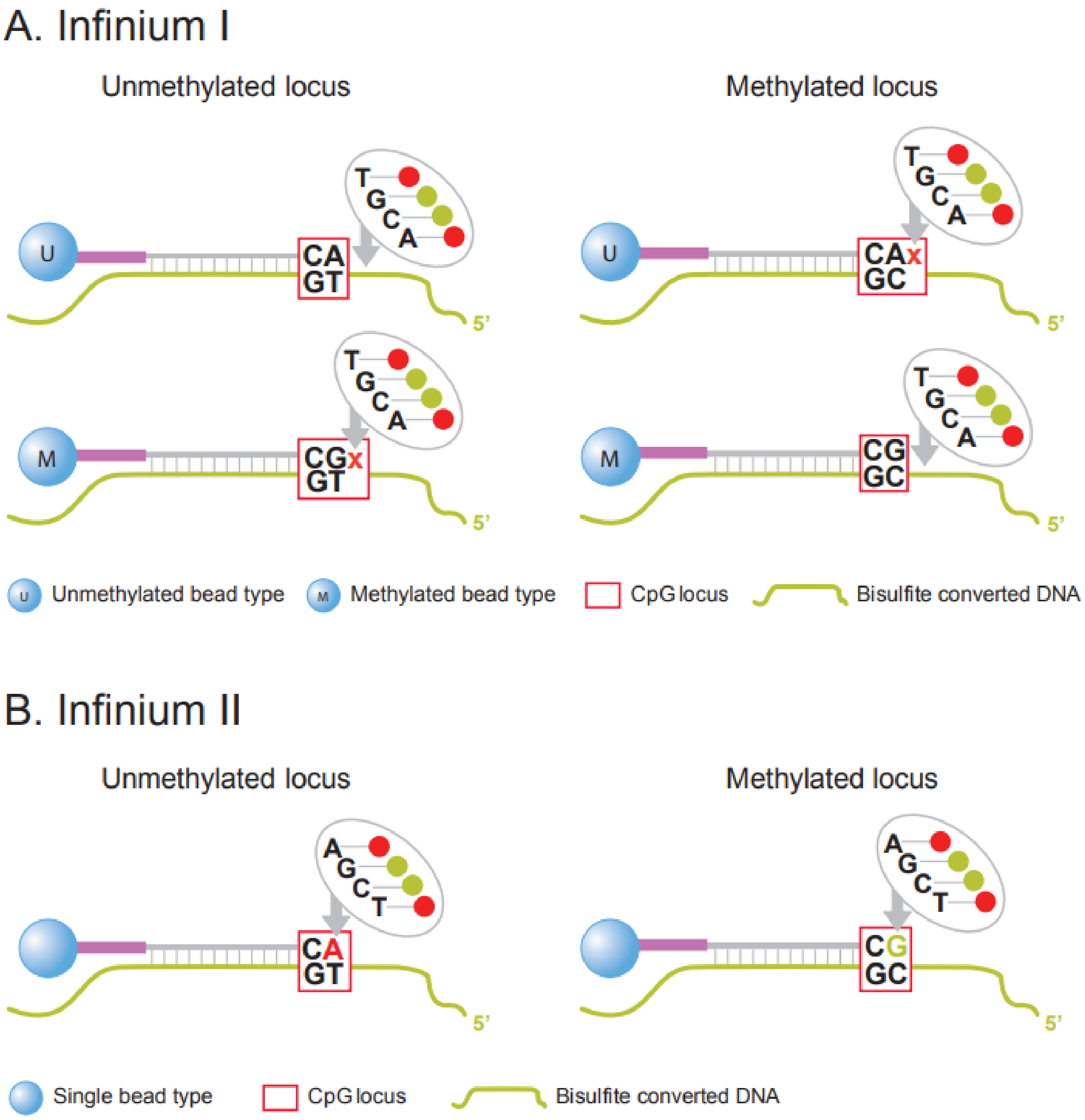


Figure 1.10: **A diagram explaining the distinction between Type I and II probes on the Illumina array sourced from (Illumina, n.d.[b]).** For Type I probes (labelled here as Infinium I), if the locus methylation and bead type do not match, there will be no fluorescent marker. In contrast, Type II probes (labelled here as Infinium II) will fluoresce regardless of methylation status, with the colour of the fluorescence signifying the methylation status; red for unmethylated, green for methylated.

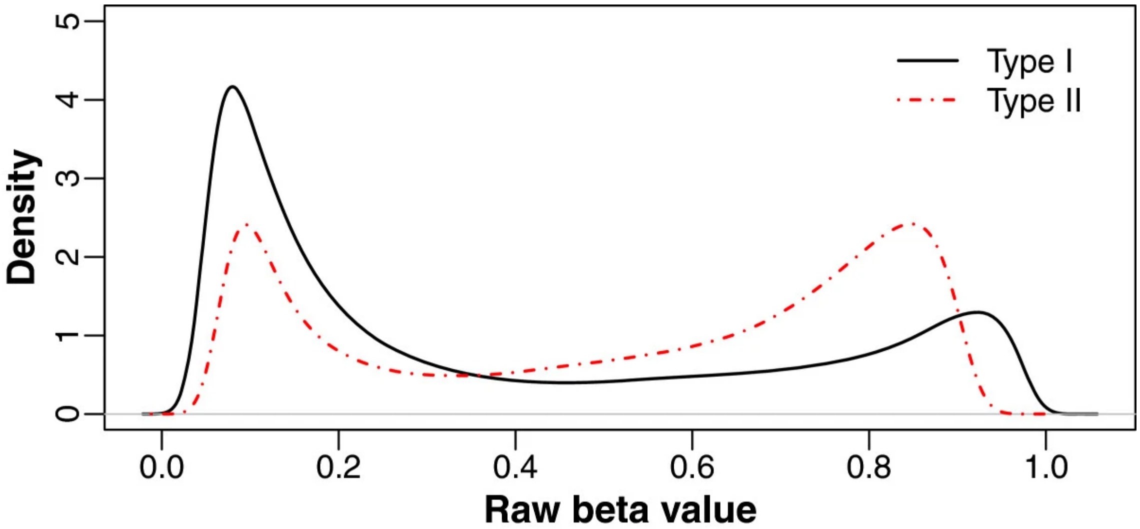


Figure 1.11: **An example of the density distribution of the β values profiled using Type I and II probes taken from Pidsley et al., 2013**.

generally considered for BS and array QC are briefly outlined.

1.3.5.1 Bisulfite sequencing data preprocessing methods

Initial steps of QC for BS data include assessing sequence quality and trimming reads, commonly carried out using FastQC (Andrews et al., 2010).

After that, a key consideration when carrying out QC of BS data is filtering by read depth. As described in Section 1.3.4.1, in BS low read depth at a DNA methylation site will reduce the sensitivity of average DNA methylation detectable. If being utilised for EWAS of complex disease, in which the DNA methylation differences between groups are commonly $<5\%$, low read depth will reduce the power to accurately detect differential DNA methylation. The read depth across an experiment will be determined by experimental design, and will be a trade off between power and cost of sequencing. For WGBS samples, the recommended sequencing depth per site is 15-30x (Ziller et al., 2015), depending on the magnitude of differential DNA methylation being tested for (see Section 1.3.7). Tools have been developed to estimate missing information across WGBS datasets, such as COMET (Libertini et al., 2016) and BSmooth (Hansen, Langmead and Irizarry, 2012), which impute DNA methylation values or smooth across genomic regions to reduce missingness. These tools are less applicable to RRBS data, in which DNA methylation sites are enriched for CGIs, resulting in non-random missingness with fewer sites profiled.

There has been limited to no assessment on how best to filter RRBS data, with studies applying varying read depth filtering thresholds, often arbitrarily selected, between 5-20 reads (Gu et al., 2010; Lutz et al., 2017; Stubbs et al., 2017; Kessler et al., 2018). Filtering data by read depth will increase the number of missing DNA methylation values and likely will result in sites with varying sample representation depending on individual level read depth. Some researchers resolve this by removing any site with missing data, however, this could result in the removal of true positives at sites that may have had sufficient power to detect differences between groups/across phenotypes. Chapter 2 aims to address this issue in BS data, exploring the relationship between read depth, effective sample size, and power and creating a novel tool utilising BS data simulation to allow users to calculate their filtering parameters for a two-group comparison in BS data to identify differentially methylated positions (DMPs).

1.3.5.2 Array data preprocessing methods

Large scale human EWAS generally utilise the Illumina array platform, and as such, specific QC and analysis pipelines have been developed. Two main R packages are used, which have common functionality: *minfi* (Jaffe and Irizarry, 2014), and *wateRmelon* (Pidsley et al., 2013), with others, such as *bigmelon* (Gorrie-Stone et al., 2019) and *meffil* (Min et al., 2018) developed to scale up previous QC pipelines to larger datasets at higher speeds, making them ideal methods for EWAS with more samples. These packages also include data analysis tools, meaning the entire process from raw data to results generation could viably be conducted in a single replicable R pipeline. This will aid in overall reproducibility (see Section 1.5), simplifying analysis and allowing more consistent methods between studies.

Specific steps of the QC pipeline most commonly used in the analysis of Illumina DNAm array data are described in Section 3.10.2. For example, array methods utilise internal control probes to ensure that the detected DNA methylation values are as expected for accurate DNA methylation quantification.

1.3.6 Considerations when choosing a platform

No DNA methylation quantification approach is perfect, and when choosing a platform the decision will ultimately be driven by the aims of a study and available resources. General considerations to be made are summarised in **Figure 1.12**. The consistency of sites profiled on the Illumina array make it highly useful for identifying site specific DNA methylation differences across a phenotype, as well as for use in biomarker development and predictive tools that utilise DNA methylation. Non-human arrays are not commercially available for many species, although they exist (a mouse array by Arneson et al., 2021 and a mammalian array by Horvath et al., 2021), but their development will not be cost effective unless they are widely used. This means that, for most researchers, it will not be viable to carry out a non-human EWAS using arrays. The Illumina array covers at most 3% of CpGs in the human genome, and sites not on the array will not be profiled, so if a wider range of sites are of importance in the given study aims, BS methods may be more suitable. BS methods can profile a larger proportion of the genome, however, doing so is costly especially at the sequencing depth required and depending on the study aim, many sites covered may not be of interest.

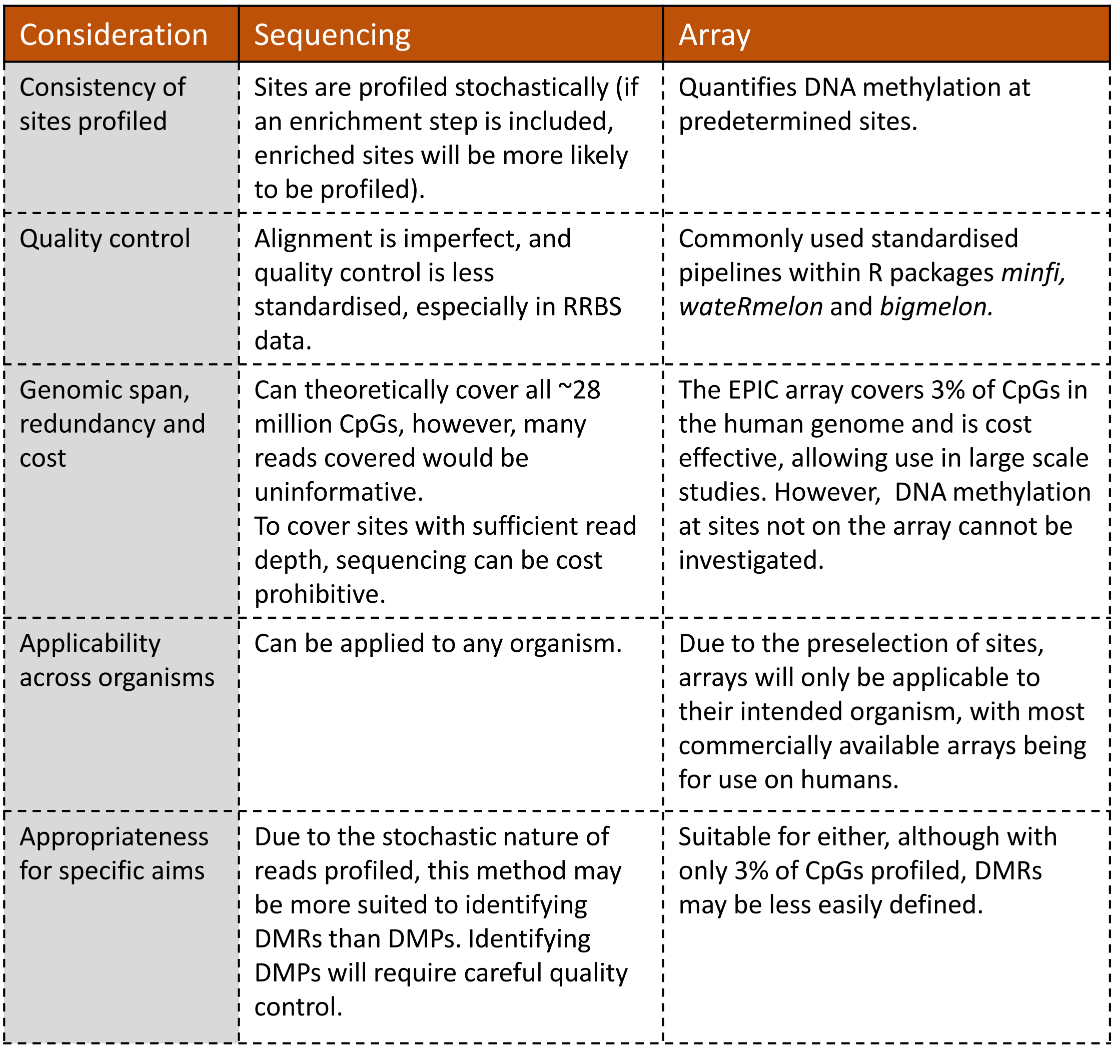
Of note, both methods described here rely on bisulfite conversion, which does not distinguish between 5mC and 5hmC, as both modifications prevent C conversion to U (Skvortsova et al., 2017). This may have implications for the interpretation of DNA methylation studies, as results may be confounded by the aggregated effect of multiple DNA modifications, although the extent may be limited due to the paucity of 5hmC across most tissue types.

1.3.7 Metrics for identifying differential DNA methylation

There are two main metrics for identifying differential DNA methylation between groups, summarised in **Figure 1.7 5A**, DMPs and differentially methylated regions (DMRs). The chosen metric(s) will depend on the study hypothesis and method used to profile DNA methylation.

The majority of EWAS first and foremost aim to identify DMPs across the DNA methylation sites for which DNA methylation has been quantified. Most commonly, linear regression is applied to test the association between site specific DNA methylation and some phenotype or disease status of interest (Mansell et al., 2019) (covariates are also routinely included in the model to minimise potential biases, as described in Section 1.3.8).

DNA methylation is not independent across DNA methylation sites; DNA methylation levels at sites within 1000-2000 bases from each other are correlated (Zhang et al., 2015). As such, DNA methylation is also investigated in the context of genomic regions. DMRs vary in exact definition, but can generally be described as a continuous genomic region across which DNA methylation differs between/across some phenotype of interest (Peters et al., 2015). Several methods for identifying DMRs exist, including but not limited to *comb-p* (Pedersen et al., 2012), *bumphunter* (Rafael, Aryee and Hansen,



Consideration	Sequencing	Array
Consistency of sites profiled	Sites are profiled stochastically (if an enrichment step is included, enriched sites will be more likely to be profiled).	Quantifies DNA methylation at predetermined sites.
Quality control	Alignment is imperfect, and quality control is less standardised, especially in RRBS data.	Commonly used standardised pipelines within R packages *minfi, wateRmelon* and *bigmelon.*
Genomic span, redundancy and cost	Can theoretically cover all ~28 million CpGs, however, many reads covered would be uninformative. To cover sites with sufficient read depth, sequencing can be cost prohibitive.	The EPIC array covers 3% of CpGs in the human genome and is cost effective, allowing use in large scale studies. However, DNA methylation at sites not on the array cannot be investigated.
Applicability across organisms	Can be applied to any organism.	Due to the preselection of sites, arrays will only be applicable to their intended organism, with most commercially available arrays being for use on humans.
Appropriateness for specific aims	Due to the stochastic nature of reads profiled, this method may be more suited to identifying DMRs than DMPs. Identifying DMPs will require careful quality control.	Suitable for either, although with only 3% of CpGs profiled, DMRs may be less easily defined.

Figure 1.12: **An overview of the considerations to be made when deciding between sequencing and array based platforms for DNA methylation quantification.**

2017), *DMRcate* (Peters et al., 2015), *Dmrff* Suderman et al., n.d., and *GlobalP* (Lent et al., 2018). Some of these take DMP results and look for continuous signals (which may or may not be due to correlation between the sites), where others first identifying regions using the DNA methylation matrix, then testing for DMRs. The maximum (and minimum) size of region across which differential DNA methylation is sought varies from method to method, with *DMRcate* counting a DMR as any two differentially methylated sites within 1000 base pairs, whereas *comb-p* searches only sites within a 200 base pair window, and using *bumphunter* the window is user defined and dependent on the application at hand.

1.3.8 Adjusting for confounding variables

As discussed earlier, DNA methylation is highly associated with cell and tissue type, but is further influenced by a number of other sample characteristics. When designing a study specifically seeking to test DNA methylation in association with a trait of interest, careful consideration must be taken to account for potential sources of variation within and across sample groups. There are a number of variables, which, if correlated with the phenotype of interest, either mechanistically or by chance will induce differential DNA methylation. These are defined as confounders, the most important of which are shown in **Figure 1.7 5B**.

Ideally, compared groups in an EWAS are balanced for potential covariates, however, where this is not the case, confounding variables and cryptic stratification within a population may result in false positives, i.e. significantly differential DNA methylation that does not relate to the phenotype being investigated. Where balanced study design is not possible, these identified sources of unwanted variation can be included as covariates in the statistical model used to identify differential DNA methylation, although their inclusion does not necessarily mean that the influence of the confounding variable has been eliminated.

The following Sections describe the main potential confounders that are commonly adjusted for in EWAS: experimental batch effects, age, sex, environmental exposures, genetics, and cellular heterogeneity.

1.3.8.1 Batch effects

Batch effects are defined as systematic variation in DNA methylation data attributable to technical sources in sample processing, which are especially pertinent where a study contains a large number of samples that cannot be processed together. These include, but are not limited to: operator effects, array sample position, reagent lots, laboratory conditions and time of experimentation. In general, batch effects are discussed more when it comes to the array, in which multiple samples are organised in a grid formation on each chip to be profiled. Batch effects can be observed both within individual chips and across chips (Buhule et al., 2014; Price and Robinson, 2018). A lack of randomisation during data generation could lead to these batch effects confounding the results (Mill and Heijmans, 2013) and so the best solution is careful study design (Price and Robinson, 2018).

The commonly applied array QC step, normalisation, can combat batch effects, although the effectiveness of any method will depend on the data and study design in question (Sun et al., 2011). Normalisation aims to reduce batch effects and maximise the sensitivity to detect true differences between experimental groups (Pidsley et al., 2013) by making the distributions of array intensities or βs the same across samples, most commonly matching the distribution of the data quantiles.

1.3.8.2 Age

Aging is a complex physiological process characterised by a progressive loss of tissue functionality and an increased risk of death (Ciccarone et al., 2018). As described in Section 1.2.4, DNA methylation plays an important role in mammalian development (Li, Bestor and Jaenisch, 1992), and is an important driver of inter-individual variation across the life course (Fraga et al., 2005). Many DNA methylation changes occur during development; Spiers et al., 2015 investigated fetal brain development, observing that 7% of DNA methylation sites tested (out of $\sim$400000) significantly changed between 23 - 184 days post conception, with most becoming hypomethylated, with differential DNA methylation as high as 50% (Spiers et al., 2015). A longitudinal comparison of blood DNA methylation between birth and adolescence identified differential DNA methylation at over half of the CpGs investigated (53% of ~450000 sites), with 36% of CpGs becoming more hypomethylated. The largest absolute change was an over 60% decrease in DNA methylation, with changes found not to be randomly distributed across the genome; the loss of DNA methylation being enriched in gene bodies and enhancers, and gain enriched in promoter regions (Mulder et al., 2021). While the DNA methylation changes that occur in later development do not necessarily have as large a magnitude as those in early development, they have been observed consistently enough across individuals for differential DNA methylation to be considered a hallmark of aging (López-Otín et al., 2013).

The theory of epigenetic aging is comprised of two parts: epigenetic drift and locus specific differential DNA methylation (Ciccarone et al., 2018). Epigenetic drift is described as a gradual stochastic increase or decrease in DNA methylation that occurs over time (Issa, 2014; Fraga et al., 2005). Given the random nature of the change it is thought be be at least in part caused by imperfect mitotic inheritance of DNA methylation (Ming et al., 2020). This leads to diverging patterns of DNA methylation between individuals, even in genetically identical twins (Fraga et al., 2005). Paradoxically, at older age, epigenetic drift converges resulting in reduced epigenetic variability, observed in very old identical twins (Talens et al., 2012). Locus specific changes, on the other hand, are highly reproducible across individuals. They are so reproducible, in fact, that they can be used to reliably predict age across many tissue types and species, the models for which are known as 'epigenetic clocks' (Hannum et al., 2013; Horvath, 2013; Shireby et al., 2020; Steg et al., 2021; Horvath et al., 2021; Raj et al., 2021; Stubbs et al., 2017; Thompson et al., n.d.).

Together, the differential DNA methylation acquired across the life course results in variation both within individuals (depending on the age at which they are profiled) and across individuals of the same age and can be observed across tissue types (Horvath et al., 2015). Therefore, if age is not matched across groups in EWAS, it will need to be adjusted for within the model used.

1.3.8.3 Sex

Due to the essential role that DNA methylation plays in X chromosome inactivation (Riggs, 1975), large DNA methylation differences can be seen between males and females across the X chromosome, with higher global levels of DNA methylation in females (McCarthy et al., 2014).

It's not only the X chromosome at which sex specific differential DNA methylation has been observed, however; approximately 5% of autosomal CpGs compared using the HM27 have been observed to have sex specific DNA methylation differences (McCarthy et al., 2014; Liu et al., 2010). A more recent study looking at sex differences using the HM450 identified 5762 significatly differentially methylated autosomal CpG sites with absolute differences up to 46.1% (Davegårdh et al., 2019). Differences have been found to be tissue dependent and can coincide with differential DNA methylation that relates to other phenotypes, however, methods such as ComBat (Leek et al., 2012), are highly successful in adjusting for sex effects (McCarthy et al., 2014).

1.3.8.4 Environment

DNA methylation is said to be one of the mechanisms by which environmental factors can influence gene expression. Evidence for environment driven differential DNA methylation can be studied using monozygotic twins, in which genetic code is identical. Fraga et al., 2005 observed DNA methylation patterns to be more divergent in twins that did not share the same habits or environment (Fraga et al., 2005).

Many environmental cues have been associated with differential DNA methylation, including maternal effects (maternal smoking (Joubert et al., 2016; Richmond et al., 2015), maternal weight (Sharp et al., 2015)), lifestyle factors (e.g. exercise (Rönn et al., 2013), alcohol consumption (Liu et al., 2018), smoking status (Zeilinger et al., 2013), hair dye (Langevin et al., 2011), educational attainment (Dongen et al., 2018), loneliness (Phillips et al., 2019), socioeconomic position (Hughes et al., 2018)), and exposures (viral infections (Schäfer and Baric, 2017), UV radiation (Oliveira, Souza and Coêlho, 2020)).

An individuals environment will influence their individual-specific methylome and would need to be adjusted for if not carefully balanced in study design. This is especially challenging in a scenario for which a higher occurrence of some environmental or behavioural factor is more likely with disease, for example, the increased prevalence of smoking in SZ patients, which can confound findings of a SZ EWAS (Hannon et al., 2021a). Furthermore, environmental factors can be more challenging to measure compared to biological factors, such as age, or may be unknown and as such, more difficult to control for.

1.3.8.5 Genetics

There is much evidence for the partial genetic control of epigenetics; A twin study across adipose tissue using mono and dizygotic twins showed that as much as 37% of DNA methylation variance could be attributed to genetic factors (Grundberg et al., 2013). In fact, the heritability of DNA methylation across generations may be explained by the contribution of genetic variation, with $\sim$20% of individual DNA methylation variation being due to differences in non-coding DNA sequence (McRae et al., 2014). Furthermore, studies across multiple tissues, including whole blood and brain, have demonstrated the contribution of genetic variation in DNA methylation levels (Gaunt et al., 2016; Olsson et al., 2014; Drong et al., 2013; Gamazon et al., 2012; Gibbs et al., 2010; Hannon et al., 2015b; Hannon et al., 2015a). Similarly, in mouse models, in which genetics can be carefully controlled, the majority of differential DNA methylation observed has been attributed to genetic differences (Orozco et al., 2014). Genetic effects on DNA methylation have primarily been shown to be *cis*, that is, local to a single nucleotide polymorphism (SNP) ($\leq$ 500kb) in the DNA sequence (Hannon et al., 2018) but can also be *trans*, that is, further away from a SNP (Bonder et al., 2016).

The genetic effect on DNA methylation can also be observed across ethnic backgrounds, where population specific DNA methylation patterns have been detected at over a third of all genes, suggesting extensive divergence is genetic control of DNA methylation. This may largely be due to differences in allele frequency, although environmental interactions undoubtedly also contribute (Fraser et al., 2012).

1.3.8.6 Cellular heterogeneity

As described in Section 1.2.4, DNA methylation profiles are cell type specific and, when comparing between them, the magnitude of DNA methylation differences uncovered are commonly upwards of 60%. The DNA methylation differences uncovered in non-cancer complex disease EWAS mostly have magnitude of <5%, and so for studies performed in bulk tissue, of which most are, if the cell proportions differ between cases and controls, it will most likely result in false positives. A graphical representation of this effect can be seen in **Figure 1.13**, in which a tissue is comprised of three cell types, each of which containing a unique DNA methylation pattern. In the **Figure**, when comparing two bulk tissue samples that comprise of different proportions of each cell type, differential methylation can be observed despite there being no other difference between the samples. This can be especially detrimental if the phenotype of interest correlates with a change in cell proportions. For example, when investigating AD in post-mortem brain tissue, the neuronal loss observed with AD progression, as well as a shift in immune cell proportions (Prinz and Priller, 2017), can confound the results meaning it is unclear if any shift in DNA methylation uncovered is due to within cell type changes associated with AD, or just due to a change in cellular proportions.

To adjust for differences in cellular composition of bulk tissues type, cellular proportions can be included in the statistical model as covariates. The proportions can be calculated empirically, or estimated using cellular deconvolution algorithms, the methods for which are detailed in Section 1.4. Assessing the reliability of cellular deconvolution prediction is the overarching aim of Chapter 3, and generating an improved model for the deconvolution of PFC is the main aim of Chapter 4.

1.4 Cell type deconvolution

As described above, cellular heterogeneity can be a major confounder of EWAS carried out in bulk tissue, such as brain or blood, for which the cellular proportions can be variable from individual to individual (blood and brain composition are described in Section 3.1.2). To get around the issue of cellular heterogeneity across a study, some

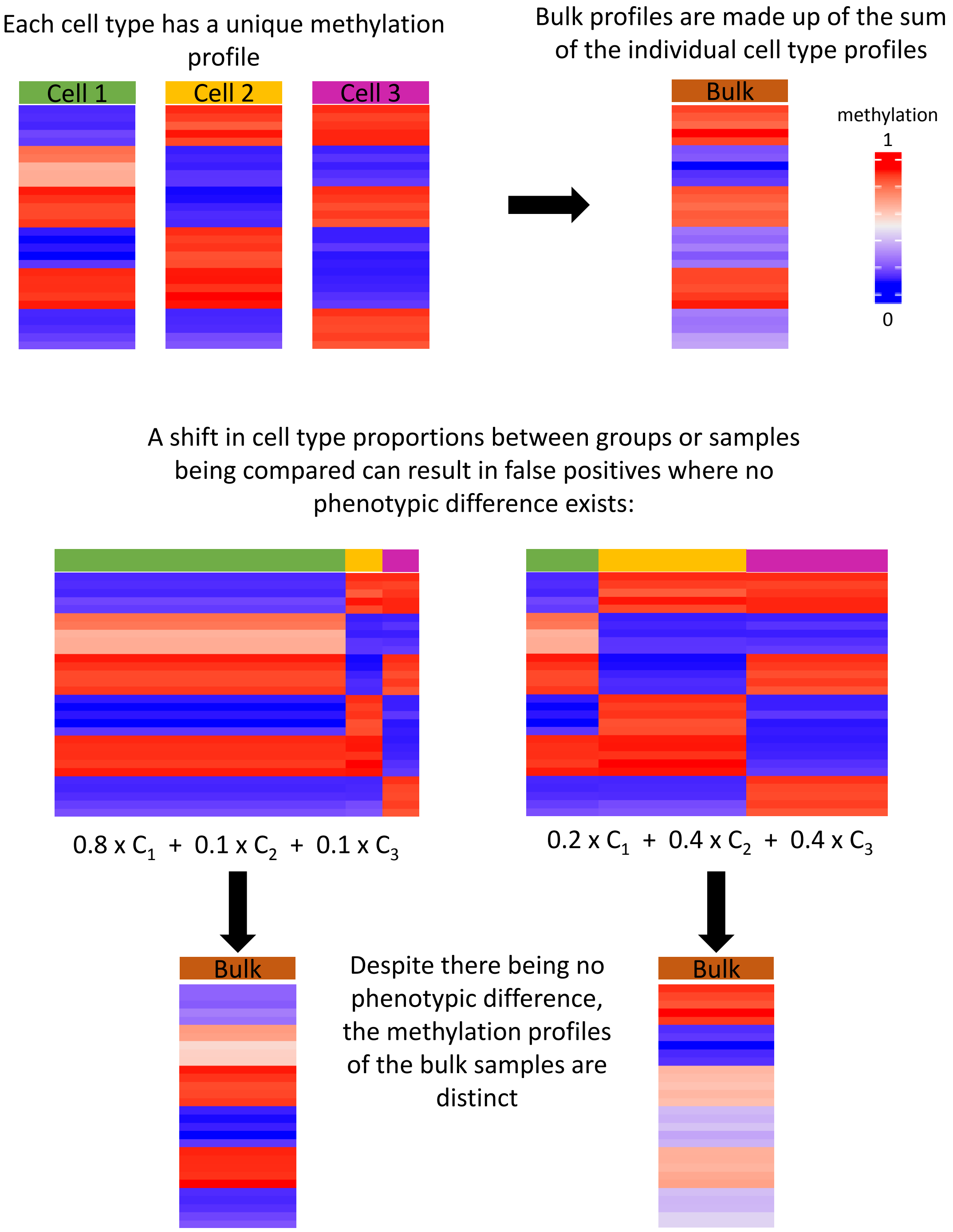

Each cell type has a unique methylation profile
Cell 1
Cell 2
Cell 3
Bulk profiles are made up of the sum of the individual cell type profiles
Bulk
methylation
1
0
A shift in cell type proportions between groups or samples being compared can result in false positives where no phenotypic difference exists:
0.8 x C1 + 0.1 x C2 + 0.1 x C3
0.2 x C1 + 0.4 x C2 + 0.4 x C3
Bulk
Despite there being no phenotypic difference, the methylation profiles of the bulk samples are distinct
Bulk

Figure 1.13: A diagram showing how cell type heterogeneity can confound DNA methylation comparisons. This example is a simplification which assumes that the bulk tissue is made up of three cell types, and uses a subset of 30 CpGs that differ with cell type to demonstrate the point. The heatmaps represent DNA methylation level, where red is fully methylated and blue is unmethylated, and each row is a CpG, each column a sample or proportion of a sample for bulk or purified data, respectively. C_i represent the ith cell type profile.

EWAS are performed on purified cellular populations (Mendizabal et al., 2019; Tulloch et al., 2018) (as described in Section 1.3.3), however, purification methods (described in Section 4.1.1.1) are not always feasible or scalable, and therefore computational solutions to estimate cellular proportions from bulk tissue DNA methylation profiles have been developed. Estimates can subsequently be used in regression analysis to adjust for cellular composition (as with other confounders, their inclusion in the model may not mean that their influence has been eliminated). Cellular deconvolution algorithms come in two classes, reference based and reference free (although semi-reference free methods also exist,i.e. a reference is used after deconvolution) (see **Table 1.1** for a summary of deconvolution algorithms).

1.4.1 Reference based deconvolution

Reference based deconvolution algorithms require a DNA methylation reference dataset comprised of the DNA methylation profiles of the purified cell types within the tissue to be deconvoluted. As such the applicability of the models generated using these algorithms will be dependent on the availability of a high quality reference dataset, and each model will be tissue specific.

Reference data exists for a number of tissue types, including cortex, containing two purified populations (Guintivano, Aryee and Kaminsky, 2013), whole blood, containing six purified populations (Reinius et al., 2012; Jaffe, 2018), cord blood, containing seven purified populations (Bakulski et al., 2016a), and saliva, containing two purified populations (Middleton et al., 2020).

In general, the development of a reference based deconvolution model can be summarised into two stages. First, is the data generation stage, in which sourcing and QC of the reference DNA methylation profiles of purified cell types is carried out. Second, is the model generation stage, in which a reference based deconvolution algorithm is used to generate the deconvolution model using the reference data, which can predict the cell type proportions of an input sample with unknown composition.

Reference based deconvolution relies on the following relationship between bulk and cell type specific DNA methylation profiles:

Table 1.1: **A table summarising deconvolution algorithms, adapted from Teschendorff and Zheng, 2017.** Abbreviations: constrained projection (CP), deep learning (DL), iterative method (IM), independent surrogate variable analysis (ISVA), linear regression (LR), linear slope model (LSM), quadratic programming (QP), robust partial correlation (RPC), removing unwanted variance (RUV), surrogate variable analysis (SVA), support vector regressions (SVR).

Statistical algorithm	Inference paradigm	Output	Tissues successfully applied	Adjusts for other confounders	Availability	Citation
Houseman's CP/QP	Reference based	Cell-type fractions	Whole blood, PBMC, cord blood, breast	No	R	Houseman et al., 2012; Jaffe and Irizarry, 2014
CIBERSORT SVR	Reference based	Cell-type fractions + DMPs	Whole blood, PBMC, breast	No	Web-based tool, JAVA, R	Newman et al., 2015; Teschendorff et al., n.d.
EpiDISH RPC	Reference based	Cell-type fractions + DMPs	Whole blood, PBMC, breast	No	R	Newman et al., 2015; Teschendorff et al., n.d.
CETS LSM	Reference based	Cell-type fractions	Cortex	No	R	(Guintivano, Aryee and Kaminsky, 2013)
IDOL IM/QC	Reference based	Cell-type fractions	Whole blood	No	R	Koestler et al., 2016
RUV	Semi-reference free	DMPs	Whole blood, PBMC	No, unless control set is modified.	R	Gagnon-Bartsch and Speed, 2012
BayesCCE	Semi-reference free	Cell-type fractions	Any	No	R	Rahmani et al., 2018

Statistical algorithm	Inference paradigm	Output	Tissues successfully applied	Adjusts for other confounders	Availability	Citation
MethylNet DL	Semi-reference free	Cell-type fractions + DMPs	Any	Yes	Python	Levy et al., 2020
Edec NMF-CP/QP	Semi-reference free	Cell-type fractions + DMPs	Complex tumor tissue	No	R	Onuchic et al., 2016
ISVA	Reference free, supervised	DMPs	Any	Yes	R	Teschendorff, Zhuang and Widschwendter, 2011
Tsisal	Reference free (with reference based cell labelling)	Cell-type fractions	Any	Not fully assessed	R	Zhang, Wu and Li, 2021
RefFreeEWAS	Reference free, supervised	DMPs	Any	Yes, if other confounders carry high variance.	R	Houseman, Molitor and Marsit, 2014
SVA	Reference free, supervised	DMPs	Any	Yes	R	Leek and Storey, 2007; Leek et al., 2012
EWASher	Reference free	DMPs	Whole blood	Yes, if other confounders carry high variance.	Python, R	Zou et al., 2014
ReFACTor	Reference free	DMPs	Whole blood	Yes, if other confounders carry high variance.	Python, R	Rahmani et al., 2016

Statistical algorithm	Inference paradigm	Output	Tissues successfully applied	Adjusts for other confounders	Availability	Citation
RefFreeCellMix NMF–CP/QP	Reference free	Cell-type fractions + DMPs	Any	Not fully assessed	R	Houseman et al., 2016
MeDeCom NMF	Reference free	Cell-type fractions + DMPs	Any	Not fully assessed	R and web-based tool	Lutsik et al., 2017

$$M_{BULK} = \sum_{i=1}^{n} P_i M_{CT_i} \tag{1.1}$$

where:

- M_{BULK} is a matrix with one column containing the genome wide DNA methylation values for a bulk tissue sample, each row of which contains the DNA methylation value at a profiled DNA methylation site.
- i is the index for cell types within the tissue, where $i \in [1, n]$ and n is the total number of cell types
- P_i is ith value of vector $\underline{P}$ containing cell proportions for all cell types in the bulk sample
- M_{CT_i} is the ith column within matrix M_{CT}, containing the genome wide DNA methylation values for cell type i at the same DNA methylation sites as M_{BULK}

Equation 1.1 has constraint:

$$\sum_{i=1}^{n} P_i \leq 1 \tag{1.2}$$

Reference based deconvolution algorithms solve the equation 1.1 for $\underline{P}$, with each algorithm using a different approach, such as CPQP, SVR, RPC, and LR (**Table 1.1**). The algorithm most commonly used is Houseman's CPQP method, henceforth referred to as Houseman's algorithm. It is commonly used due to it's easy applicability through QC R packages *minfi* and *wateRmelon*, and is one of the go to algorithms for predicting cellular proportions of brain or blood. The specific steps used in deconvolution are described in Section 3.1.1.

Reference based models return cellular proportion estimates that sum close to one (to allow for cell types that may be absent from the reference data), returning the closest possible solution to the above equation, even when not biologically meaningful (for example, if using a blood model to deconvolute a non-blood tissue). The accuracy of reference based deconvolution cannot be easily assessed without datasets in which cellular composition is already known. Furthermore, even when cellular composition is empirically calculated or cell purification methods have been utilised, techniques may not be completely accurate. To allow for the assessment of deconvolution accuracy, Chapter 3 focuses on developing and validating a reference based deconvolution error metric, named Cetygo (Cell type goodness).

Even with the optimal deconvolution algorithm, reference based deconvolution models generated can only be as good as the reference data in which they were generated. The current standard for cortical deconvolution contains only two cell types, neuronal and non-neuronal, due to the challenges of brain tissue purification. Finer granularity, that is, the inclusion of more specific cell types, would allow for more comprehensive cell type predictions. To that end, Chapter 4 utilises a novel sorted reference dataset to generate a three cell type PFC deconvolution model.

1.4.2 Reference free deconvolution

Reference free deconvolution, can be applied to tissues where no relevant panel of reference data exists. In general, unsupervised reference free methods adjust for unwanted variation within the data, assuming that cell type heterogeneity will be a primary source of this unexplained variance (Houseman, Molitor and Marsit, 2014; Zou et al., 2014). However, it is not guaranteed that cell type composition will always be the main source of variation, for example, in cancer datasets, and where this assumption is not met these methods would lack effectiveness for the intended purpose. To avoid the removal of variance of interest, supervised methods such as surrogate variable analysis (Leek et al., 2012; Teschendorff, Zhuang and Widschwendter, 2011) allow users to input phenotype data which reduces the removal of biological signal. In general, reference based methods more effectively account for cellular composition (Teschendorff and Zheng, 2017), and so where a reference dataset is available, it is preferable to use it.

1.5 Reproducibility

Science has a reproducibility issue; according to Freedman, Cockburn and Simcoe, 2015, at least 50% of pre-clinical life science research findings in the United States (US) are not reproducible, defined as the ability to replicate the same results using the same materials and methods (Freedman, Cockburn and Simcoe, 2015). Irreproducibility can have important implications. Firstly, it can result in wasted resources that could have been more effectively used, for example, at least US$28 billion are spent on science that cannot be replicated in the US alone (Freedman, Cockburn and Simcoe, 2015). Furthermore, downplaying scientific uncertainty erodes trust in science in the long run (Vazire, 2017; Kreps and Kriner, 2020). Evidence of this erosion can be seen in the relationship between academia and the pharmaceutical industry, where pharmaceutical companies now typically run in-house validations of potential drug targets published by academic researchers to ensure validity (Prinz, Schlange and Asadullah, 2011; Aschwanden, 2015; Jasny et al., 2017), as the methodologies applied are not always sufficient or transparent enough. Not only does this further increase cost, but the lack of reproducibility also creates an additional barrier to the discovery of drugs that may improve the lives of millions. Public trust in science is also highly important, as seen in the COVID-19 pandemic. A lack of transparency in research methodology and limitations can overstate the importance of findings which, if later shown to be false, reduce trust in the field as a whole (Kreps and Kriner, 2020). This may have long term ramifications on policy and scientific funding, as well as the uptake of potentially lifesaving treatments such as vaccines. Scientists trust in research also matters, as each study builds on the findings of a myriad of others; irreproducible science degrades this knowledge base (Lushington and Chaguturu, 2016).

In the context of EWAS, there are two main mechanisms of irreproducibility: a lack of clear methodology, resulting in results that cannot be replicated even in the same dataset, and the inability to replicate in a separate study, which may be due to inadequate study design, analysis or sample variation leading to false positive or false negative results. To address the first mechanism, there has been an increased drive for reproducible and open science practices, which includes greater transparency in the reporting of methodologies. To validate findings and address the second mechanism, DNA methylation cohorts are often meta-analysed to assess consistency of results and demonstrate reproducibility (Smith et al., 2021; Policicchio et al., 2020b; Joubert et al., 2016; Hannon et al., 2021a). Potential drivers of irreproducibility in EWAS still exist, however, and include insufficient study power (Section 1.3.1), which may be driven by small sample size and or low read depth (in BS studies), and inadequate adjustment for study confounders causing spurious results (Section 1.3.8), including cellular heterogeneity. The general theme of this thesis is to develop tools that will enable users to improve the reproducibility of EWAS, either allowing for the careful QC of BS data (Chapter 2), assessing the accuracy of reference based cellular deconvolution (Chapter 3), or allowing users to deconvolute cortex tissue with more granularity (Chapter 4). Each of these aims is described in more detail in Section 1.6.

Reproducibility does not relate only to the replicability of results but also to general practices such as open science, in which research is shared more transparently. To that end, the completed R scripts developed for the research within this thesis have been made publicly available at https://github.com/ds420/.

1.6 Thesis aims

The overarching aim of this thesis is to develop novel methods and computational tools that enhance existing preprocessing pipelines for DNA methylation data and provide confidence in downstream analyses that arise from that dataset.

The quantification of DNA methylation is a routine experiment performed in many large epidemiological cohorts and is used to address a broad range of research questions in health and disease. Methodologies to do so are generally sound, however, where aberrant DNA methylation profiles occur due to technical or biological reasons, it is important to address these issues prior to analysis. As such, the main aims of this thesis are to establish resources for researchers to assess study power (Chapter 2) and commonly used covariate quality (Chapter 3), allowing for more appropriate dataset specific analysis decisions, and to to expand available the reference dataset for epigenetic cell deconvolution in the PFC for more complete deconvolution (Chapter 4).

The main objectives of this PhD are divided into the following Chapters:

- **Chapter 2** aims to investigate the relationship between power and the properties of a BS dataset, including read depth and sample size, focusing on how each influence the ability to perform EWAS. Leveraging these findings and data simulation, this chapter aims to guide studies on appropriate QC and develop a publicly available tool for calculating the optimum data filtering for BS data.
- **Chapter 3** aims to establish a methodology for quantifying the accuracy of reference based deconvolution of DNA methylation array data, Cetygo. The Chapter aims to profile it's performance in order to provide guidelines on it's utility in epidemiological studies.
- **Chapter 4** aims to expand existing deconvolution models for PFC tissue such that they can estimate the cellular proportions of three distinct cell types. The Chapter aims to assess model performance and applicability across a range of datasets.

2. Characterising the properties of bisulfite sequencing data: maximizing power and sensitivity to identify between-group differences in DNA methylation

This Chapter is presented in the form of a peer reviewed manuscript that has been published in BMC Genomics (Seiler Vellame et al., 2021). Supplementary Figures are included at the end of the Chapter, after main manuscript. Supplementary Tables (Additional File 2 in the published version) can be found in Section 7, **Table 7.1**.

2.1 Background and publication aims

Bisulfite sequencing (BS) is a method for quantifying DNA methylation, where the bisulfite conversion of DNA changes unmethylated cytosine (C) bases into uracil (U) and leaves methylated C intact (see **Figure 1.7 3A**). Short read sequencing is then utilised to determine the sequence of the sample by generating reads that represent fragments of DNA that are aligned to a reference genome to determine their genomic location (see Section 1.3.4.1 for details). The DNA methylation of each site in a read will be binary, originating from an individual cell. DNA methylation is commonly measured across multiple cells (i.e. a tissue) at a time and as such, to estimate the DNA methylation of a DNA methylation site across a population of cells the proportion is calculated across reads.

BS is commonly used for the quantification of DNA methylation in epidemiological research (Gertz et al., 2011; Bundo et al., 2020; Tang et al., 2018; Fernández-Santiago et al., 2019; Rizzardi et al., 2019b). The method is often valuable for DNA methylation studies as, in contrast to the array-based alternative, it is applicable across species and profiles a larger proportion of the genome. Whole genome bisulfite sequencing (WGBS) has the widest genomic coverage, with the potential to quantify DNA methylation at all ∼28 million CpGs in the human genome (Cokus et al., 2008; Lister et al., 2009; Laurent et al., 2010; Urich et al., 2015). However, it can be cost prohibitive for many studies and is, as the following analyses show, potentially not currently optimal for large epidemiological analyses. Furthermore, as DNA methylation studies are mainly only interested in a subset of Cs (and more specifically CpG sites) where DNA methylation is variable, a high number of WGBS reads are uninformative. Reduced representation bisulfite sequencing (RRBS), in contrast, involves a target enrichment step using the methylation-insensitive enzyme MspI to target CpG-rich regions of the genome (Gu et al., 2011) prior to bisulfite conversion, increasing the proportion of informative sequencing reads.

Despite the common usage of BS methods, little empirical work has been done to investigate how the statistical properties of sequencing data influences how amenable they are for addressing epidemiological research questions. One especially important facet specific to sequencing data is read depth (i.e. number of reads at a DNA methylation site), as the DNA methylation value for each DNA methylation site profiled is calculated as the average of the binary methylation status of each read. As such, the sensitivity of each DNA methylation value is dependent on the read depth. For example, a site captured by only two reads could have only three possible proportions, 0, 0.5 or 1, regardless of the actual DNA methylation level. This lack of precision at low read depths will inevitably influence the ability to detect differential DNA methylation patterns across a trait of interest, especially where effect size is small. It is common practice to remove DNA methylation values estimated from a low number of reads, however, the threshold for read depth is selected on an arbitrary basis and as such may not be optimal. Ideally an appropriate threshold should be informed by the properties of the data to maximise statistical power to detect differences.

Another common challenge encountered in processing BS data is missing DNA methylation values. There are two drivers of missingness, firstly, no reads being sequenced at a DNA methylation site for that sample, and secondly, due to the aforementioned removal of sites with insufficient read depth. Missingness of a DNA methylation value at a subset of samples effectively reduces a DNA methylation sites analytical sample size.

To assess optimum filtering and sample size, power calculations can be used with the aim of choosing thresholds that maximise study power. Power calculations have been utilised for array based studies (Mansell et al., 2019), where it is assumed that the sample size will be consistent and DNA methylation will be estimated with the same level of precision across all sites. Power calculations for BS studies are more complex with effective sample size varying per DNA methylation site and precision of DNA methylation estimates being dependent on read depth, and is yet to be fully explored. The relationship between power and read depth has been investigated previously by Ziller et al., 2015 for WGBS, however, the relationship with sample size was not explored (Ziller et al., 2015). This Chapter provides a more comprehensive assessment of the impact of read depth and establishes a framework for statistical analysis of BS studies.

This Chapter aims to:

1. characterise the statistical and genomic properties of RRBS data
2. investigate the relationship between power to detect differences between groups and
    - read depth
    - sample size
    - the magnitude of difference in DNA methylation between groups
3. utilise data simulation to estimate the power of a BS study with set filtering parameters

2.2 Summary of methods and results

Two datasets were utilised for the analysis within this paper:

- **Dataset mRRBS** was an RRBS dataset containing the DNA methylation profiles of 125 frontal cortex samples dissected from mice aged 2–10 months old. The quality of the sequencing data was assessed using FastQC (Andrews et al., 2010), before reads were aligned to the mm10 reference (GRCm38) genome using Bismark (Krueger and Andrews, 2011). The dataset was utilised to characterise the properties of RRBS data and to estimate the necessary parameters for the simulation of RRBS data.

- **Dataset mArray** was a secondary dataset utilised in the paper analyses, containing the DNA methylation profiles of 80 of the 125 samples from **Dataset mRRBS**, with DNA methylation quantified using the Illumina Beadchip vertebrate DNAm array (Arneson et al., 2021). The dataset was used as a 'true' measure of DNA methylation to which **Dataset RRBS** DNA methylation values could be compared across variable read depths.

The primary finding of this paper was that the relationship between study parameters and power were complex and dependent upon each other, with both read depth and sample size directly influence the statistical power for a specific DNA methylation site. As such, no one filtering threshold will be appropriate across all studies. The relationship between these parameters and power was non-linear, with sample size having a larger impact on power than read depth.

2.3 Contribution to the field

This publication is the first to characterise the nuances of performing statistical comparisons between groups where DNA methylation was profiled using RRBS. Raising the awareness of the properties of BS data, and the implications that has for the analysis and study design should encourage more transparency in the methodology of BS studies, specifically when it comes to data filtering. Additionally, more consideration as to an appropriate read depth filter, and the consequences of that choice, should lead to more reliable results that are reproducible. As a result of the analysis in this Chapter, software was developed that implemented the simulation framework for a user's bespoke study design enabling them to calculate the study-specific power of a two-group statistical comparison using BS data, applicable to both WGBS and RRBS datasets. The software allows for the calculation of the optimal read depth and minimum DNA methylation site sample size threshold. The code for POWEREDBiSeq is openly available at https://github.com/ds420/POWEREDBiSeq.

2.4 Personal contribution to the work

Dataset mRRBS (see **Table 6.3**) was aligned and preprocessed by DSV. All analyses were carried out using R by DSV, the code for which can be found at https://github.com/ds420/Characterizing-the-properties-of-bisulfite-sequencing-data. The interpretation and writing of the manuscript was carried out by DSV under the supervision of Dr Eilis Hannon and Prof. Jonathan Mill. All figures and schematics used to present the data were generated and conceived by DSV.

BMC Genomics

RESEARCH Open Access

Characterizing the properties of bisulfite sequencing data: maximizing power and sensitivity to identify between-group differences in DNA methylation

Dorothea Seiler Vellame[1*], Isabel Castanho[1,2,3], Aisha Dahir[1], Jonathan Mill[1*†] and Eilis Hannon[1*†]

Abstract

Background: The combination of sodium bisulfite treatment with highly-parallel sequencing is a common method for quantifying DNA methylation across the genome. The power to detect between-group differences in DNA methylation using bisulfite-sequencing approaches is influenced by both experimental (e.g. read depth, missing data and sample size) and biological (e.g. mean level of DNA methylation and difference between groups) parameters. There is, however, no consensus about the optimal thresholds for filtering bisulfite sequencing data with implications for the reproducibility of findings in epigenetic epidemiology.

Results: We used a large reduced representation bisulfite sequencing (RRBS) dataset to assess the distribution of read depth across DNA methylation sites and the extent of missing data. To investigate how various study variables influence power to identify DNA methylation differences between groups, we developed a framework for simulating bisulfite sequencing data. As expected, sequencing read depth, group size, and the magnitude of DNA methylation difference between groups all impacted upon statistical power. The influence on power was not dependent on one specific parameter, but reflected the combination of study-specific variables. As a resource to the community, we have developed a tool, POWEREDBiSeq, which utilizes our simulation framework to predict study-specific power for the identification of DNAm differences between groups, taking into account user-defined read depth filtering parameters and the minimum sample size per group.

Conclusions: Our data-driven approach highlights the importance of filtering bisulfite-sequencing data by minimum read depth and illustrates how the choice of threshold is influenced by the specific study design and the expected differences between groups being compared. The POWEREDBiSeq tool, which can be applied to different types of bisulfite sequencing data (e.g. RRBS, whole genome bisulfite sequencing (WGBS), targeted bisulfite sequencing and amplicon-based bisulfite sequencing), can help users identify the level of data filtering needed to optimize power and aims to improve the reproducibility of bisulfite sequencing studies.

Keywords: DNA methylation, Bisulfite sequencing, RRBS, Epigenetics, Power, Read depth, Sample size

* Correspondence: ds420@exeter.ac.uk; j.mill@exeter.ac.uk; e.j.hannon@exeter.ac.uk
†Jonathan Mill and Eilis Hannon contributed equally to this work.
[1]College of Medicine and Health, University of Exeter, Royal Devon and Exeter Hospital, Exeter EX2 5DW, UK
Full list of author information is available at the end of the article

Background

Epigenetic processes regulate gene expression via modifications to DNA, histone proteins and chromatin without altering the underlying DNA sequence, and there is increasing interest and understanding of the role that epigenetic variation plays in development and disease [1]. The most extensively studied epigenetic modification is DNA methylation (DNAm), the addition of a methyl group to the fifth carbon position of cytosine that occurs primarily, although not exclusively, in the context of cytosine-guanine (CpG) dinucleotides. Despite being traditionally regarded as a mechanism of transcriptional repression, DNAm is actually associated with both increased and decreased gene expression depending upon the genomic context [2], and also plays a role in other transcriptional functions including alternative splicing and promoter usage [3].

Inter-individual variation in DNAm has been associated with cancer [4], brain disorders [5–8], metabolic phenotypes [9, 10] and autoimmune diseases [11]. A number of high-throughput methods have been developed to quantify genome-wide patterns of DNAm, although these differ with regard to enrichment strategy, quantification accuracy and analytical approach [12]. Many approaches are based on the treatment of genomic DNA with sodium bisulfite, which converts unmethylated cytosines into uracil (and subsequently to thymine after amplification) while methylated cytosines are unaffected. The field of epigenetic epidemiology in human cohorts has been facilitated by the development of cost effective, standardized commercial arrays such as the Illumina EPIC Beadchip [13]. Data generated using this platform is relatively straightforward to process and analyze, with a number of standardized software tools and analytical pipelines [14, 15]. These arrays are only currently commonly available for human samples and are limited to capturing predefined genomic positions making up only ~ 3% of CpG sites in the human genome [16].

For studies requiring greater coverage of the genome, or for the quantification of DNAm in non-human organisms, it is typical to employ highly parallel short read sequencing of bisulfite-treated DNA libraries. A key step in the analytical pipeline of such data is the mapping or alignment of these short sequences back to the genome of interest, a process that is complicated by the degenerated sequence complexity of bisulfite-treated DNA [17]. As well as the need to determine accurately where in the genome a read originates from, the analysis of bisulfite sequencing data involves distinguishing reads mapping to methylated alleles from those mapping to unmethylated alleles. For each cytosine, the level of DNAm is estimated by quantifying the proportion of methylated (C) to unmethylated (T) cytosines from the sequenced reads overlapping that position. Bisulfite sequencing data provides information about cytosine methylation occurring in three distinct sequence contexts: CpG, CHH or CpH sites.

In this paper, we sought to characterize the properties of bisulfite sequencing data with the goal of exploring the experimental variables that influence statistical power and sensitivity to identify differences in DNA methylation in population-based analyses. We define ‘DNAm sites’ as vectors, such that each DNAm site has a ‘DNAm point’ per sample, which incorporates ‘read depth’ (i.e. the total number of reads covering that DNAm site), and ‘DNAm value’ (i.e. the proportion of methylated reads at that DNAm site). As with all sequencing applications, the total coverage, defined here as the total number of reads across the genome, is critical to the success of an experiment, as it will result in a higher average read depth at any individual DNAm point. Read depth influences both accuracy and statistical power. DNAm is measured as a proportion, therefore, when read depth is low there are only a finite number of possible values and the sensitivity of bisulfite sequencing is constrained. For example, a DNAm point covered by only four reads can only have five possible configurations of the ratio of methylated to unmethylated reads (4:0, 3:1, 2:2, 1:3, 0:4) resulting in the possible DNAm proportions of 0.00, 0.25, 0.50, 0.75, or 1.00. This lack of sensitivity has a direct effect on the magnitude and accuracy of differences that can be detected between groups, meaning that DNAm points with low average read depth may not have sufficient power for the detection of small or even moderate changes in DNAm. This is particularly pertinent as many studies of differential DNAm in complex phenotypes and disease typically identify changes of < 5% [8, 18]; such small differences are likely to require precise proportions of the DNAm to be detected.

An additional challenge for the interpretation of bisulfite sequencing data compared to array-based methods, which have a fixed content, is that the precise regions of the genome covered by sequencing reads generated in any given experiment can be highly variable. This means that DNAm sites captured in a sequencing experiment may not contain many DNAm points, and that even where the DNAm points have been assayed across many of the samples, the read depth is potentially highly variable. This results in a matrix of DNAm values with a high proportion of missing data, effectively lowering the sample size at that DNAm site, in turn reducing the power to detect associations in analysis.

The gold standard bisulfite-sequencing method is whole genome bisulfite sequencing (WGBS) [19], although this can be cost prohibitive for many studies and is not yet amenable for large epidemiological analyses. Furthermore, in a study where the main interest is cytosines, in particular at CpG sites, a high number of WGBS reads are uninformative. Reduced representation bisulfite sequencing (RRBS), in contrast, involves a target enrichment step using the methylation-insensitive enzyme MspI to target CpG-rich regions of the genome [20] prior to bisulfite conversion. This increases the proportion of informative sequencing reads, and RRBS typically interrogates DNAm sites in 85–90% of CpG islands [21, 22].

While multiple tools exist for the alignment and quantification of DNAm from bisulfite-sequencing data (e.g. *Bismark* [17], *GSNAP* [23], *BSMAP* [24], *BS-Seeker3* [25]), there is no consensus about the optimal approach for determining the appropriate minimum read depth or number of DNAm points required to ensure high-quality data for a well-powered statistical analysis. For example, existing studies have utilized a huge variety of read depth thresholds; a relatively arbitrary value between 5 and 20 reads per DNAm point is often used in filtering steps [26–29], most commonly with no justification provided for the use of that threshold. There is also no consensus as to what to do with DNAm sites that have very few DNAm points. Part of this inconsistency arises from a lack of guidelines or studies exploring how read depth and missingness influence statistical power.

The aim of this study was to determine the relationship between read depth and the accuracy of DNAm quantification, as well as the effect of missing DNAm points on statistical power for identifying group differences in DNAm with a particular focus on RRBS studies. Using properties derived from a large RRBS dataset generated by our group, we designed a simulation framework to explore how accuracy changes as a function of read depth, as well as comparing the DNAm level estimated from RRBS data with levels quantified using a novel Illumina array [30]. We then extended our simulation framework to investigate how statistical power to identify differences in DNAm level between groups varies as a function of read depth and sample size while also considering the effect of i) the level of DNAm at individual DNAm sites, ii) the expected difference in DNAm between groups, and iii) the balance of sample sizes between comparison groups. Our data-driven approach highlights the importance of filtering by minimum read depth and minimum number of DNAm points per DNAm site, and illustrates how the choice of threshold is influenced by the specific study design and the expected differences between groups being compared. Finally, we present an approach for estimating statistical power for a bisulfite sequencing study for a given read depth and minimum DNAm points filtering threshold which can be used to improve the detection of true positives and reproducibility of findings. Our tool, POWer dEtermined REad Depth filtering for Bisulfite Sequencing (POWEREDBiSeq), is available at https://github.com/ds420/POWEREDBiSeq as a resource to the community.

Results

Read depth in RRBS data follows a negative binomial distribution, while the level of DNAm is bimodally distributed

As part of an ongoing study of aging, we profiled DNAm in 125 frontal cortex samples dissected from mice aged 2–10 months old using the original RRBS protocol [20] (see Methods). Prior to quality control filtering, a mean of 41,199,876 (SD = 6,753,486) single end reads were generated per sample (Additional file 2). The quality of the sequencing data was assessed using *FastQC* [31], before reads were aligned to the *mm10* reference (GRCm38) genome using *Bismark* [17]. Here, we define DNAm sites as vectors, such that each DNAm site has a DNAm point per sample, containing read depth and DNAm values. That is, DNAm site = {DNAm point$_1$ = {m$_1$, rd$_1$}, …, DNAm point$_i$ = {m$_i$, rd$_i$}, …, DNAm point$_n$ = {m$_n$, rd$_n$}}, for i in 1 to n samples, where m$_i$ represents the proportion of DNAm at a DNAm point$_i$, and rd$_i$ is the read depth, defined here as the total number of reads at the DNAm point. If rd$_i$ is 0, there will be no DNAm point associated with sample i (Fig. 1). Across all samples, there was a total of 64,199,621 distinct DNAm points covered (including CpG, CpH and CHH sites), with a total of 3,419,677 different DNAm sites assayed, and each sample containing a mean of 2,170,454 (SD = 124,281) DNAm points across all DNAm sites. We characterized the distribution of read depth for each sample across DNAm points, observing a unimodal discrete distribution, skewed to the left and characterized by a long tail (Fig. 2a). This distribution is typical of count data and is expected in sequencing datasets where the vast majority of DNAm points are covered by relatively few reads and a

DNAm site $= \{\{m_1, rd_1\}, ..., \{m_i, rd_i\}, ..., \{m_n, rd_n\}\}$, where i is the sample number for a study containing n samples, m_i is the DNAm value for sample i at that DNAm site, and rd_i is the read depth.

For example:

DNAm site	Samples					# DNAm points
	S_1	S_2	S_3	S_4	S_5	
cg_1	$\{0,1\}$	$\{-,0\}$	$\{1,7\}$	$\{0.43,7\}$	$\{1,24\}$	4
cg_2	$\{1,32\}$	$\{1,21\}$	$\{1,2\}$	$\{1,4\}$	$\{1,4\}$	5
cg_3	$\{-,0\}$	$\{0.20,41\}$	$\{-,0\}$	$\{-,0\}$	$\{0.5,2\}$	2
cg_4	$\{0.88,49\}$	$\{0.82,17\}$	$\{0.93,15\}$	$\{-,0\}$	$\{-,0\}$	3
cg_5	$\{1,1\}$	$\{-,0\}$	$\{-,0\}$	$\{-,0\}$	$\{-,0\}$	1

The number of samples, n, in this example is 5, however, due to the nature of BS data, not every sample will have a DNAm point for each DNAm site, and so the number of DNAm points per DNAm site refers to the effective sample size at that given DNAm site. Here, DNAm sites have 4, 5, 2, 3, and 1 DNAm points, respectively, due to some samples having a read depth of zero at that DNAm site. The maximum number of DNAm points per DNAm site is equal to the sample size, n. The number of DNAm points at a DNAm site may decrease further if read depth filtering is applied, for example, if a read depth filter of 10 is applied to cg_1, the DNAm site will go from having four DNAm points to having only one.

Fig. 1 An overview and example of the term 'DNAm point' used in our analysis

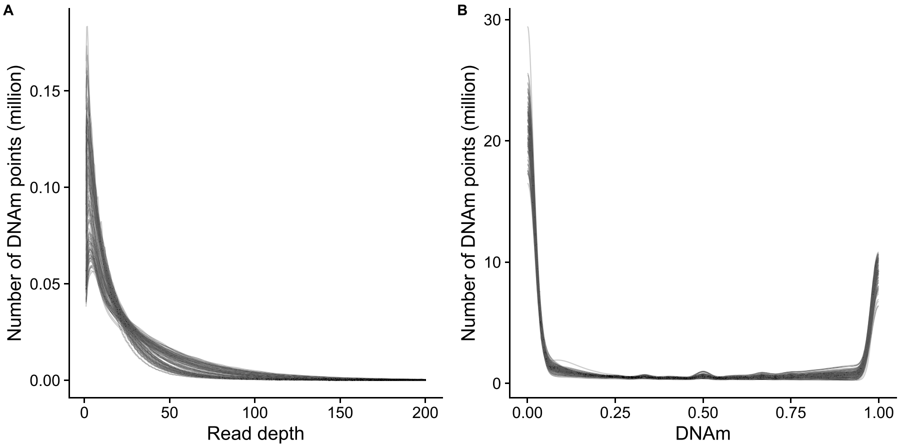


Fig. 2 Characterization of read depth and mean DNAm across the DNAm points profiled by RRBS. The distribution of **a** read depth across DNAm points and **b** proportion of DNAm across DNAm points. Each line represents one sample. Read depth plots were capped at a read depth of 200 to facilitate the interpretation of plots, with less than 0.5% (1140174) of DNAm points being characterized by > 200 reads

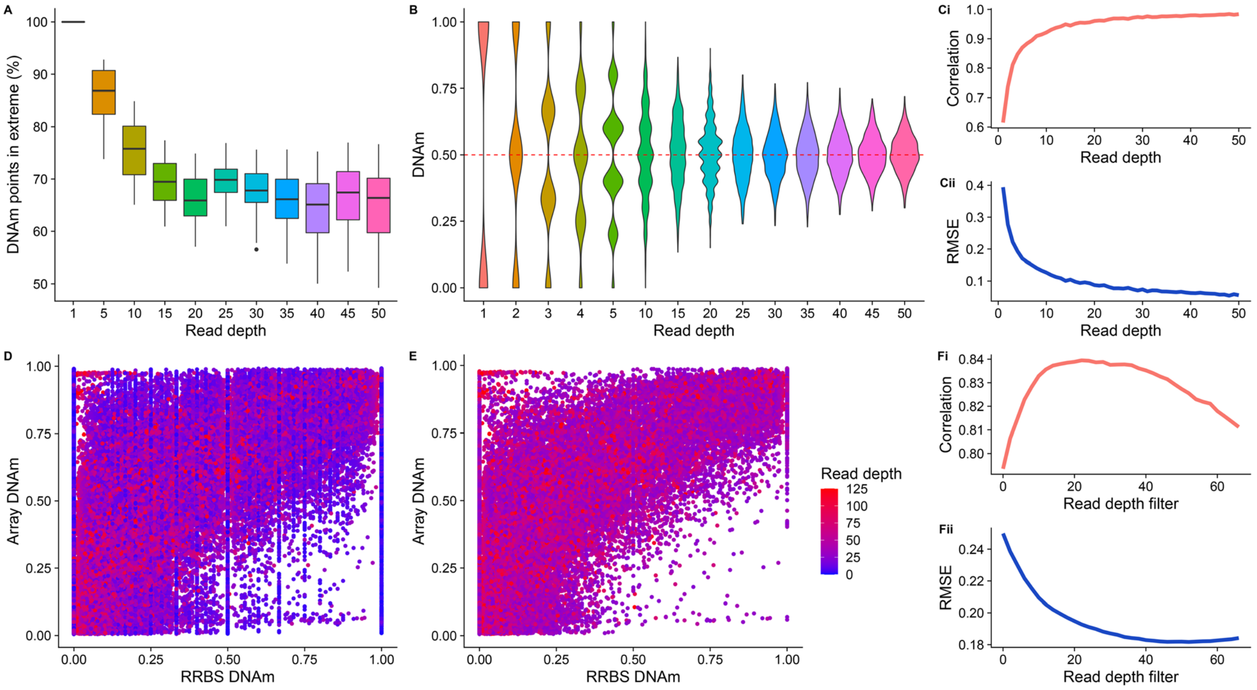


Fig. 3 The consequence of 'missingness' in RRBS data demonstrated by array and simulation bisulfite-sequencing data. **A**) A boxplot showing the proportion of DNAm points that have 'extreme' DNAm (0.05 < DNAm < 0.95) calculated for DNAm points with different read depths (x axis). **B**) Violin plots showing the distribution of estimated DNAm values from a simulated bisulfite sequencing experiment for a DNAm site where the true value is 0.50, as a function of read depth. Line graphs showing the Pearson correlation (**Ci**) and root mean squared error (RMSE) (**Cii**) between simulated and 'real' DNAm values for 1000 DNAm points as a function of read depth. These analyses used a subset of real data selected to contain DNAm points with read depth > 10 and evenly distributed DNAm (see Methods). Scatterplots of DNAm values quantified using RRBS (x-axis) and a custom vertebrate Illumina DNAm array [30] (y-axis) in matched samples (*n* = 80) for **D**) all DNAm points and **E**) the subset of DNAm points with read depth greater than the peak Pearson correlation read depth in **Fi** (i.e. 22 reads). Line graphs showing **Fi**) the Pearson correlation and **Fii**) error (RMSE) of RRBS data and array data as a function of the read depth filter applied to the RRBS dataset

minority of DNAm points are covered by a large number of reads. Across all DNAm points, 22.1% (60,117,549) had less than or equal to than 5 reads and 3.30% (8,941,868) had more than 100 reads. Next, we visualized the distribution of DNAm levels across all DNAm points, observing the expected bimodal distribution, with the majority of DNAm sites being either completely methylated (50% of DNAm sites > 0.95) or unmethylated (49% of DNAm sites < 0.05) [32] (Fig. 2b).

Read depth has a dramatic, non-linear effect on accuracy of DNAm estimates

One consequence of low read depth in RRBS data is reduced accuracy for the quantification of DNAm at DNAm points. While DNAm points that are either completely methylated or unmethylated can theoretically be characterized precisely with a single read, this is not the case for DNAm points with intermediate levels of DNAm, which may be inaccurately classed as methylated or unmethylated at low read depths. To understand the extent of this problem, we compared the proportion of DNAm values at extremes (less than 0.05 or greater than 0.95), with increasing read depths across DNAm points (Fig. 3A). As expected, the proportion of DNAm sites estimated to have extreme levels of DNAm was greater at lower read depths; 86.1% (SD = 4.94) of sites were estimated to have DNAm > 0.95 or < 0.05 at a read depth of 5, compared to 64.7% (SD = 6.90) at a read depth of 50. This suggests that, compared to DNAm points with a read depth of 50, more than 20% of DNAm points with a read depth of 5 may have been inaccurately classified as having an extreme level of DNAm.

To formally quantify the error in estimating DNAm, we used simulations of increasing read depth to estimate DNAm for a hypothetical DNAm site with an intermediate level of DNAm (0.50), calculating the difference between the estimated and true DNAm level. For read depths < 10, we observed a discrete distribution of estimated DNAm (Fig. 3B), with the range of predictions spanning 0.00–1.00 but centered on 0.50. In line with the Central Limit Theorem, we observe that as read depth increases, the distribution of estimated DNAm levels becomes more continuous and normally distributed around a DNAm value of 0.50. We expanded these simulations to consider DNAm sites with DNAm levels across the full distribution of possible values. We simulated 10,000 DNAm points with DNAm uniformly sampled between 0.00–1.00 and sampled 10,000 RRBS DNAm points with matched DNAm levels for comparison (see Methods). We found that as read depth increases, the correlation across DNAm points between estimated and actual DNAm level tends towards 1.00 (Fig. 3Ci) and the root mean squared error (RMSE) tends towards 0.00 (Fig. 3Cii). However, these effects are non-linear, with more dramatic improvements in accuracy occurring at lower read depths; i.e. there is a jump from a correlation of 0.589 to 0.926 between 1 and 10 reads with relatively minimal gains after that. Similarly, the RMSE drops from 0.404 at a read depth of 1.00 to 0.124 at a read depth of 10.

RRBS and Illumina arrays DNAm values correlate highly

Commercial DNAm arrays, such as the Illumina EPIC BeadChip array, are commonly utilized as an alternative strategy to bisulfite sequencing approaches in large human studies, due to their relatively low cost and the ease of interpreting data [33]. To further characterize the accuracy and sensitivity of RRBS, we performed a comparison with DNAm levels quantified using a novel Illumina Beadchip vertebrate DNAm array [30] on an overlapping set of 80 mouse frontal cortex DNA samples. A total of 3552 unique DNAm sites were quantified in both the RRBS and array datasets, with each RRBS sample containing a mean of 2263 overlapping DNAm data points (SD = 104). First, we compared the distribution of DNAm estimates across all DNAm points between the two technologies, observing the expected bimodal distribution with both approaches (Supplementary Figure 1). Of note, the array data contains a higher proportion of DNAm sites with intermediate levels of DNAm (0.05–0.95), and the unmethylated and methylated peaks are shifted inwards from the boundaries, highlighting the reduced sensitivity of the array for quantifying extreme levels of DNAm [16]. In contrast, the peaks in the RRBS data are at 0.00 and 1.00. The array samples also have less variability between samples, with distributions looking nearly identical, due to DNAm points being consistently characterized for each DNAm site. Directly comparing the estimated level of DNAm between the two assays, we observed a strong positive correlation (Pearson correlation = 0.794) even with no read depth filtering in the RRBS data (Fig. 3D). The correlation between assays increases as more stringent read depth filtering is applied to the RRBS data, with the maximum correlation (Pearson correlation = 0.840) obtained at a read depth threshold of 22 (Fig. 3E, Fi). Although this correlation indicates a relatively strong relationship between the estimates of DNAm quantified using RRBS and the Illumina array, it does not necessarily indicate that the DNAm estimates generated by the two platforms are equal. Closer inspection showed that the relationship between RRBS- and array-derived DNAm estimates is not linear (Fig. 3D), and therefore we also explored absolute differences in DNAm estimates between the two assays. We observed a notable skew, with DNAm estimates from the array being generally higher than those from RRBS (mean difference = 0.112, SD = 0.223), and this relationship was observed regardless of read depth (Supplementary Figure 2). As expected, the RMSE between DNAm estimates generated using array and RRBS decreases as the stringency of read depth filtering in the RRBS dataset increases (Fig. 3Fii), plateauing at a read depth of ~30. Of note, the minimum RMSE observed was 0.180, suggesting some systemic differences between the two platforms in estimated DNAm levels. Our findings corroborate previous findings in which DNAm estimates generated using Illumina arrays and BS data are strongly correlated [34–37].

RRBS enrichment results in a subset of DNAm sites that have consistent read depth across DNAm points

In order to perform a statistical analysis of DNAm differences between groups (e.g. in a study of cases vs controls), multiple samples, usually representing biological replicates, are required. We have demonstrated the importance of filtering RRBS data by read depth on obtaining accurate estimates of DNAm, however, this has the consequence of increasing the number of missing DNAm points (Fig. 4a). As expected, we found that read depth is not random across DNAm sites, but highly correlated between pairs of samples (Fig. 4b). To demonstrate this further, we iteratively increased the number of samples and calculated the proportion of DNAm points shared across DNAm sites (Fig. 4c). The proportion of DNAm points present decreases in a non-linear manner before plateauing at 0.20, demonstrating that there is a subset of DNAm sites for which read depth is greater than 0 across all or most DNAm points. DNAm sites containing all possible DNAm points, that is, each DNAm point had a read depth > 1, were found to have consistently higher read depth, with a strong correlation in read depths between DNAm points (Fig. 4d). This correlation in read depth between samples is a result of the enrichment strategy used in RRBS, meaning that specific CpG-rich regions are dramatically overrepresented in the sequencing data across all samples. As expected, the common DNAm sites containing all possible DNAm points were enriched in CpG islands compared to all DNAm sites (Fig. 4e) reflecting the *MspI*-based enrichment strategy used in RRBS [20].

Simulated data demonstrates the consequence of read depth, sample size, and mean DNAm difference per group on power

Statistical power to identify differences in DNAm between two groups (e.g. cases vs controls), defined as the proportion of successfully detected true positives, will vary across DNAm sites and is influenced by multiple variables. In bisulfite sequencing studies, these include read depth, the number of samples in each group, the ratio of group sizes, the mean DNAm level, and the expected difference in DNAm between groups. We explored how each of these variables influences power by simulating bisulfite sequencing data for a given DNAm site following the framework laid out in Fig. 5. Briefly, a two group comparison was simulated, with sample size,

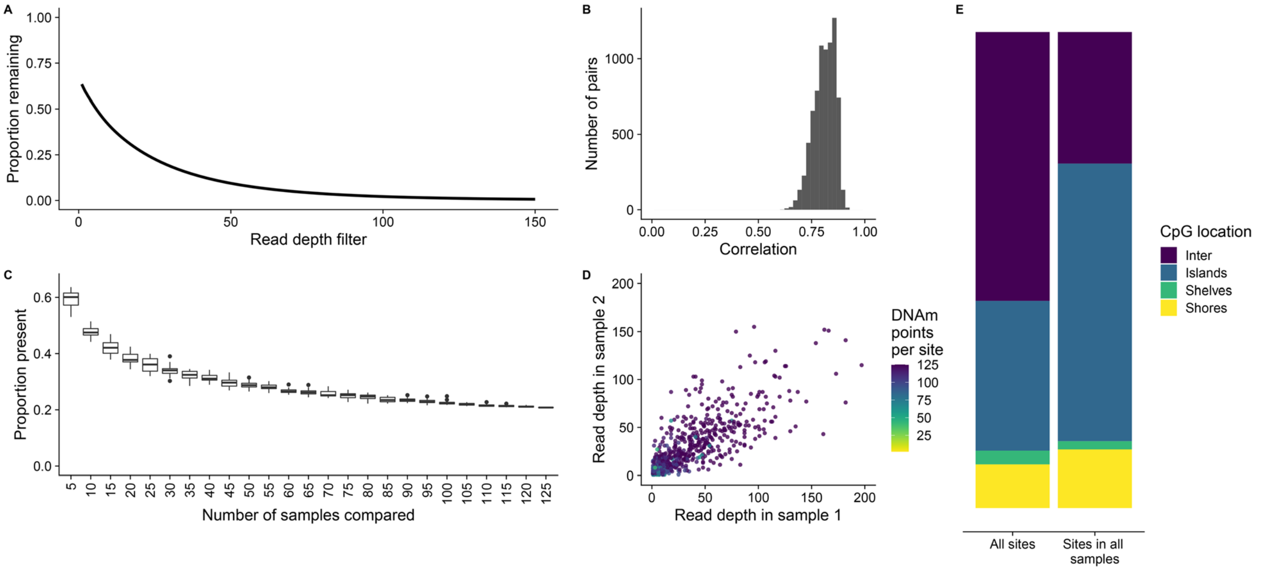
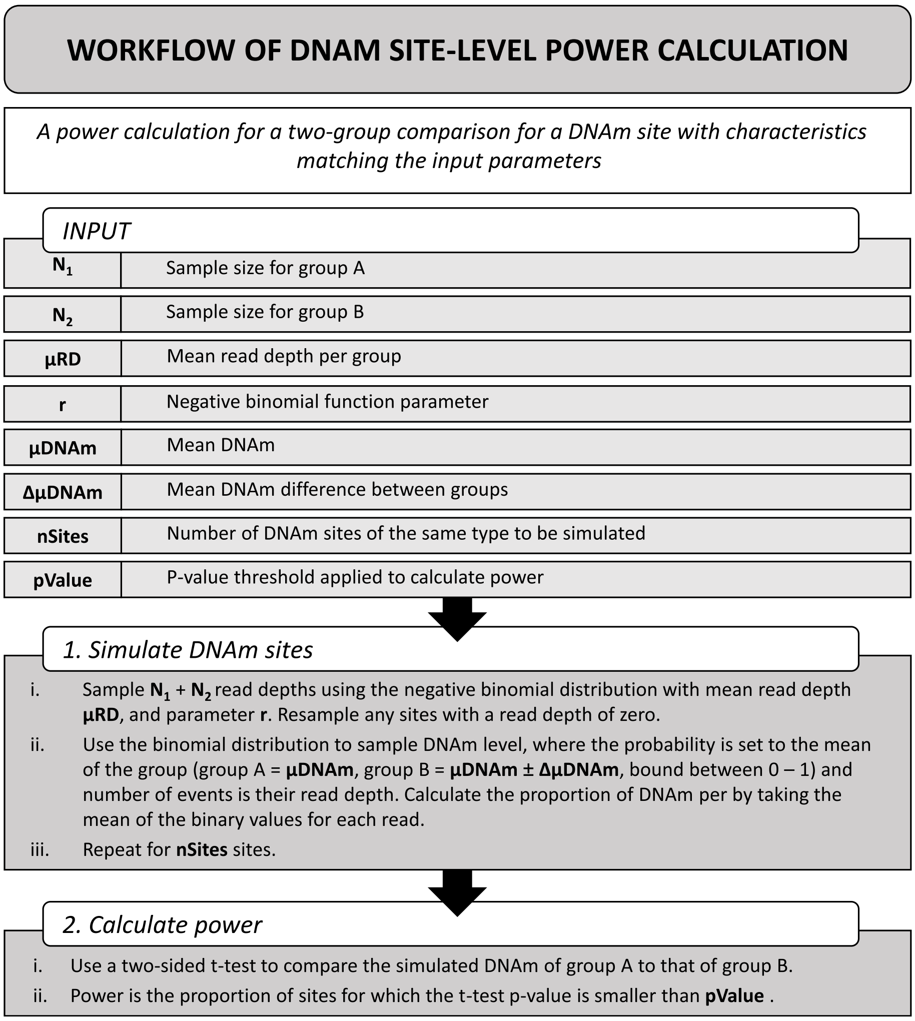


Fig. 4 A subset of higher read depth DNAm sites are over-represented in RRBS datasets. **a** A line graph of the mean proportion of DNAm points remaining (y-axis) after filtering by increasing read depth thresholds (x-axis). **b** The Spearman's correlation of read depth between all pairs of samples. **c** The proportion of overlap in the DNAm points present across an increasing number of samples compared. **d** Read depth plotted from two randomly selected samples, colored by the number of DNAm points that the DNAm site that have a read depth > 0. 1000 DNAm points were randomly selected and read depth is plotted up to 200 to facilitate the interpretation of plots. **e** The proportion of DNAm sites in intergenic regions (purple), CpG islands (blue), shelves (green) and shores (yellow) for all DNAm sites and all DNAm sites with read depth > 1 across all samples



WORKFLOW OF DNAM SITE-LEVEL POWER CALCULATION

A power calculation for a two-group comparison for a DNAm site with characteristics matching the input parameters

INPUT

N_1	Sample size for group A
N_2	Sample size for group B
µRD	Mean read depth per group
r	Negative binomial function parameter
µDNAm	Mean DNAm
ΔµDNAm	Mean DNAm difference between groups
nSites	Number of DNAm sites of the same type to be simulated
pValue	P-value threshold applied to calculate power

1. Simulate DNAm sites

i. Sample $\mathbf{N_1}$ + $\mathbf{N_2}$ read depths using the negative binomial distribution with mean read depth **µRD**, and parameter **r**. Resample any sites with a read depth of zero.

ii. Use the binomial distribution to sample DNAm level, where the probability is set to the mean of the group (group A = **µDNAm**, group B = **µDNAm ± ΔµDNAm**, bound between 0 – 1) and number of events is their read depth. Calculate the proportion of DNAm per by taking the mean of the binary values for each read.

iii. Repeat for **nSites** sites.

2. Calculate power

i. Use a two-sided t-test to compare the simulated DNAm of group A to that of group B.

ii. Power is the proportion of sites for which the t-test p-value is smaller than **pValue** .

Fig. 5 Outline of the framework for simulating bisulfite-sequencing data and assessing power in a DNAm site. This framework can be expanded to simulate a range of different DNAm sites by varying the input parameters

mean read depth, µDNAm (the mean DNAm across the DNAm point) and ΔµDNAm (the mean difference in DNAm between groups) used as input variables that were either kept constant or varied to observe the effect on power. Each exemplar DNAm site was simulated 10, 000 times, containing all DNAm points for the given sample size. A two-sided t-test was used to compare groups and power calculated as the proportion of *p*-values smaller than 5×10^{-6}. It is important to note that all parameters, including r, the *p*-value threshold for power, and number of DNAm sites simulated, were selected with the aim of visualising how power might change with each variable in turn. Subsequent findings are based on exemplar DNAm sites, and exact values should be taken as such; they may not be representative of a wider study, as our aim was solely to characterize the relationship between each variable and statistical power. The values used to generate the results for each variable shown in Fig. 6 can be found in Supplementary Table 1.

As expected, increased read depth had a positive effect on power across each of the scenarios we considered, however, the potential gains are highly dependent upon the specific combination of parameters (Fig. 6a). For example, in a scenario where each group contains 30 samples and the mean DNAm level is 0.25, there is a relatively dramatic increase in power to detect a DNAm difference of 0.20 between groups as read depth increases, with 80% power at a mean read depth of 37, although there are minimal gains with read depths > 50. In contrast the gain in power with increased read depth is much less pronounced when detecting a mean DNAm difference of 0.10, and there is very little power at any read depth to detect a DNAm difference of 0.05. Therefore, if small effect sizes are relevant for the phenotype under study, power will need to be increased through

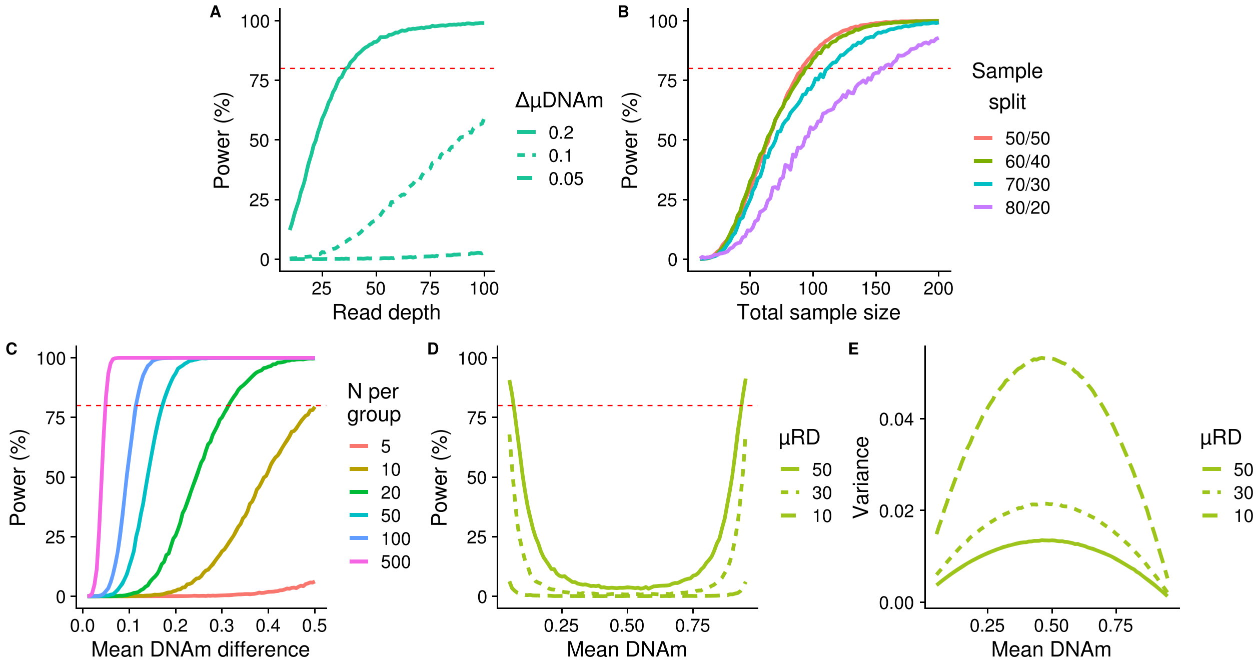


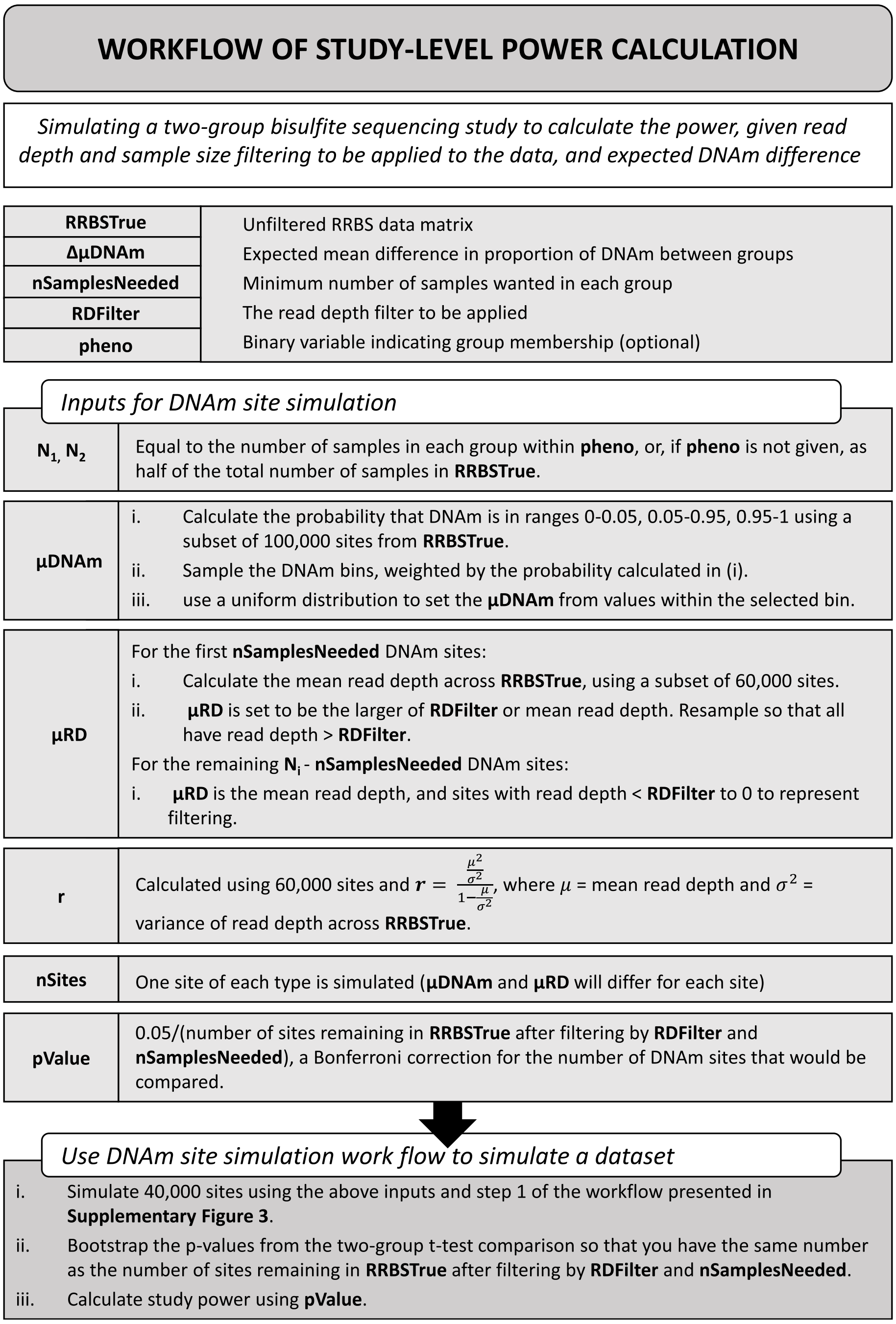
Fig. 6 Power is influenced by read depth, sample size, and mean DNAm level in two-group comparisons. Power curves plotting statistical power to detect significant differences in DNAm between two groups as a function of **a** read depth, **b** sample size and the effect of an unbalanced sample size between groups, **c** the mean difference in DNAm between the groups and **d** the mean DNAm at simulated DNAm sites. **e** The variance for the simulated data shown in panel **d**. Simulations were performed 10,000 times with a negative binomial parameter of r = 1.5

other methods, e.g. increased sample size, as read depth filtering alone will not be sufficient. The relationship between read depth and power was previously also found in WGBS data, with higher total sequencing depth increasing the true positives detected [38].

We next investigated the effect of sample size and the ratio of group sizes on power (Fig. 6b), concluding that the optimal design in terms of maximizing power is to have equal sized comparison groups, assuming that the total sample size is constant. Fixing mean read depth to be 20 and a mean DNAm level of 0.25, our simulations showed that to have 80% power to detect a DNAm difference of 0.20 between groups a total sample size of 94 is required when the sample size ratio between groups is 60:40 (56 and 38 samples, respectively), which is only two more samples than required when the sample size ratio is balanced (i.e. 50:50). In the most extreme scenario we considered, an 80:20 ratio between groups, a total of 154 samples (123 and 31, respectively) are needed to have 80% power to detect a DNAm difference of 0.20 between groups. This has implications for the handling of DNAm sites where DNAm points are missing; it suggests that there may be a tolerable level of 'missingness' when comparing DNAm between groups that can be 'rescued' by having a greater sample size in the second comparison group. As with read depth (Fig. 6a), we found a non-linear relationship with power for both sample size (Fig. 6b) and mean DNAm difference between groups (Fig. 6c). Where each of these variables is the limiting factor, we found that the greatest gains in power occurred initially, with diminishing returns at higher levels and an eventual plateau. Where other variables act to reduce the overall power, the power curve is flattened and a plateau is not reached. One interesting observation from our simulations was the U-shaped relationship between power and mean level of DNAm at a given site (Fig. 6d). Power is highest at DNAm sites with either very low or very high levels of DNAm, and decreases to a minimum at intermediate levels of DNAm. We hypothesize that this reflects the relationship between the mean and variance in DNAm [39] (Fig. 6e), where the variance is lowest at the extremes, an artefact of DNAm being measured as proportion bounded at 0.00 and 1.00.

Simulated bisulfite sequencing studies can be utilized to estimate power given suggested filtering

Our results indicate that, given the complex interplay of multiple experimental parameters, the choice of threshold for filtering DNAm sites is not always straightforward and will depend on the specific research question being addressed. Furthermore, the power calculations presented so far only consider a single DNAm site, whereas genome-wide comparisons of DNAm typically involve the analysis of hundreds of thousands of DNAm sites; given the effect of the properties of DNAm sites (e.g. in mean DNAm level) on power, no DNAm site can be considered to be 'representative' of the others. Therefore, we extended our simulation framework to quantify a study-level power statistic that considered all DNAm sites, allowing for the calculation of power given an RRBS dataset, and the read depth and minimum DNAm points per DNAm site filtering to be carried out. The extension of the simulation framework can be seen in Fig. 7 and is described in Methods. Briefly, an actual RRBS data set was used to estimate the simulation parameters (namely, sample size, µDNAm, µRD and negative binomial parameter, r) so that simulations reflect the real data. We compared the real and simulated data finding that the distribution of simulated read depths is highly comparable to real data for lower read depths (Fig. 8 Ai). Higher read depths do not seem to be



WORKFLOW OF STUDY-LEVEL POWER CALCULATION

Simulating a two-group bisulfite sequencing study to calculate the power, given read depth and sample size filtering to be applied to the data, and expected DNAm difference

RRBSTrue	Unfiltered RRBS data matrix
ΔµDNAm	Expected mean difference in proportion of DNAm between groups
nSamplesNeeded	Minimum number of samples wanted in each group
RDFilter	The read depth filter to be applied
pheno	Binary variable indicating group membership (optional)

Inputs for DNAm site simulation

N_1, N_2	Equal to the number of samples in each group within **pheno**, or, if **pheno** is not given, as half of the total number of samples in **RRBSTrue**.
µDNAm	i. Calculate the probability that DNAm is in ranges 0-0.05, 0.05-0.95, 0.95-1 using a subset of 100,000 sites from **RRBSTrue**. ii. Sample the DNAm bins, weighted by the probability calculated in (i). iii. use a uniform distribution to set the **µDNAm** from values within the selected bin.
µRD	For the first **nSamplesNeeded** DNAm sites: i. Calculate the mean read depth across **RRBSTrue**, using a subset of 60,000 sites. ii. **µRD** is set to be the larger of **RDFilter** or mean read depth. Resample so that all have read depth > **RDFilter**. For the remaining **N_i** - **nSamplesNeeded** DNAm sites: i. **µRD** is the mean read depth, and sites with read depth < **RDFilter** to 0 to represent filtering.
r	Calculated using 60,000 sites and $r = \frac{\frac{\mu^2}{\sigma^2}}{1-\frac{\mu}{\sigma^2}}$, where μ = mean read depth and σ^2 = variance of read depth across **RRBSTrue**.
nSites	One site of each type is simulated (**µDNAm** and **µRD** will differ for each site)
pValue	0.05/(number of sites remaining in **RRBSTrue** after filtering by **RDFilter** and **nSamplesNeeded**), a Bonferroni correction for the number of DNAm sites that would be compared.

Use DNAm site simulation work flow to simulate a dataset

i. Simulate 40,000 sites using the above inputs and step 1 of the workflow presented in **Supplementary Figure 3**.

ii. Bootstrap the p-values from the two-group t-test comparison so that you have the same number as the number of sites remaining in **RRBSTrue** after filtering by **RDFilter** and **nSamplesNeeded**.

iii. Calculate study power using **pValue**.

Fig. 7 Flow diagram describing the framework for simulating bisulfite sequencing studies utilized in POWEREDBiSeq. An application of the framework described in Fig. 5, used to assess the power of a two-group bisulfite sequencing study given different filtering parameters

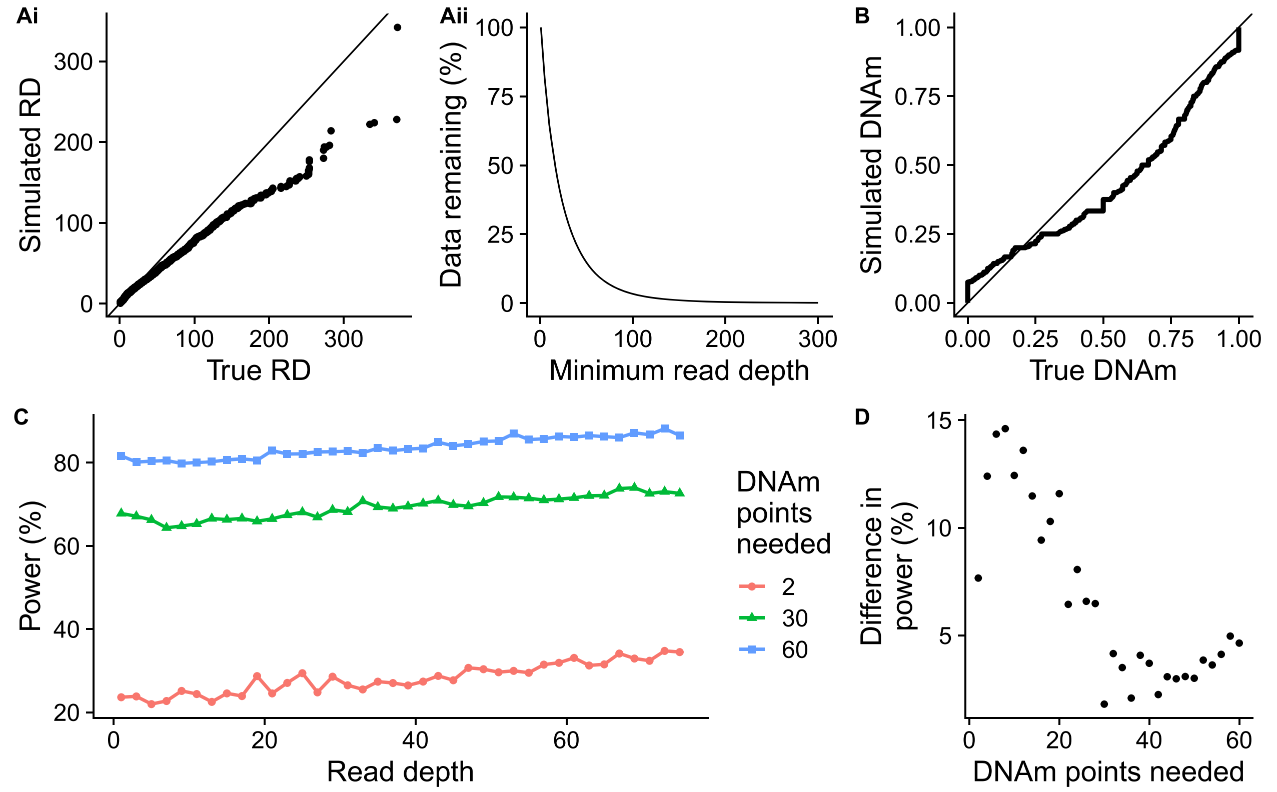


Fig. 8 Summarizing the simulation and predictions of POWEREDBiSeq. Ai) A QQplot comparing the read depth (RD) of 10,000 simulated DNAm points to 10,000 randomly sampled true DNAm points from an RRBS dataset. **Aii**) The proportion of DNAm points remaining in the RRBS dataset with read depths >x. **B**) A QQplot comparing the DNAm of 10,000 simulated DNAm points to a 10,000 randomly sampled true DNAm points. **C**) The relationship between the difference in power predicted by POWEREDBiSeq at different minimum sample sizes ($n = 2$, 30 and 60) as the minimum read depth threshold is increased, with a mean difference between groups of 0.06. **D**) The relationship between the increase in power to detect a mean difference in DNAm between groups of 0.06 predicted by POWEREDBiSeq at a read depth of 75 compared to power at a read depth of 1 as a function of the number minimum of samples per group

captured as accurately by the negative binomial distribution, however, given that 95% of DNAm points have read depth < 85 (Fig. 8 Aii), this should be less important to the simulation. Overall, simulated DNAm estimates were similar to real DNAm levels across DNAm points, although there was some deviation, for example, a slight under representation of DNAm points with DNAm proportions above 0.25 and an overrepresentation of DNAm points with DNAm proportions above 0.25 (Fig. 8B).

To demonstrate the methodology, we considered a hypothetical study design with a total of 125 samples, specifying an expected mean DNAm difference between groups of 0.06, picked arbitrarily to allow for power visualization. To profile how read depth influences the power of the study, we incrementally increased the minimum read depth from 1 to 75, and to investigate the effect of the minimum number of DNAm points needed to find a difference between groups (i.e. the minimum effective sample size at any DNAm site given filtering by read depths and the often sparse nature of RRBS), we chose three arbitrary values: 2, 30 and 60. Power only increased subtly as read depth filtering became more stringent (Fig. 8C), compared to the gain of increasing the number of DNAm points. However, the gain is not consistent across all study designs, with greater gains in smaller studies (Fig. 8D). For example, with a minimum of two DNAm points per group, increasing the read depth threshold from 1 to 75 resulted in an increase in power of 10.9%, compared to a smaller increase of 4.83 and 4.89%, respectively, when the minimum DNAm points were set at 30 or 60. Our analysis reaffirms the interplay between all study-specific experimental variables. However, it should be noted that even with the most extreme read depth filtering, the maximum power for a group with a minimum of two DNAm points is still dramatically lower that the power of a study with a larger minimum and no or negligible filtering. Finally, we summarized our study wide power calculation in the R function POWer dEtermined REad Depth filtering for Bisulfite Sequencing (POWEREDBiSeq), which is available as a resource to the community at https://github.com/ds420/POWEREDBiSeq. The calculation results in largely consistent and normally distributed predictions of power, however, outliers can occur, suggesting that multiple iterations should be performed (Supplementary Figure 3).

Discussion

In this paper, we systematically characterize the properties of a representative RRBS dataset, assessing the distribution of read depth and missing data across DNAm sites. Using our framework of bisulfite sequencing data simulation, we investigate the impact of various study variables (e.g. read depth, group size, skewness in group size, and magnitude of DNAm difference) on the accuracy of DNAm quantification, and power to detect DNAm differences between two groups. As a resource to the community, we have developed a tool (POWEREDBiSeq), which utilizes our findings to predict power for individual study designs, accounting for the filtering to be applied.

When comparing to simulated data, we found that the accuracy to detect a given DNAm difference between groups improves with increased read depth. This likely reflects the fact that count data is only able to represent continuous data if the number of counts (i.e. sequencing reads) is high enough. Overall, we found a strong correlation in DNAm estimates derived from RRBS and Illumina DNAm array data; this relationship increases with minimum read depth filtering and reaches a maximum when excluding DNAm sites covered by less than 22 reads. The high correlation between platforms and the relationship with read-depth concurs with previous analyses comparing RRBS and the Illumina 450 K array [34], RRBS and the Illumina 27 K array [35], targeted bisulfite sequencing and the Illumina EPIC array [36], and WGBS and the Illumina EPIC array [37]. This finding has implications for studies using RRBS to identify differences in DNAm as it highlights the importance of read depth filtering in generating an accurate measure of the true DNAm level.

We investigated the impact of various experimental variables on power, defined as the proportion of true positives detected in a two-group comparison, in a bisulfite-sequencing study utilizing simulated data. We observed that these variables (read depth, sample size, DNAm difference between groups and mean DNAm at a given DNAm site) act together to influence power. Read depth, sample size and DNAm difference between groups will all limit power in a certain range, with power plateauing at 100% when they are no longer the limiting factor. DNAm level at a DNAm site has a U-shaped relationship with power, where DNAm points with extreme DNAm (near 0 and 1) are more powered to identify between-group differences primarily because the variance in DNAm at these DNAm sites is smaller. Our findings highlight the importance of data filtering for maximizing power; the minimum number of DNAm points needed across each DNAm site to be compared has a dramatic effect on power, as it dictates the minimum effective sample size at any one DNAm site. Read depth also influences power, although we observed that read depth filtering alone cannot overcome an inadequate study design (i.e. too few samples). As a resource to the community, we have summarized our data simulations so that others can apply them to their data to calculate the power to identify between-group differences in DNAm within the context of their specific study design. Our scripts are packaged into the POWEREDBiSeq application (https://github.com/ds420/POWEREDBiSeq) which allows users to optimize their power by, for example, simulating the effects of increasing their sequencing read depth filtering threshold or minimum DNAm points across groups.

Although our analyses and simulations focused on RRBS datasets, many of our conclusions are valid for other types of bisulfite sequencing data. For example, the relationship between read depth and accuracy applies to any bisulfite sequencing based DNAm experiment that profiles DNAm at a single nucleotide resolution. Additionally, the relationship between power and read depth, sample size, DNAm difference, and mean DNAm is also relevant for other sequencing based DNAm studies. Various methods differ in read depth and the distribution of DNAm sites sequenced across the genome. Targeted bisulfite sequencing (TBS) and amplicon-based sequencing, for example, typically profile a more restricted set of DNAm sites than RRBS, as only regions of interest are enriched. This results in a more uniform distribution of reads across DNAm points, which acts to improve power across the study. In whole genome bisulfite sequencing (WGBS) studies, however, while more DNAm sites are interrogated across the genome as a whole, the read depth per DNAm point tends to be lower than that obtained using RRBS or TBS. POWEREDBiSeq can be applied to other bisulfite sequencing types because the internal variables, such as DNAm distribution and number of DNAm sites, are calculated based on input data. For the same reason, POWEREDBiSeq is also applicable to DNAm at CHH and CGH sites, which are often covered in bisulfite sequencing studies but have dramatically different properties to DNAm at CpG sites, although it is important to verify that the simulated and real data distributions are alike. In datasets with a frequent occurrence of high read depths across DNAm points (> 100), some caution in the use of POWEREDBiSeq is warranted, as we found that the negative binomial distribution underestimates higher read depths when simulating data. This was not pertinent in our case as the 95% of sites had a read depth below 85. For the simulation of read depth, Poisson [40] or negative binomial distributions [41] have been used; we chose the negative binomial approach as it allows the variance to differ from the mean. The binomial distribution has been utilized to model DNAm in previous studies [40, 42].

The results of POWEREDBiSeq will be dependent on the planned filtering stringency of the user, as well as the biological question that the bisulfite sequencing experiment aims to address; for example, a study looking into DNAm changes between cancer and non-cancer samples will have higher power due to the comparatively large DNAm differences between groups [43] compared to those observed in many complex disease case and control studies [8, 18]. Bisulfite sequencing data generated in cell lines and genetically identical mouse models will be comparatively less 'noisy' than analyses of diverse human populations using heterogeneous tissues such as blood, resulting in increased power. Retaining poor quality (i.e. low read depth) DNAm sites in a bisulfite sequencing dataset increases the multiple testing burden, meaning it will be harder to identify true between-group differences in DNAm at higher quality, more adequately powered, DNAm sites. A limitation of POWEREDBISEQ and our data simulations is that they are based on a two-group comparison (e.g. cases vs controls), meaning our findings are not specifically applicable to more complex study designs. One question not addressed by our analysis is whether, for a given amount of available resource, it is optimal to sequence more samples at the same level or increase sequencing depth for a smaller number of samples. This was explored previously by Ziller and colleagues [38] using WGBS data; they concluded that with a low total sequencing depth of 10x, the best sensitivity was achieved by including an additional replicate per group with 5x coverage. If total sequencing depth potential was higher, the most optimal sensitivity was gained by increasing the number of replicates, rather than increasing sequencing depth above 10x. An equivalent study has not been carried out in RRBS data due to a lack of additional RRBS studies with sufficient coverage.

To our knowledge, this is the first attempt to develop recommendations for bisulfite sequencing experiments based on sequencing read depth, minimum number of DNAm points and statistical power. We believe findings from this work will improve the reproducibility of bisulfite sequencing studies; we encourage researchers working in this field to clearly detail any data filtering steps and ensure an appropriate filter for read depth and other parameters has been applied, with justification for the choice of threshold.

Methods

DNAm quantification by RRBS

Genomic DNA was isolated from mouse cortex [44] using the AllPrep DNA/RNA Mini Kit (QIAGEN) and assessed for quality and quantity using the NanoDrop 8000 spectrophotometer (Thermo Fisher Scientific) and the Qubit high sensitivity assay (Qubit dsDNA HS Assay, Thermo Fisher Scientific). RRBS libraries were prepared using the Premium RRBS kit (Diagenode). The final RRBS library pools were distributed across thirty-two HiSeq2500 (Illumina) lanes and subjected to 50 bp single-end sequencing as previously described [20].

Preprocessing the dataset

RRBS sequencing quality was assessed using *FastQC* (version v0.11.7) [31] with all samples characterized by high quality base calls (quality score > 28 across all bases). Sequences were trimmed using *TrimGalore* (version 0.4.4_dev) [45], with a quality score of 20 and an error rate of 0.2 used to remove poor quality bases at the ends of reads. Reads with fewer than 20 base pairs after trimming were then removed. Reads were aligned to the mm10 (GRCm38) mouse genome [46] using *Bismark* v0.19.0 with default parameters [17], which implements *SAMtools* 1.8 [47] and *Bowtie2* v2.3.4.1 [48]. The total number of aligned reads and cytosines can be found in Additional file 2.

Statistical methods

All subsequent analysis was carried out in R 3.5.2 (2018-12-20) [49] using the R packages *ggplot2* (version 3.2.1) [50], *Cowplot* (version 1.0.0) [51], *Tidyr* (version 1.0.0) [52], *Viridis* (version 0.5.1), *viridisLite* (version 0.3.0) [53], *colortools* (version 0.1.5) [54], and *reshape2* (version 1.4.3) [55].

Annotating RRBS to the CpG islands

R packages *annotatr* (version 1.8.0) [56] and *GenomicRanges* (version 1.34.0) [57] were used to annotate CpGs to features for the analyses shown in Fig. 3E. The *annotatr* package assigned CpG islands as per the mm10 reference annotation, with CpG shores defined as 2Kb upstream/downstream from the ends of the CpG islands, and CpG shelves as another 2Kb upstream/downstream of the farthest upstream/downstream limits of the CpG shores. The remaining genomic regions make up the inter-CGI annotation.

DNAm quantification quantified by array

A subset of 80 DNA samples were additionally profiled using a custom Illumina DNAm array (the "HorvathMammalMethylChip40" [30]). Briefly, this array includes ~ 36 k CpGs that are located in genomic regions highly-conserved across 50 mammalian species. Data was loaded from idat files into an RGChannelSet object using the *minfi* package (version 1.28.4) [58–64] and processed through the following steps: 1) checking the methylated and unmethylated intensities and excluding samples < 800, 2) confirming successful bisulfite conversion excluding samples with low conversion rates (< 80%), 3) confirming correct sex using profiles from the X chromosome, and 4) confirming tissue type, excluding any sample predicted incorrectly based on DNAm profile. Prior to analysis data was normalised using the *Sesame* package (version 1.4.0) [65], and filtered to DNAm sites classed as mapping uniquely to the mouse genome, leaving 23,633 DNAm sites.

Framework for simulating RRBS data

We developed an analytical framework to profile the power of RRBS DNAm sites, enabling us to vary different parameters such that we could explore a number of research questions. The DNAm site-level simulation workflow is described in Fig. 5, which aims to compare the DNAm between two groups, A and B. For each DNAm site simulated, there are 8 parameters to consider: N_1 and N_2 are the sample size each group, respectively, μRD is the mean read depth of the DNAm site to be simulated, r is a negative binomial parameter, described in more detail below, μDNAm is the mean DNAm across the DNAm site, *ΔμDNAm* is the mean difference in DNAm between groups, *nSites* is the number of DNAm sites to be simulated, and *pValue* is the p-value used to assess power.

When simulating a DNAm site, the first step is to simulate read depth. Read depth could be assigned an arbitrary value, or, where realistic variation across DNAm points was required, could be sampled from a negative binomial distribution [41]. The negative binomial distribution is defined by the parameters r and p, although within the R function *rnbinom()* can be defined by *μRD* and r, which can calculated from real data using eq. (6), the derivation of which is as follows:

The negative binomial equations are:

$$\mu = \frac{pr}{1-p} \tag{1}$$

$$\sigma^2 = \frac{pr}{(1-p)^2} \tag{2}$$

Where μ is the mean (in this case, *μRD*) and σ^2 is the variance of the read depth data calculated across all samples. We want r in terms of μ and σ^2. Multiply (2) by 1 − p and equate that and (1) to get:

$$\sigma^2(1-p) = \mu \tag{3}$$

Rearrange for p:

$$p = 1-\frac{\mu}{\sigma^2} \tag{4}$$

Substitute (4) into (1) and simplify:

$$\begin{aligned} \mu &= \frac{\left(1-\frac{\mu}{\sigma^2}\right)r}{1-\left(1-\frac{\mu}{\sigma^2}\right)} \\ \mu &= \frac{\left(1-\frac{\mu}{\sigma^2}\right)r}{\frac{\mu}{\sigma^2}} \end{aligned} \tag{5}$$

Rearrange (5) for r:

$$r = \frac{\frac{\mu^2}{\sigma^2}}{1-\frac{\mu}{\sigma^2}} \tag{6}$$

Once read depth was established, a binary value representing DNAm status was assigned to each read using the binomial distribution. For each read in group A, the probability of being methylated was *μDNAm*, and for group B was *μDNAm* ± *ΔμDNAm*, where the probability was bound between 0 and 1. The proportion of DNAm was calculated as the mean DNAm at each DNAm point.

The process was repeated for *nSites*. To calculate power, a two-sided t-test was performed between groups A and B. Power was defined as the proportion of DNAm sites for which the t-test p-value was smaller than *pValue*.

Profiling the accuracy of RRBS data

To investigate how the distribution and accuracy of DNAm changed with increasing read depth, we considered a range of read depths (1–50). To profile accuracy across levels of DNAm in an RRBS study, we simulated 10,000 DNAm points per read depth, with DNAm sampled uniformly between 0 and 1. Ten thousand DNAm points with matching DNAm were sampled from the RRBS data and correlation and RMSE were calculated between the true and the estimated DNAm points for each read depth.

Profiling the power of RRBS data

To calculate the power of RRBS DNAm sites, we investigated a hypothetical two-group comparison study design (e.g. a case vs control analysis). We aimed to explore the effects of read depth, mean DNAm level, the sample size and sample size balance of groups, and the mean DNAm difference between groups on power. To this end, we utilized the simulation framework described above and in Fig. 5 to simulate specific DNAm sites so that the resulting shift in power, given a change in a variable or combination of variables, could be visualized. The parameters assigned can be seen in Supplementary Table 1, where the variable parameter took a range of discrete values as seen in the x axes in Fig. 6. μRD set was used as a negative binomial parameter, from which read depth (> 0) was sampled. For group A, μDNAm was used as the probability of DNAm, sampled from the binomial distribution. For each set of parameters chosen, 10,000 DNAm sites were simulated. The *r* value was 1.5, and *pValue* 5×10^{-6}, which were chosen arbitrarily to allow for the visualization of changing power.

Profiling the power of RRBS studies given data filtering

We aimed to create a power calculator to determine the statistical power of a bisulfite sequencing study with specified read depth and minimum DNAm point filtering thresholds and specified mean DNAm difference between groups across a two-group study design. To this end, we utilized the simulation framework described above and in Fig. 5 to simulate filtered data. The following input data was required (also described in Fig. 7): *RRBSTrue* - the unfiltered matrix of RRBS data, *Δ*μDNAm - the mean difference in DNAm between groups expected given the biology of the samples, *nDNAmPoint* - the minimum number of DNAm points needed per DNAm site, *RDFilter* – the minimum read depth filter to be applied, *pheno* – an optional variable dictating group membership.

These were used to estimate the variables for the framework in Fig. 5: N_1 and N_2 were assigned using *pheno*, or if *pheno* was not given, assigned as half of the number of samples in *RRBSTrue*. The data being simulated represented data that remained was post-filtering, therefore, given that we need at least *nDNAmPoint* DNAm points with sufficient read depth, *μRD* was calculated separately for the first *nDNAmPoint* DNAm points to the latter. For the first *nDNAmPoint* DNAm points, *μRD* was the larger of the mean read depth across *RRBSTrue* (estimated using 60,000 DNAm sites) and *RDFilter*, and subsequent read depth must be > *RDFilter*. For the remaining DNAm points, the mean read depth was used, where all simulated read depths < *RDFilter* were assigned a read depth of 0 to represent that they would get filtered out of the data. *r* was estimated using eq. 6 and a subset of 60,000 DNAm sites from *RRBSTrue*. To estimate *μDNAm*, we first estimated the probability that a filtered DNAm site falls into one of the following ranges: 0–0.05, 0.05–0.95, 0.95–1, using a subset of 100,000 DNAm sites from *RRBSTrue*. The ranges were sampled using the probabilities calculated and a uniform distribution used to set *μDNAm* from the values across the selected range. To ensure that the subsets of *RRBSTrue* used to estimate variables were enough, we investigated the decline in prediction variability for each (Supplementary Figures 4, 5 and 6).

Forty thousand DNAm sites were simulated, using the above inputs and step 1 of the workflow presented in Fig. 5 and above. The resulting *p*-values were bootstrapped to result in the same number as the number of DNAm sites remaining in *RRBSTrue* after filtering by *RDFilter* and *nDNAmPoint*. The power was calculated using a Bonferroni correction for the number of DNAm sites remaining.

We created POWEREDBiSeq so that others can calculate their statistical power in bisulfite sequencing studies. The R function is available at https://github.com/ds420/POWEREDBiSeq.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07721-z.

Additional file 1: Supplementary Figure 1. The distribution of DNAm levels across the genome profiled using RRBS or a custom array. **Supplementary Figure 2.** DNAm estimates derived from RRBS are on average lower than those from the array platform. **Supplementary Figure 3.** A histogram of POWEREDBiSeq calculations showing variability in estimated power. **Supplementary Figure 4.** r is more accurately estimated when using a larger number of DNAm sites. **Supplementary Figure 5.** DNAm priors are more accurately estimated when using more DNAm sites. **Supplementary Figure 6.** The proportion of DNAm sites remaining is more accurately estimated when using more DNAm sites. **Supplementary Table 1.** A summary of parameters used in simulation analysis.

Additional file 2. RRBS information on total number of reads aligned, unaligned ambiguously aligned, and total number of reads, as well as the number of methylated and unmethylated CpGs, CpH, and CHH's, and total number of cytosines.

Acknowledgements

NA

Authors' contributions

DSV, JM and EH wrote the manuscript. IC performed the lab work. Analysis was led by DSV, who was assisted by AD. The author(s) read and approved the final manuscript.

Funding

DSV is funded by a BBSRC CASE PhD studentship. JM and EH are supported by a UK Medical Research Council Grant (MR/R005176/1 awarded to JM). Sequencing data was funded in part through the Medical Research Council (MRC) Proximity to Discovery: Industry Engagement Fund (Precision Medicine Exeter Innovation Platform reference MC_PC_14127) and a research grant from Alzheimer's Research UK (ARUK-PG2018B-016). Sequencing infrastructure was supported by a Wellcome Trust Multi User Equipment Award (WT101650MA) and Medical Research Council (MRC) Clinical Infrastructure Funding (MR/M008924/1).

Availability of data and materials

Data has been deposited in GEO under accession number GSE169235. All scripts for this paper are publicly available and can be found in the Characterizing-the-properties-of-bisulfite-sequencing-data repository at https://github.com/ds420/Characterizing-the-properties-of-bisulfite-sequencing-data.

Declarations

Ethics approval and consent to participate

This study analyzed data generated as part of another study (not as part of this study). The original work was approved by the UK Home Office. All animal procedures were carried out at Eli Lilly and Company, in accordance with the UK Animals (Scientific Procedures) Act 1986 and with approval of the local Animal Welfare and Ethical Review Board.

Consent for publication

NA

Competing interests

None of the authors have any competing interests relevant to this study.

Author details

[1]College of Medicine and Health, University of Exeter, Royal Devon and Exeter Hospital, Exeter EX2 5DW, UK. [2]Department of Pathology, Beth Israel Deaconess Medical Center, 330 Brookline-Avenue, Boston, Massachusetts, USA. [3]Harvard Medical School, Boston, Massachusetts, USA.

Received: 20 January 2021 Accepted: 13 May 2021
Published online: 15 June 2021

References

1. Murphy TM, Mill J. Epigenetics in health and disease: heralding the EWAS era. Lancet. 2014;383(9933):1952–4. https://doi.org/10.1016/S0140-6736(14)60269-5.
2. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. 2014;15(2):R37. https://doi.org/10.1186/gb-2014-15-2-r37.
3. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, Dsouza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466(7303):253–7. https://doi.org/10.1038/nature09165.
4. Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer. 2004;4(2):143–53. https://doi.org/10.1038/nrc1279.
5. Hannon E, Dempster E, Viana J, Burrage J, Smith AR, Macdonald R, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. Genome Biol. 2016;17(1):176. https://doi.org/10.1186/s13059-016-1041-x.
6. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci. 2014;17(9):1156–63. https://doi.org/10.1038/nn.3786.
7. Lunnon K, Smith R, Hannon E, De Jager PL, Srivastava G, Volta M, et al. Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. Nat Neurosci. 2014;17(9):1164–70. https://doi.org/10.1038/nn.3782.
8. Iurato S, Carrillo-Roa T, Arloth J, Czamara D, Diener-Hölzl L, Lange J, et al. DNA methylation signatures in panic disorder. Transl Psychiatry. 2017;7(12):1287. https://doi.org/10.1038/s41398-017-0026-1.
9. Petersen AK, Zeilinger S, Kastenmüller G, Werner RM, Brugger M, Peters A, et al. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. Hum Mol Genet. 2014;23(2):534–45. https://doi.org/10.1093/hmg/ddt430.
10. Huang Y, Hui Q, Walker DI, Uppal K, Goldberg J, Jones DP, et al. Untargeted metabolomics reveals multiple metabolites influencing smoking-related DNA methylation. Epigenomics. 2018;10(4):379–93. https://doi.org/10.2217/epi-2017-0101.
11. Carnero-Montoro E, Alarcón-Riquelme ME. Epigenome-wide association studies for systemic autoimmune diseases: the road behind and the road ahead. Clin Immunol. 2018;196:21–33. https://doi.org/10.1016/j.clim.2018.03.014.
12. Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. Epigenetics Chromatin. 2016;9(1):26. https://doi.org/10.1186/s13072-016-0075-3.
13. Illumina. Illumina Support. http://support.illumina.com. Accessed 2 May 2021.
14. Takeuchi F, Kato N. Nonlinear ridge regression improves robustness of cell-type-specific differential expression studies. BMC Bioinforma. 2021;22(1):1–25. https://doi.org/10.1186/s12859-021-03982-3.
15. Pidsley R, CCY W, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013;14(1):293. https://doi.org/10.1186/1471-2164-14-293.
16. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17(1):208. https://doi.org/10.1186/s13059-016-1066-1.
17. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571–2. https://doi.org/10.1093/bioinformatics/btr167.
18. Smith RG, Pishva E, Shireby G, Smith AR, Roubroeks JAY, Hannon E, et al. Meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights 220 differentially methylated loci across cortex; 2020. https://doi.org/10.1101/2020.02.28.957894.
19. Plongthongkum N, Diep DH, Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nat Rev Genet. 2014;15(10):647–61. https://doi.org/10.1038/nrg3772.
20. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc. 2011;6(4):468–81. https://doi.org/10.1038/nprot.2010.190.
21. Smith ZD, Gu H, Bock C, Gnirke A, Meissner A. High-throughput bisulfite sequencing in mammalian genomes. Methods. 2009;48(3):226–32. https://doi.org/10.1016/j.ymeth.2009.05.003.
22. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008;454(7205):766–70. https://doi.org/10.1038/nature07107.
23. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26(7):873–81. https://doi.org/10.1093/bioinformatics/btq057.
24. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10(1):232. https://doi.org/10.1186/1471-2105-10-232.
25. Huang KYY, Huang YJ, Chen PY. BS-Seeker3: ultrafast pipeline for bisulfite sequencing. BMC Bioinformatics. 2018;19(1):111. https://doi.org/10.1186/s12859-018-2120-7.
26. Gu H, Bock C, Mikkelsen TS, Jäger N, Smith ZD, Tomazou E, et al. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. Nat Methods. 2010;7(2):133–6. https://doi.org/10.1038/nmeth.1414.
27. Kessler NJ, Waterland RA, Prentice AM, Silver MJ. Establishment of environmentally sensitive DNA methylation states in the very early human embryo. 2018. http://advances.sciencemag.org/. Accessed 22 Nov 2019.
28. Lutz PE, Tanti A, Gasecka A, Barnett-Burns S, Kim JJ, Zhou Y, et al. Association of a history of child abuse with impaired myelination in the anterior cingulate cortex: convergent epigenetic, transcriptional, and morphological evidence. Am J Psychiatry. 2017;174(12):1185–94. https://doi.org/10.1176/appi.ajp.2017.16111286.
29. Stubbs TM, Bonder MJ, Stark A-K, Krueger F, von Meyenn F, Stegle O, et al. Multi-tissue DNA methylation age predictor in mouse. Genome Biol. 2017;18(1):68. https://doi.org/10.1186/s13059-017-1203-5.
30. Arneson A, Haghani A, Thompson MJ, Pellegrini M, Bin Kwon S, Vu H, et al. A mammalian methylation array for profiling methylation levels at conserved sequences. bioRxiv. 2021;2021.01.07.425637. https://doi.org/10.1101/2021.01.07.425637.
31. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. FastQC. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
32. Hannon E, Chand AN, Evans MD, Wong CCY, Grubb MS, Mill J. A role for CaV1 and calcineurin signaling in depolarization-induced changes in neuronal DNA methylation. Neuroepigenetics. 2015;3:1–6. https://doi.org/10.1016/j.nepig.2015.06.001.
33. Fan S, Chi W. Methods for genome-wide DNA methylation analysis in human cancer. Brief Funct Genomics. 2016;15:432–42. https://doi.org/10.1093/bfgp/elw010.
34. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013;23(3):555–67. https://doi.org/10.1101/gr.147942.112.
35. Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol. 2010;28(10):1106–14. https://doi.org/10.1038/nbt.1681.
36. Heiss JA, Brennan KJ, Baccarelli AA, Téllez-Rojo MM, Estrada-Gutiérrez G, Wright RO, et al. Battle of epigenetic proportions: comparing Illumina's EPIC methylation microarrays and TruSeq targeted bisulfite sequencing. Epigenetics. 2020;15(1-2):174–82. https://doi.org/10.1080/15592294.2019.1656159.

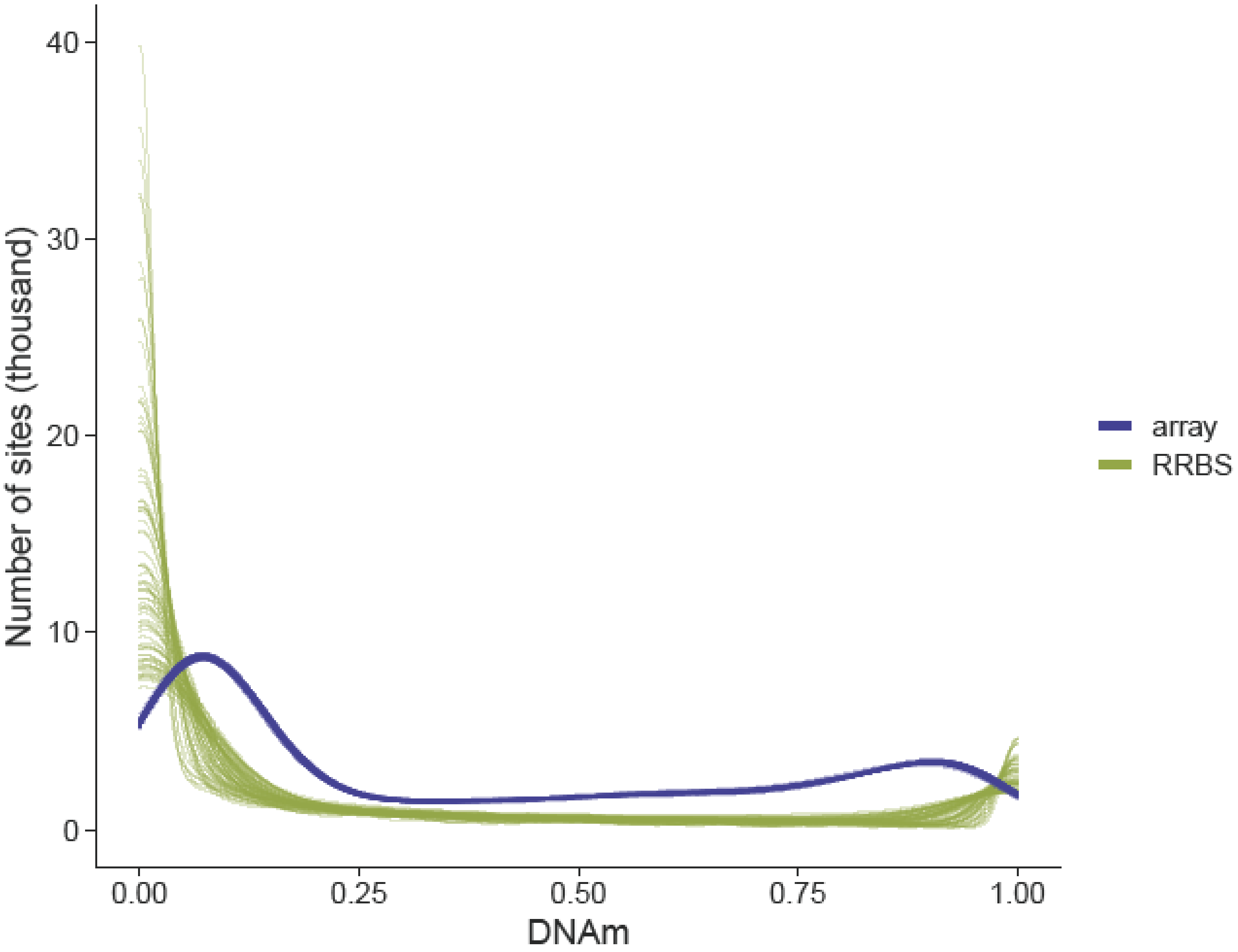
37. Wang T, Guan W, Lin J, Boutaoui N, Canino G, Luo J, et al. A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. Epigenetics. 2015;10(7):662–9. https://doi.org/10.1080/15592294.2015.1057384.
38. Ziller MJ, Hansen KD, Meissner A, Aryee MJ. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. Nat Methods. 2015;12(3):230–2. https://doi.org/10.1038/nmeth.3152.
39. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11(1):587. https://doi.org/10.1186/1471-2105-11-587.
40. Rackham OJL, Dellaportas P, Petretto E, Bottolo L. WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools. Bioinformatics. 2015;31(14):2371–3. https://doi.org/10.1093/bioinformatics/btv114.
41. Chen Y, Pal B, Visvader JE, Smyth GK. Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR. F1000Research. 2018;6:2055.
42. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014;42(8):e69. https://doi.org/10.1093/nar/gku154.
43. Xu Z, Bolick SCE, Deroo LA, Weinberg CR, Sandler DP, Taylor JA. Epigenome-wide association study of breast cancer using prospectively collected sister study samples. J Natl Cancer Inst. 2013;105(10):694–700. https://doi.org/10.1093/jnci/djt045.
44. Castanho I, Murray TK, Hannon E, Jeffries A, Walker E, Laing E, et al. Transcriptional Signatures of Tau and Amyloid Neuropathology. Cell Rep. 2020;30:2040–2054.e5.
45. Krueger F. Trim Galore. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed 7 Nov 2020.
46. Mouse genome mm10 (GRCm38). https://www.ncbi.nlm.nih.gov/assembly/GCF_000001635.20/. Accessed 7 Nov 2020.
47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
48. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
49. R Core Team. R: A Language and Environment for Statistical Computing. 2018. https://www.r-project.org/.
50. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. https://ggplot2.tidyverse.org.
51. Wilke CO. Cowplot R package. https://cran.r-project.org/web/packages/cowplot/index.html. Accessed 7 Nov 2020.
52. Wickham H. tidyr R package. https://cran.r-project.org/web/packages/tidyr/index.html. Accessed 7 Nov 2020.
53. Garnier S, Ross N, Rudis B, Sciaini M, Scherer *C. viridis* R packaage. https://cran.r-project.org/web/packages/viridis/index.html. Accessed 7 Nov 2020.
54. Sanchez G. colortools R package. https://cran.r-project.org/web/packages/colortools/index.html. Accessed 7 Nov 2020.
55. Wickham H. Reshaping data with the reshape package. J Stat Softw. 2007;21:1–20 http://www.jstatsoft.org/v21/i12/.
56. Cavalcante RG, Sartor MA. Annotatr: genomic regions in context. Bioinformatics. 2017;33(15):2381–3. https://doi.org/10.1093/bioinformatics/btx183.
57. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118. https://doi.org/10.1371/journal.pcbi.1003118.
58. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–9. https://doi.org/10.1093/bioinformatics/btu049.
59. Maksimovic J, Gordon L, Oshlack A. SWAN: subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol. 2012;13(6):R44. https://doi.org/10.1186/gb-2012-13-6-r44.
60. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 2014;15(11):503. https://doi.org/10.1186/s13059-014-0503-2.
61. Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA methylation BeadArrays. Nucleic Acids Res. 2013;41(7):e90. https://doi.org/10.1093/nar/gkt090.
62. Fortin JP, Hansen KD. Reconstructing a/B compartments as revealed by hi-C using long-range correlations in epigenetic data. Genome Biol. 2015;16(1):180. https://doi.org/10.1186/s13059-015-0741-y.
63. 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(1):1–21. https://doi.org/10.1186/s13072-016-0107-z.
64. Fortin JP, Triche TJ, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinformatics. 2017;33:558–60. https://doi.org/10.1093/bioinformatics/btw691.
65. Zhou W. SeSAMe. https://github.com/zwdzwd/sesame. Accessed 18 Nov 2020.

Publisher's Note

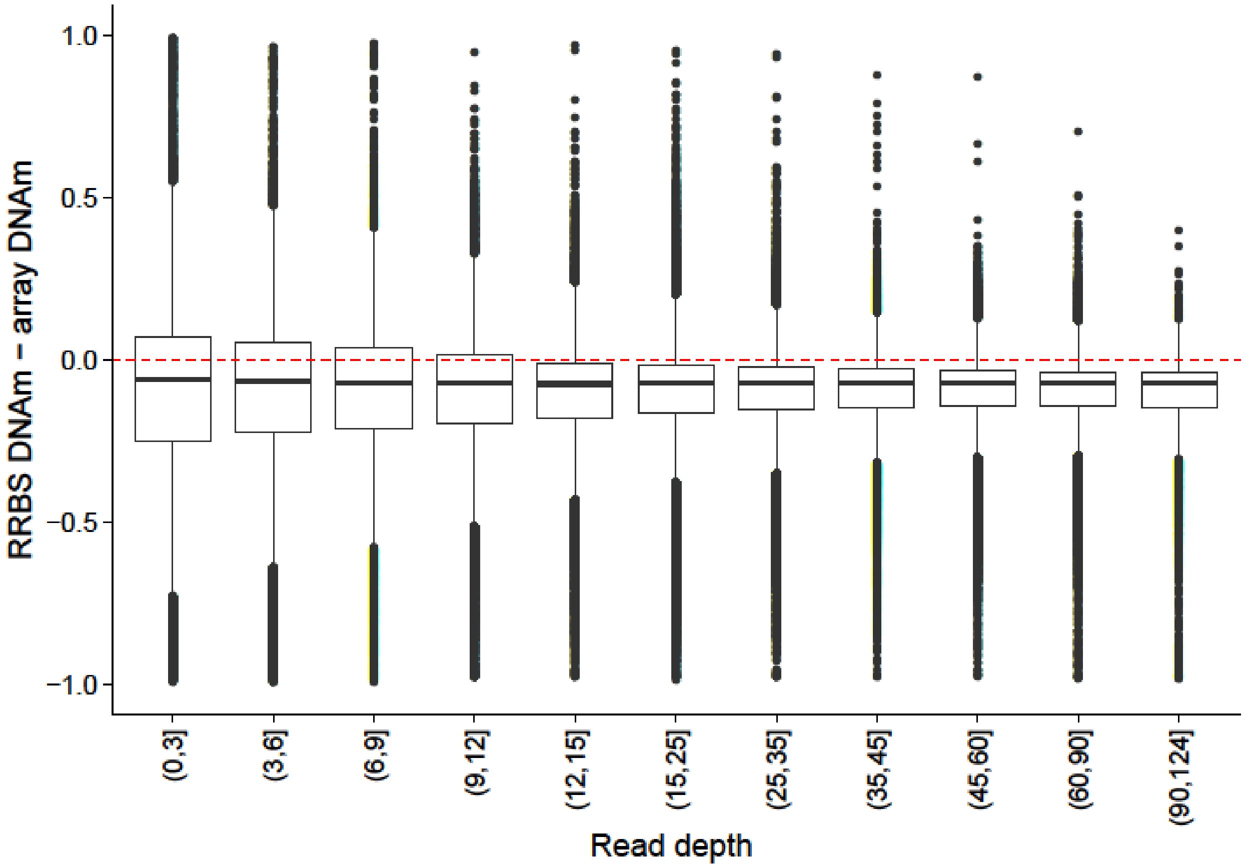
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Figures

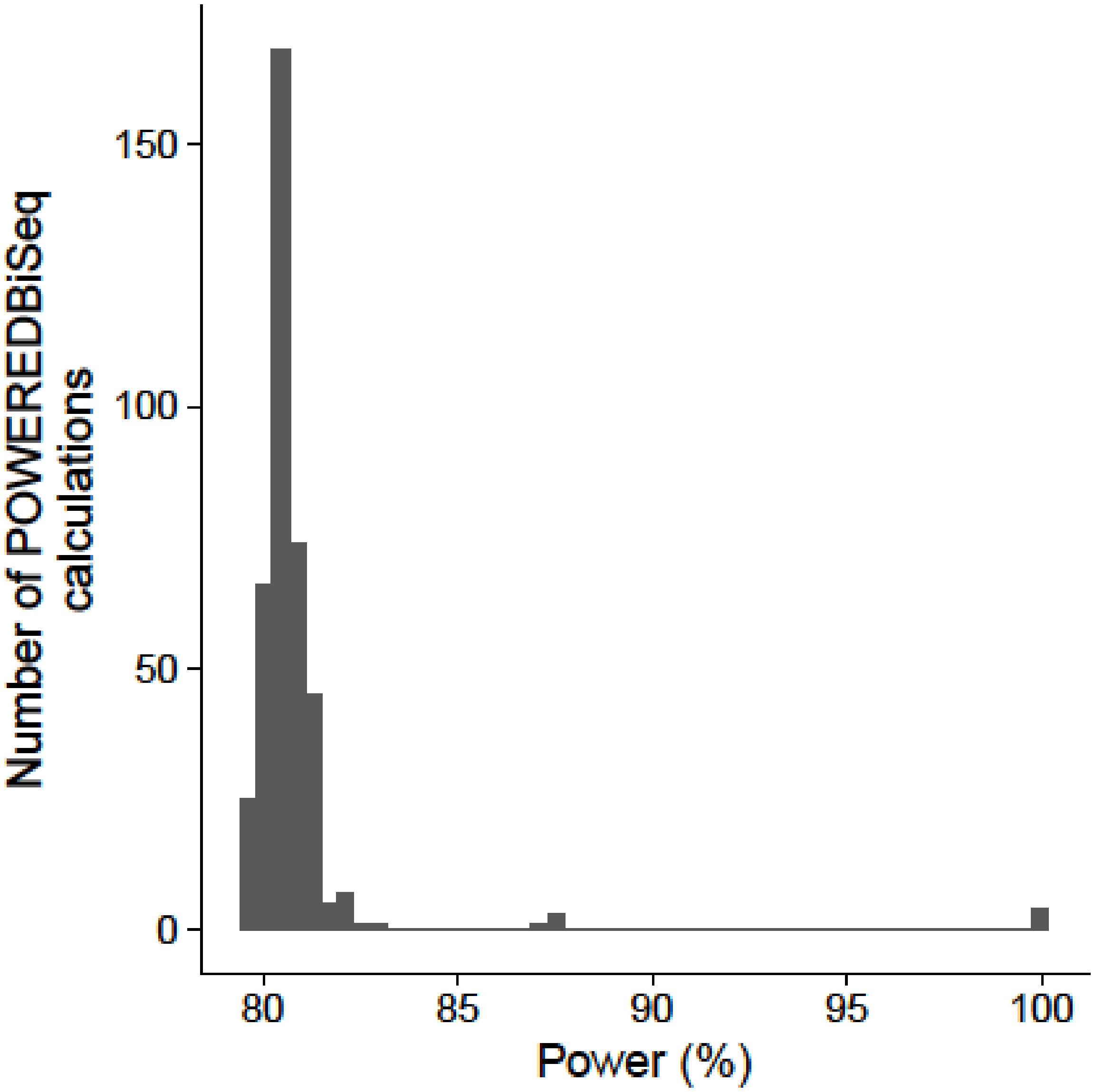
Supplementary Figure 1: The distribution of DNAm levels across the genome profiled using RRBS or a custom array. Density plot of the distribution of DNAm level across all DNAm sites, where each line represents one of the 80 overlapping samples profiled using a custom Illumina array [1] (blue)or RRBS (green).



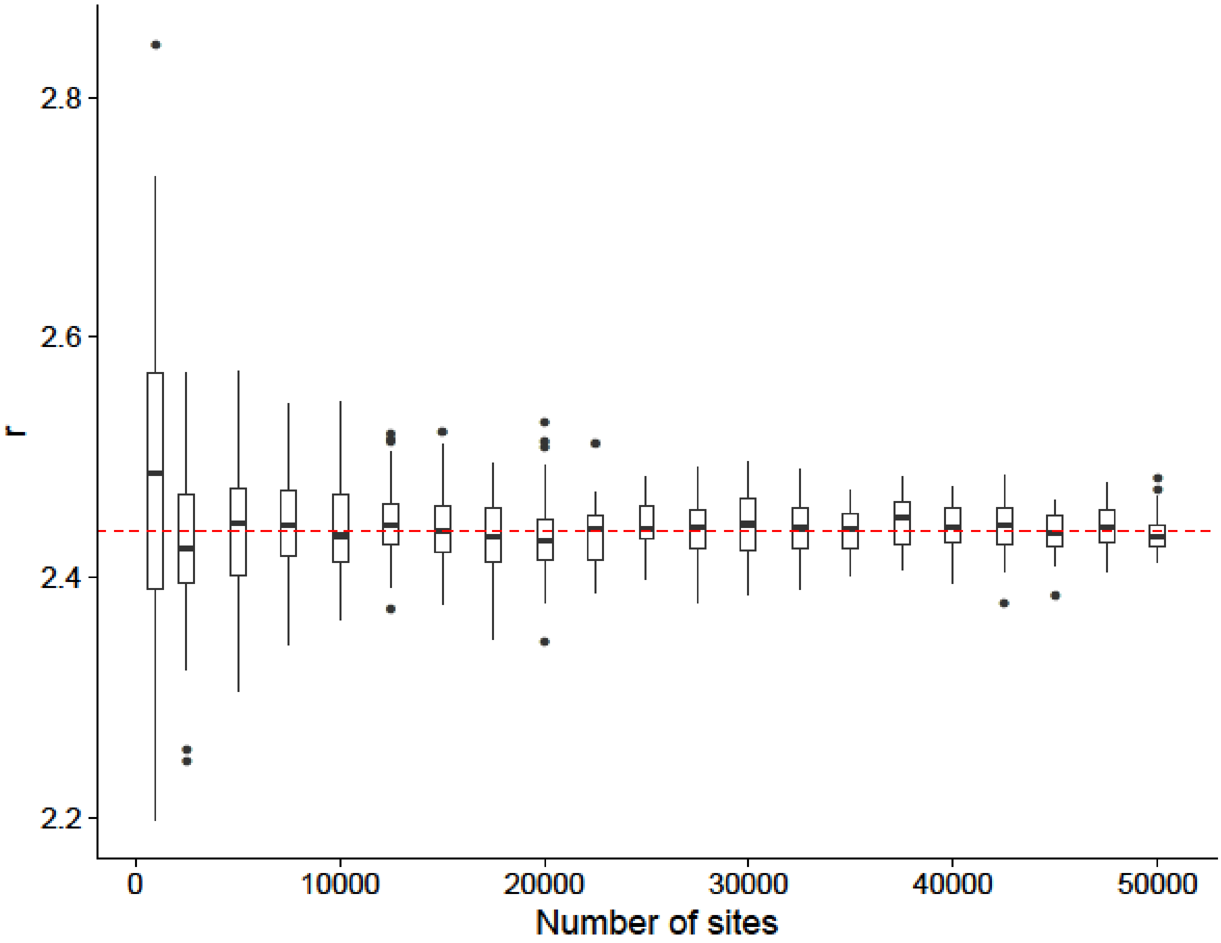
Supplementary Figure 2: DNAm estimates derived from RRBS are on average lower than those from the array platform. Shown are boxplots of the difference in DNAm estimated from the Illumina mammalian array and RRBS DNAm data across all overlapping DNAm points, with DNAm points grouped based on their read depths in the RRBS data.



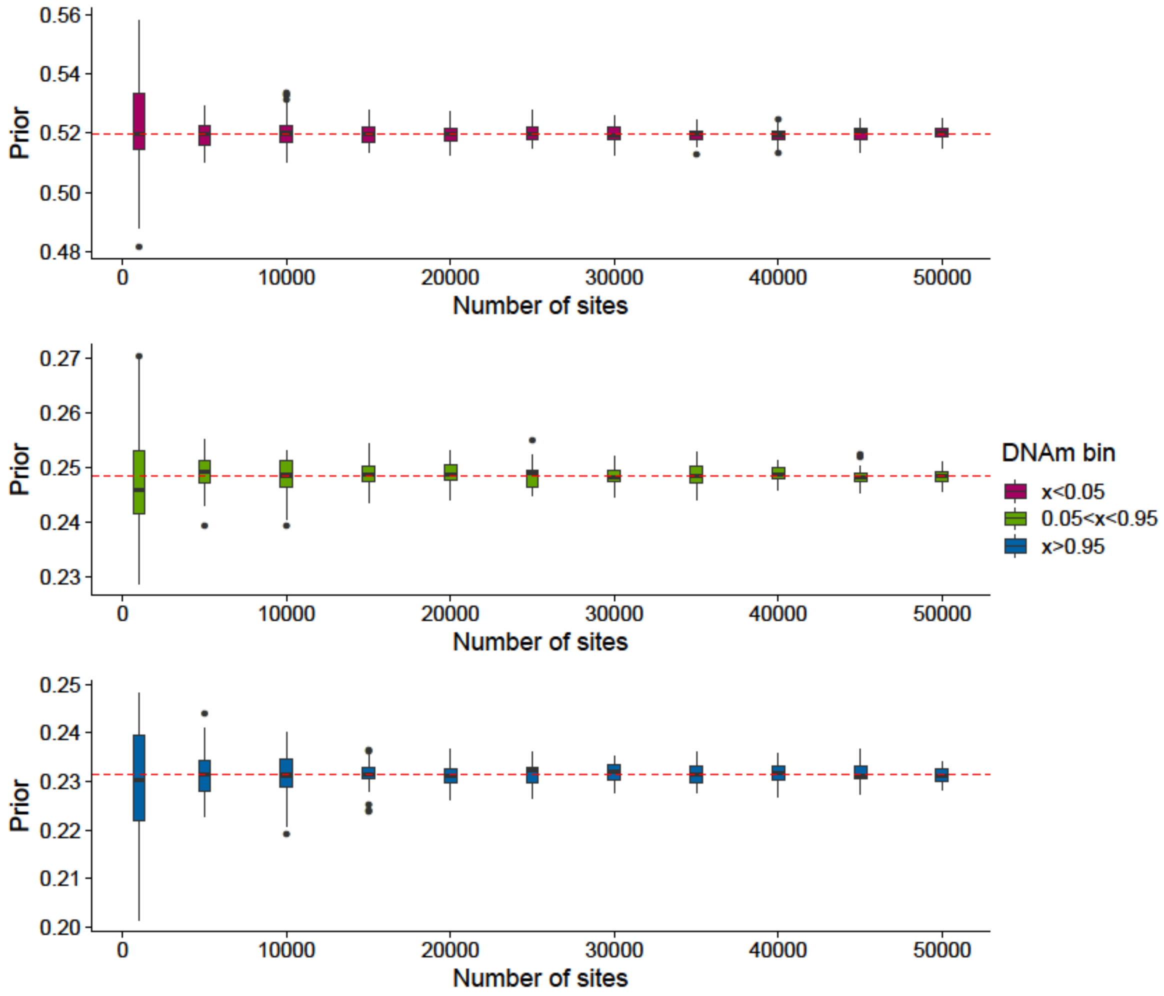
Supplementary Figure 3: A histogram of POWEREDBiSeq calculations showing variability in estimated power. The estimated power of a study was calculated 400 times, using a mean DNAm difference between groups of 0.06 and minimum sample size of 60. Users can calculate bespoke power estimates using their own study-specific parameters in POWEREDBiSeq.



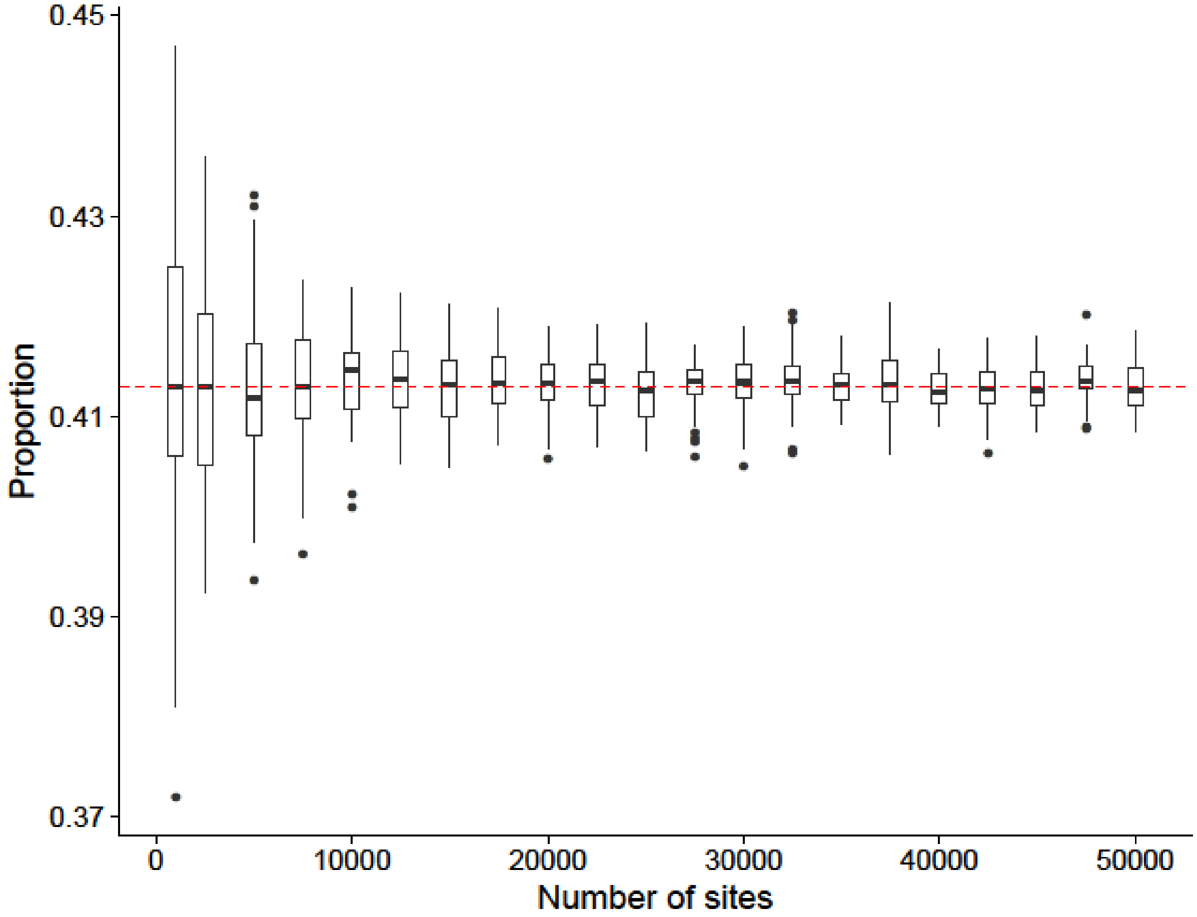
Supplementary Figure 4: r is more accurately estimated when using a larger number of DNAm sites. The negative binomial parameter, r, was calculated from an increasing number of DNAm sites using RRBS data from 125 samples. The red dashed line is the true value of r, calculated from the entire dataset.



Supplementary Figure 5: DNAm priors are more accurately estimated when using more DNAm sites. DNAm priors are the probability that the DNAm value of selected DNAm sites fall within the DNAm bins 0-0.05, 0.05 - 0.95 and 0.95 – 1, shown in maroon, green and blue, respectively. Priors were calculated across 125 samples. The true value for each prior, calculated across the entire dataset, is shown in red.



Supplementary Figure 6: The proportion of DNAm sites remaining is more accurately estimated when using more DNAm sites. The proportion of DNAm sites remaining after filtering by minimum read depth and minimum number of samples calculated from an increasing number of DNAm sites from an RRBS dataset. n was calculated across 125 samples. The red dashed line is the true value of n, calculated from the entire dataset.



Supplementary Table 1: **A summary of parameters used in simulation analysis.**

Plot	μRD	N_1, N_2	ΔμDNAm	μDNAm
A	Variable	30	0.2	0.25
			0.1	
			0.05	
B	20	Variable	0.2	0.25
		5		
		10		
C	25	20	Variable	0.25
		50		
		100		
		500		
D, E	50	80	0.05	Variable
	30			
	10			

Each row refers to a plot in **Figure 6**. Values were chosen so that the variable of interest could be seen to influence power within the scale of the figure.

3. Profiling the accuracy of reference based cellular deconvolution models

3.1 Introduction

DNA methylation studies, including epigenome wide association studies (EWAS) (described in Section 1.3), conducted in bulk tissue can be confounded by cellular heterogeneity, defined here as variability of cell type proportions across a population (as seen in **Figure 1.13** and Section 1.3.8.6). To account for this heterogeneity, cellular deconvolution algorithms can be applied to genome wide DNA methylation profiles, generating variables that quantify cellular composition (Houseman et al., 2012; Newman et al., 2015; Teschendorff et al., n.d.; Guintivano, Aryee and Kaminsky, 2013; Rahmani et al., 2016; Zou et al., 2014; Leek and Storey, 2007; Leek et al., 2012; Houseman, Molitor and Marsit, 2014; Rahmani et al., 2018; Gagnon-Bartsch and Speed, 2012; Levy et al., 2020; Zhang, Wu and Li, 2021; Onuchic et al., 2016; Teschendorff, Zhuang and Widschwendter, 2011; Lutsik et al., 2017; Houseman et al., 2016) (summarised in **Table 1.1**). There are two main types of deconvolution algorithm: reference based, which utilise the DNA methylation profiles of purified cell types within a tissue to estimate cellular composition, and reference free, which do not require data to assess cell type driven variability in the data, although semi-reference free algorithms have also been developed, which contain aspects of both methods. A summary of the specific deconvolution algorithms can be found in Section 1.4.

Reference based deconvolution can be divided into two main stages: first is the data generation stage, in which reference data is obtained by profiling DNA methylation across purified cell types for the tissue of interest. Given the relatively large magnitude of cell type specific differential DNA methylation (Hannon et al., 2021b), only a handful of samples are required per cell type to make up a reference dataset.

In the second stage, here named the model generation stage, DNA methylation sites with cell type specific patterns learned from the reference data are leveraged by a reference based deconvolution algorithm to generate a deconvolution model. This can be used to predict the cellular composition from DNA methylation measured in an input sample, subsequently referred to as a bulk tissue sample, although there is no reason why the algorithm cannot be applied to purified cellular samples. While a number of algorithms have been proposed for reference based deconvolution, Houseman's algorithm is the most commonly used due to it's integration in the quality control (QC) packages *minfi* (Jaffe and Irizarry, 2014) and *wateRmelon* (Pidsley et al., 2013). In the following Section, the general framework for reference based deconvolution is described, highlighting the specific steps within Houseman's algorithm.

3.1.1 Houseman's algorithm

in any one sample, reference based deconvolution assumes the following relationship between bulk and cell type specific DNA methylation profiles:

$$M_{BULK} = \sum_{i=1}^{nTOT} P_i M_{CT_i} \tag{3.1}$$

where:

- M_{BULK} is a matrix with one column containing the genome wide DNA methylation values for a bulk tissue sample, each row of which contains the DNA methylation value at a profiled DNA methylation site.
- i is the index for cell types, where $i \in [1, nTOT]$ and $nTOT$ is the total number of cell types comprising the bulk tissue sample
- P_i is ith value of vector $\underline{P}$ containing cell proportions for all cell types in the bulk sample
- M_{CT_i} is the ith column within matrix M_{CT}, containing the genome wide DNA methylation values for cell type i at the same DNA methylation sites as M_{BULK}

The goal of reference based deconvolution is to estimate $\underline{P}$, denoted $\hat{\underline{P}}$, for user provided input sample M_{BULK}.

To do so, first, an estimate of M_{CT} is required: the reference data, denoted $\hat{M}_{CT}$. Reference data may not contain every possible cell type within a tissue of interest, due to challenges in data sorting using current methods (see Section 4.1.1.1), however, it is important that a sufficient proportion of the cells in the bulk tissue are represented. Here, we say that $\hat{M}_{CT}$ contains DNA methylation data for n purified cell type populations.

As the DNA methylation differences between cell types are stark, not all DNA methylation sites are needed for effective prediction. Reference based deconvolution algorithms select a subset of k cell type specific DNA methylation sites from $\hat{M}_{CT}$, denoted $\tilde{M}_{CT}$, which are used to predict $\underline{\hat{P}}$.

$$\tilde{M}_{CT} = \begin{array}{c} \\ cg_1 \\ \vdots \\ cg_j \\ \vdots \\ cg_k \end{array} \begin{array}{c} \begin{array}{ccccc} \tilde{M}_{CT_1} & ... & \tilde{M}_{CT_i} & ... & \tilde{M}_{CT_n} \end{array} \\ \begin{pmatrix} \tilde{M}_{CT}[1,1] & ... & \tilde{M}_{CT}[1,i] & ... & \tilde{M}_{CT}[1,n] \\ \vdots & & \vdots & & \vdots \\ \tilde{M}_{CT}[j,1] & ... & \tilde{M}_{CT}[j,i] & ... & \tilde{M}_{CT}[j,n] \\ \vdots & & \vdots & & \vdots \\ \tilde{M}_{CT}[k,1] & ... & \tilde{M}_{CT}[k,i] & ... & \tilde{M}_{CT}[k,n] \end{pmatrix} \end{array}$$

where

- cg_j is one of k cell type specific sites selected from the reference data $\hat{M}_{CT}$
- $\tilde{M}_{CT}[j,i]$ is the mean DNA methylation value across all reference data samples of the ith cell type, at the jth site

The default application of Houseman's algorithm is through R packages *minfi* and *wateRmelon*, in which a set of k DNA methylation sites are selected by using an analysis of variance (ANOVA) F-test to compare the DNA methylation profiles of the samples of any one cell type to all other cell types in the reference data. Two sets of DNA methylation sites are selected per cell type, with *numProbes* (an arbitrarily set integer within the R function) sites selected that have significantly higher DNA methylation than all other cell types, and *numProbes* with significantly lower DNA methylation, resulting in $k = 2$ x $numProbes$ x n where n is the total number of cell types within the reference dataset. Where $n = 2$, $k = 2$ x $numProbes$, as sites selected will be unique to both cell types. $\tilde{M}_{CT}$ will be comprised of the average DNA methylation values of each cell type across all k DNA methylation sites selected. By default, *numProbes* is set to 50, and the significance level which each site must be below to be selected is P < 1e-8; if there are not *numProbes* DNA methylation sites that have significant differences, only those that do will be used.

Substituting the values into equation 1.1 gives us the equation to be solved:

$$\tilde{M}_{BULK} = \sum_{i=1}^{n} \hat{P}_i \tilde{M}_{CT_i}$$

where $\tilde{M}_{BULK}$ is the nx1 matrix of bulk DNA methylation values at the sites contained in $\tilde{M}_{CT}$. The equation is solved for $\underline{\hat{P}}$ with the constraint that $\sum_{i=1}^{n} \hat{P}_i \leq 1$. Where all cell types within a tissue are present in the model, the constraint should hold true. Houseman's algorithm uses constrained projection and quadratic programming to solve this equation. In essence, the method applies a least squares minimisation with the above constraint.

3.1.2 Reference datasets for deconvolution

Two tissues commonly used in EWAS are blood and brain tissue, both of which are highly heterogeneous in cellular composition. As such, reference datasets have been generated, which can be used for the reference based deconvolution of bulk brain or blood DNA methylation samples.

3.1.2.1 Blood reference based deconvolution

An overview of the hierarchy of the main white blood cell types can be found in **Figure 3.1**, with cells first divided into granulocytes and agranulocytes. Subclasses of granulocytes include neutrophils, eosinophils, and basophils. Agranulocytes are then further divided into lymphocytes and monocytes, with lymphocyte subclasses being B cells, CD4^{+} T cells, CD8^{+} T cells, and natural killers (NKs) cells. Their general functions and average proportion in whole blood described are in **Figure 3.1**.

The commonly used blood reference panel, referred to in this thesis as **Dataset**

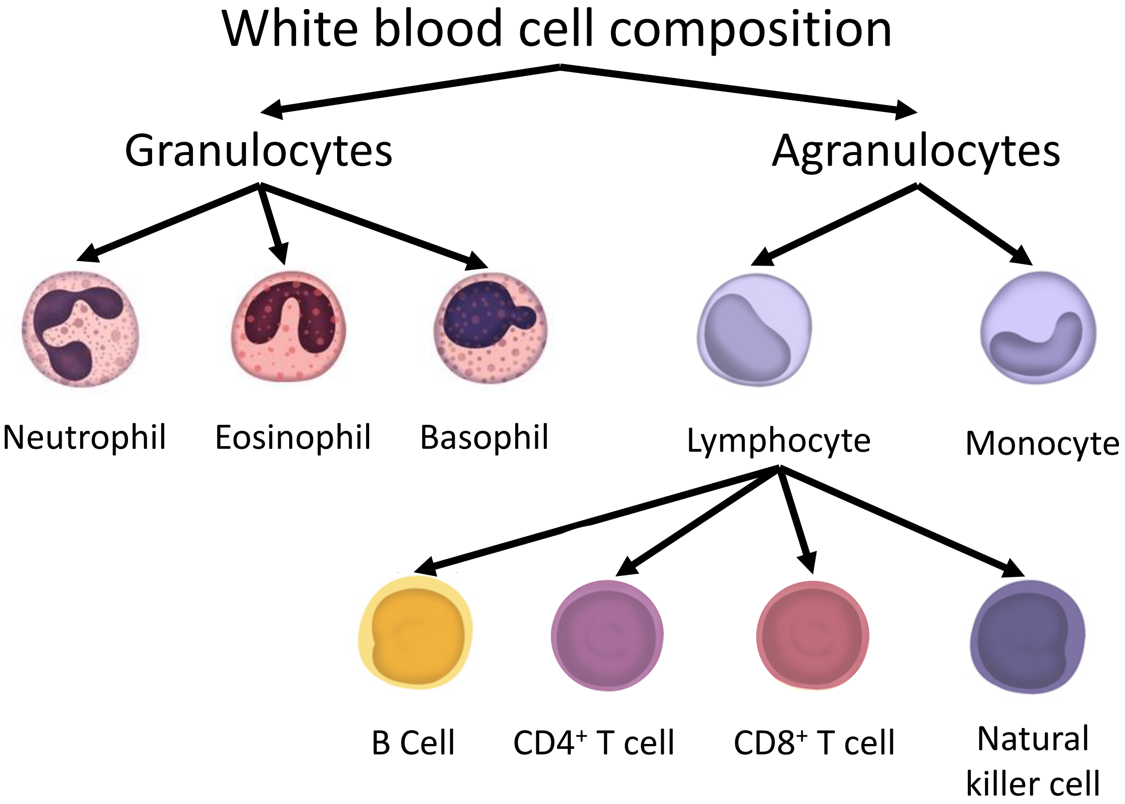


Figure 3.1: **A diagram showing the hierarchy of the main cell types within blood tissue.** Cell images are taken from Sciencefacts.net.

Reinius, contains six cell types: B cells, CD4$^+$ T cells, CD8$^+$ T cells, granulocytes, monocytes and NKs cells.

The purified cell types within this reference panel were derived using a density gradient to divide granulocytes from peripheral blood mononuclear cells (PBMCs), and then magnetic-activated cell sorting (MACS) (Schmitz et al., 1994), which uses antibodies conjugated to magnetic beads to recognise cell surface antigens on the target cell type to separate and purify cell populations within the pbmcs (**Figure 3.2**).

3.1.2.2 Prefrontal cortex reference based deconvolution

An overview of the hierarchy of the main cortical cell types can be found in **Figure 3.3**, with cells first divided into neurons and glia. Subclasses of neurons include GABAergic and glutamatergic neurons. Glia are then further divided into oligidendrocytes, astrocytes and microglia. Their general functions are described in **Figure 3.3**.

The commonly used prefrontal cortex (PFC) reference panel, referred to in this thesis as **Dataset Guintivano**, contains two cell types: neuronal and non-neuronal (glial). PFC tissue was purified using fluorescence-activated cell sorting (FACS), in which cell nuclei are isolated and stained utilising cell type specific nuclear markers, and sorted based on fluorescence. Here, the NeuN marker is utilised, dividing populations into a neuronally enriched, NeuN+, and an 'other' population, NeuN- (see Section 3.3 for full description of the reference datasets, defined as **Dataset Guintivano**). Further fluorescence-activated nuclear sorting (FANS) markers exist that can be used in conjunction with NeuN staining to further separate both the NeuN+ and NeuN- populations, described in Section 4.1.1.1.

3.1.3 Limitations to existing deconvolution algorithms

There is an assumption that, when applying reference based deconvolution, the resulting cellular composition estimates will be sufficiently accurate to be of utility as covariates in downstream analyses. What's more, it is assumed that models will perform comparably well in all cohorts comprised of the relevant tissue and across all samples in a cohort. In the majority of cases, deconvolution is used because the true cellular proportions of a sample are unknown and it is therefore not possible to validate model performance.

Table 3.1: **A summary of the nucleated cell types within blood and their functions.** Each cell type has many functions, but in general all white blood cells play a role in inflammation and disease. The hierarchy of cell types can be found in **Figure 3.1**.

Cell type	General function	Estimated proportion
Granulocytes		51-78%
Neutrophils	identify signs of infection in the body	50-70%
Eosinophils	participate in immediate allergic reactions, as well as modulating inflammatory responses	1-6%
Basophils	play an important role in immune surveillance	<1%
Lymphocytes		20-50%
B cells	production of antibodies and antigen presentation	
CD4$^+$ T cells	assist other lymphocytes, including activating the maturation of B cells	
CD8$^+$ T cells	can destroy virus-infected or tumor cells	
Natural killers	same as CD8$^+$ T cells, without requiring antibody activation	
Monocytes	precursor cells for macrophages, which can phagocytose (ingest and digest) dead cells and bacteria	2-10%

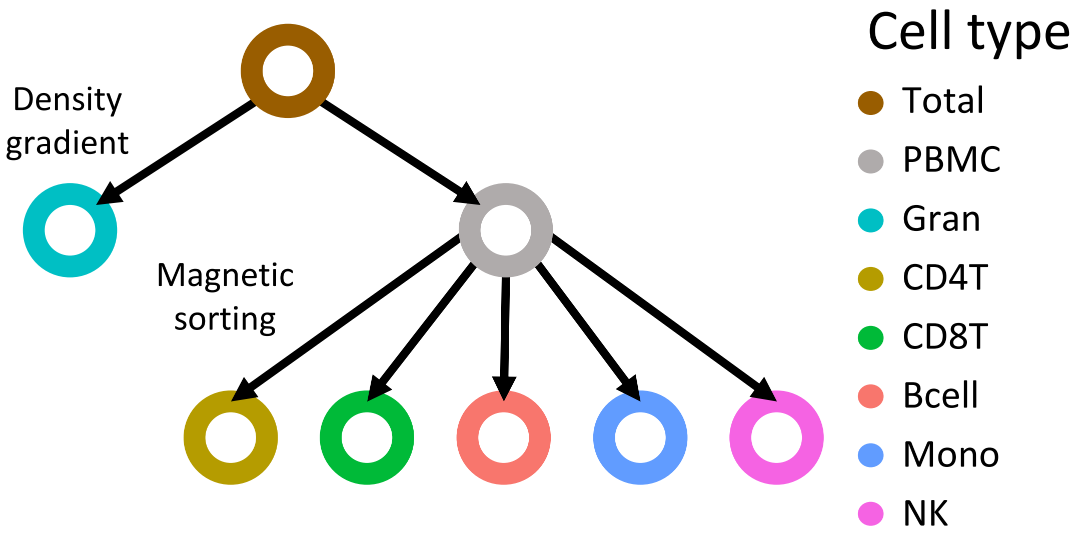


Figure 3.2: **A diagram of the cellular hierarchy of purification of Dataset Reinius, sorted using MACS.** Total - unsorted blood, PBMC, granulocyte (Gran), CD4^{+} T cell (CD4T), CD8^{+} T cell (CD8T), B cell (Bcell), monocyte (Mono), NK cell.

Table 3.3: **A summary of the cell types within the brain and their functions.** The hierarchy of cell types can be found in **Figure 3.3**. The estimated ratio of neurons to glial cells across the developed brain as a whole is 1:1 (Bartheld, Bahney and Herculano-Houzel, 2016).

Cell type	General function	Citation
Neurons		
GABAergic	exitatory neurons dependent on glial signals for their signal transmission	(Turko et al., 2019)
Glutamatergic	inhibitory neurons	
Glia		
Oligodendrocytes	producers of myelin, which insulates the neuronal axon	(Nave, 2010)
Microglia	phagocytose many products in the brain	(Jäkel and Dimou, 2017; Lenz and Nelson, 2018)
Astrocytes	most abundant type of glia, with functions including playing an important role in water and ion homeostasis	(Zeng and Sanes, 2017; Kimelberg, 2010)

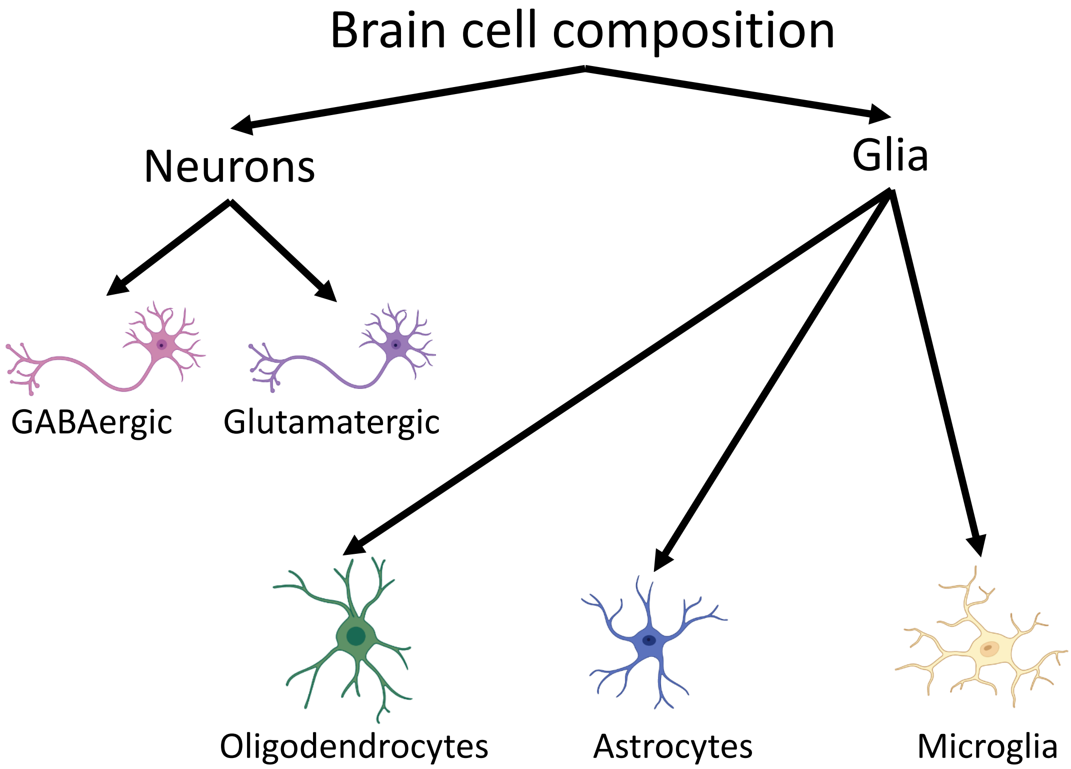


Figure 3.3: A diagram showing the hierarchy of the main cell types within brain tissue.

This may result in inaccurate composition estimates being carried forward for analysis, for example, when adjusting for cellular heterogeneity in EWAS. Inaccurate cellular composition estimates may lead to false positive results, firstly, by not adequately adjusting for cellular heterogeneity, and secondly, by increasing variance in the data and inducing false positives through incorrect adjustment.

It is not currently possible to tell how well a deconvolution model performs in datasets where the true cellular composition is unknown. However, deconvolution models are an important solution where the cellular composition of samples is unknown; both the blood and PFC models are commonly used to combat the issue of cellular heterogeneity in EWAS (Lunnon et al., 2014; Montano et al., 2016; Logue et al., 2017; Levine et al., 2018; Shireby et al., 2020; Smith et al., 2021; Policicchio et al., 2020b).

A potential driver of decreased deconvolution accuracy within a dataset, would be a disparity between DNA methylation profiles of the reference data and input samples which may be caused by common drivers of differential DNA methylation, such as biological differences, including sex (McCarthy et al., 2014; Liu et al., 2010), age (Ciccarone et al., 2018), and ethnicity (driven by genetic factors) (Fraser et al., 2012), or technical differences, such as batch or operator effects. The magnitude of DNA methylation differences are largest between cell types compared to the other possible factors listed above, and so it is assumed that deconvolution will still perform sufficiently well despite minor deviations induced by biological or technical variation. However, this may not be a valid assumption across all datasets and samples. In scenarios where the training data for the deconvolution model does not resemble the (unmeasured) cell type specific profiles for the input data, the application of the model is inappropriate and the resulting cellular composition estimates will not be biologically meaningful. Similarly, poor sample quality of the sample to be deconvoluted, where the resulting DNA methylation profile has large measurement error, is expected to lead to poorly estimated cellular composition.

Another potential reason for a reference based deconvolution model to perform less accurately is if the reference data is incomplete. As stated in Section 3.1.1, Houseman's algorithm (and others (Newman et al., 2015; Teschendorff et al., n.d.; Guintivano, Aryee and Kaminsky, 2013)) operate under the constraint that $\sum_{i=1}^{n} \hat{P}_i \leq 1$, which is the assumption that the reference data contains all, if not close to all, of the cell types present in the bulk tissue. In blood, findings by Reinius et al., 2012 suggest that 4.68% ± 2.83 of whole blood is unaccounted for by the six cell types in the blood deconvolution model described in Section 3.1.2.1 (Reinius et al., 2012) (obtained from **Table 3.5**). The importance of unprofiled cell types to deconvolution accuracy is unknown and is likely to depend on how abundant the missing cell types are.

Additionally, while reference datasets may be representive of all of the main cell types present in the bulk tissue, due to the hierarchical nature of cellular composition (see **Figures 3.1** and **3.3**), the DNA methylation profiles of purified cell types may still be heterogeneous. For example, the blood deconvolution reference data contains the DNA methylation profile for granulocytes, which is a broad class of cell comprising of three subtypes, neutrophils, eosinophils, and basophils. Similarly, PFC deconvolution classifies a single category for neurons, which could be further divided into GABAergic and glutamatergic neurons, and glial cells (i.e. non-neurons) which can be separated into oligidendrocytes, astrocytes and microglia. It is not known how this might affect the composition estimates, although one could assume that if the cellular composition of a purified but heterogeneous cell type was highly different to that in an independent sample (e.g. the PFC reference neuronal population is comprised largely of GABAergic neurons, and the test bulk sample is instead comprised largely of glutamatergic neurons) deconvolution would be negatively impacted.

3.1.4 Chapter aims

The main challenge when using reference based deconvolution to predict cell type composition is that the accuracy of predictions is unknown. When using these predictions to adjust for cellular composition in EWAS, any inaccuracy could further confound results rather than reducing cell type confounding. As it stands, publicly available reference datasets and the deconvolution models generated from them are validated in the initial publication in which they are presented, relying on data for which cellular composition is known. However, it is possible that this validation only holds in specific scenarios, and extrapolating the model to independent datasets characterised by different technical or biological variables may reduce deconvolution accuracy. Therefore, these deconvolution models may not be as accurate when applied to a novel bulk dataset. Currently there is no way to assess this; in this Chapter we aim to establish an error metric (Cetygo) for reference based deconvolution which will allow for the assessment of deconvolution model performance under various scenarios.

The specific Chapter aims are to:

1. establish a framework for quantifying deconvolution accuracy that is agnostic of tissue or reference panel
2. characterise Cetygo's ability in blood and brain tissue to:
    - detect inaccurate estimates of cellular composition
    - compare between different deconvolution models
    - confirm applicability of models to independent datasets
3. provide guidelines on applying and interpreting Cetygo to assess prediction quality

3.2 Cetygo: the concept behind the error metric

The objective of reference based deconvolution is to estimate the proportion of relevant cell types from a genome wide profile of DNA methylation.

Mathematically, deconvolution algorithms solve:

$$\tilde{M}_{BULK} = \sum_{i=1}^{n} \hat{P}_i \tilde{M}_{CT_i} \tag{3.2}$$

for $\underline{\hat{P}}$, where:

- $\tilde{M}_{BULK}$ is a matrix with one column containing the genome wide DNA methylation values across an input bulk sample at the k DNA methylation sites contained in $\tilde{M}_{CT}$
- i is the index for cell types within the tissue, where $i \in [1, n]$ and n is the total number of purified cell types in the reference data
- $\hat{P}_i$ is ith value of vector $\underline{\hat{P}}$ containing the estimated cell proportions for all n cell types available in the reference data $\tilde{M}_{CT_i}$ used to develop the deconvolution model
- $\tilde{M}_{CT_i}$ is a column within the deconvolution model matrix $\tilde{M}_{CT}$, containing the DNA methylation values at the k algorithm-selected cell type specific DNA methylation sites for each ith cell type population available in the reference data

with constraint:

$$\sum_{i=1}^{n} \hat{P}_i \leq 1$$

Once $\underline{\hat{P}}$ has been derived, it can be substituted back into equation 3.2 to calculate the expected DNA methylation profile of $\tilde{M}_{BULK}$:

$$\hat{M}_{BULK} = \sum_{i=1}^{n} \hat{P}_i \tilde{M}_{CT_i}$$

To quantify the accuracy of the cellular deconvolution, the root mean squared error (RMSE) was calculated between the experimentally observed DNA methylation levels for the bulk tissue, $\tilde{M}_{BULK}$, and the expected DNA methylation profile, $\hat{M}_{BULK}$, calculated using the estimated cell type proportions $\underline{\hat{P}}$ and reference data, i.e. the DNA methylation profile expected with that combination of cell types:

$$Cetygo = RMSE(\tilde{M}_{BULK}, \hat{M}_{BULK})$$

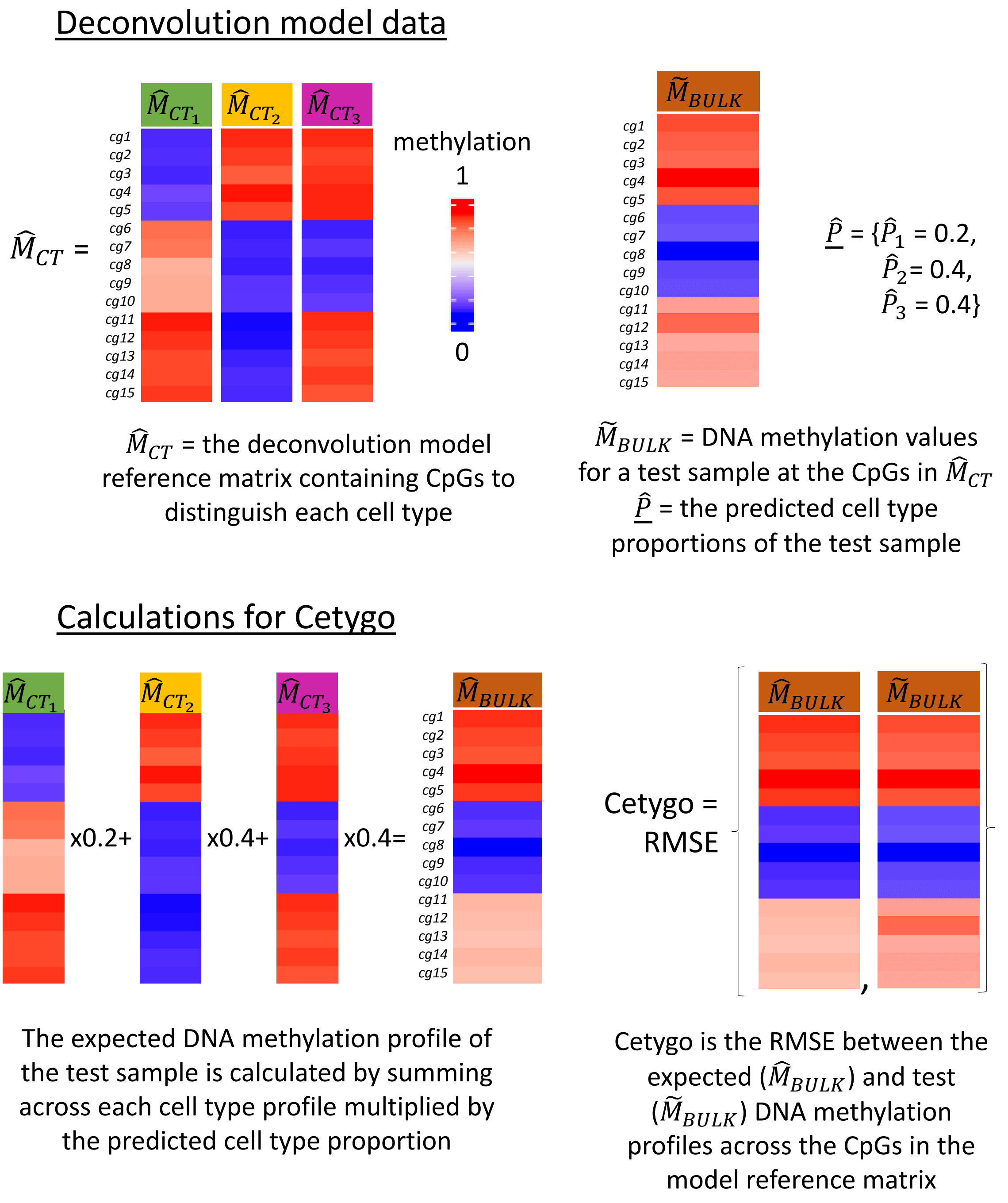

Deconvolution model data
$\widehat{M}_{CT_1}$
$\widehat{M}_{CT_2}$
$\widehat{M}_{CT_3}$
$\widehat{M}_{CT}$ =
cg1
cg2
cg3
cg4
cg5
cg6
cg7
cg8
cg9
cg10
cg11
cg12
cg13
cg14
cg15
methylation
1
0
$\widetilde{M}_{BULK}$
$\underline{\widehat{P}}$ = {$\widehat{P}_1$ = 0.2, $\widehat{P}_2$= 0.4, $\widehat{P}_3$ = 0.4}
$\widehat{M}_{CT}$ = the deconvolution model reference matrix containing CpGs to distinguish each cell type
$\widetilde{M}_{BULK}$ = DNA methylation values for a test sample at the CpGs in $\widehat{M}_{CT}$
$\underline{\widehat{P}}$ = the predicted cell type proportions of the test sample
Calculations for Cetygo
x0.2+
x0.4+
x0.4=
$\widehat{M}_{BULK}$
Cetygo = RMSE
$\widehat{M}_{BULK}$
$\widetilde{M}_{BULK}$
The expected DNA methylation profile of the test sample is calculated by summing across each cell type profile multiplied by the predicted cell type proportion
Cetygo is the RMSE between the expected ($\widehat{M}_{BULK}$) and test ($\widetilde{M}_{BULK}$) DNA methylation profiles across the CpGs in the model reference matrix

Figure 3.4: A diagram explaining the Cetygo workflow. The heatmaps represent DNA methylation proportion, where red is fully methylated and blue is unmethylated. 'Deconvolution model data' contains the data input and output applying a deconvolution model: $\tilde{M}_{BULK}$ as the input sample, $\hat{M}_{CT}$ as the model reference matrix, and $\underline{\hat{P}}$ as the predicted cell type proportions of $\tilde{M}_{BULK}$. RMSE - root mean squared error.

3.3 An overview of the deconvolution models and datasets utilised in Chapter 3

For proof of principle of Cetygo, this Chapter utilises the Houseman algorithm (described in detail in Section 3.1.1) to generate deconvolution models in two different bulk tissues, whole blood and PFC.

A blood deconvolution model, referred to as **Model 6CellBlood**, was generated using **Dataset Reinius** (Reinius et al., 2012), which contains 36 purified blood samples from 6 individuals across 6 cell types: B cells, CD4$^+$ T cells, CD8$^+$ T cells, granulocytes, monocytes and NK cells. The mean age of individuals is 38 ± 13.6 years and all individuals were male. Cell type populations were purified using a density gradient and MACS, after which DNA methylation was profiled using the Human Methylation 450K BeadChip (HM450) array. The data was obtained as matrix of unnormalised βs. The dataset is the default HM450 reference dataset for blood within *minfi* and *wateRmelon*. In this Chapter, this dataset was utilised as the reference data to train and test **Model 6CellBlood**, a cell type deconvolution model for blood. **Model 6CellBlood** was generated using five of the six individuals (details on model generation can be found in Section 3.10.1.1), keeping a single sample from each cell type aside for testing from one individual. The resulting model contained 600 predictive DNA methylation sites. This model was then used to assess Cetygo as an error metric for cellular deconvolution.

A PFC deconvolution model, referred to as **Model 2CellPFC**, was generated using **Dataset Guintivano** (Guintivano, Aryee and Kaminsky, 2013), which contains 58 samples derived from post-mortem PFC tissue. FANS sorting was applied across 29 individuals, utilising the antibody NeuN, resulting in 29 NeuN+ and 29 NeuN- samples. DNA methylation profiling was carried out using the HM450 array and was obtained as a matrix of unnormalised βs. The mean age of individuals was 32.1 ± 15.9 years, the ratio of males to females was 14:15, and the ratio of African to Caucasian individuals was 6:23. This dataset makes up the reference data for deconvolution algorithm and model, CETS (Guintivano, Aryee and Kaminsky, 2013), as well as being the reference data applied within *minfi* and *wateRmelon*, and is the current standard for cortex deconvolution. For the purposes of model generation in this thesis, the individuals in the dataset had been divided into training and testing (details on model generation are in Section 3.10.1.2), resulting in 30 training and 28 testing samples. The training samples were utilised to generate **Model 2CellPFC**, predicting the proportion of NeuN+ and NeuN- across cortex tissue. **Model 2CellPFC** contained 100 predictive DNA methylation sites.

Model 6CellBlood and **Model 2CellPFC** were utilised to assess how Cetygo varied across blood and PFC datasets, respectively. The testing data from both **Dataset Reinius** and **Dataset Guintivano** are utilised in this Chapter to simulate bulk blood and PFC samples, respectively (see Section 3.4). This Chapter also utilises independent datasets for further testing of Cetygo, which are described below:

Dataset GEO contains 20960 samples from 225 datasets, collated by Tyler Gorrie-Stone from publicly available HM450 array data stored on the Gene Expression Omnibus (GEO) server (the GEO accession numbers and sample sizes for which can be found in **Table 7.4**). The data was obtained as a matrix unnormalised βs. All tissue annotations were obtained through a regular expression search for the term "Tissue" or "Source". The dataset is used to test whether Cetygo can distinguish when a tissue with no cell types in the deconvolution model used has been inaccurately deconvoluted. For that reason, all samples without a tissue annotation were removed from the dataset, resulting in 43 tissue types, including 11514 blood and 1589 brain samples. The data was filtered to the subset of samples that contained all the predictive DNA methylation sites in the model to be applied, resulting in 16662 when applying **Model 6CellBlood**, and 19131 samples when applying **Model 2CellPFC**.

Dataset E-Risk was generated to assess the covariation of DNA methylation across peripheral cells and tissues (Hannon et al., 2021b). It contains 173 from purified blood cell type samples (B cells (n = 28), CD4^{+} T cells (n = 28), CD8^{+} T cells (n = 28), monocytes (n = 28), and granulocytes (n = 29)). Samples were 18 or 19 years old and the ratio of male to female individuals was 12:17. DNA methylation was profiled using the EPIC array and unnormalised βs were used.

Dataset PPMI, from the Parkinson's Progression Marker Innitiative (PPMI) cohort, contains 524 whole blood samples taken from healthy controls, patients with early Parkinson's disease (PD) and subjects without evidence of dopaminergic deficiency (Marek et al., 2011). The mean age of individuals is 61.6 ± 10 years and the ratio of male to female individuals was 352:172. DNA methylation was profiled using the EPIC array and unnormalised βs were used.

Dataset EXTEND was acquired from the National Institute for Health Research (NIHR) funded Exeter 10,000 study, in which whole blood samples were obtained from healthy individuals. Here, 1234 samples had DNA methylation quantified on the EPIC array. The data was preprocessed and normalised using the pipeline described in Section 3.10.2. The mean age of individuals is 56.3 ± 11.7 years and the ratio of male to female individuals was 489:686.

Dataset Understanding Society was acquired from the Understanding Society cohort, *'The UK household longitudinal study'*, in which whole blood samples were obtained (Hannon et al., 2018). The dataset contains 1175 samples, profiled using the EPIC array, preprocessed and normalised using the pipeline described in Section 3.10.2. The mean age of individuals is 58 ± 15 years and the ratio of male to female individuals was 591:643.

Dataset IoP, generated to investigate DNA methylation differences associated with schizophrenia (SZ) and first episode psychosis (FEP) (Hannon et al., 2021a), contains 799 samples, profiled using the HM450 array, preprocessed and normalised using the pipeline described in Section 3.10.2. Using genotype information, the dataset was separated into 376 European and 423 non-European samples (methodology described in Section 3.10.3). The mean age of individuals is 44 ± 11.3 years and the ratio of male to female individuals was 503:296.

Dataset EUGEI, generated to investigate DNA methylation differences associated with FEP (Hannon et al., 2021a), contains 934 samples, profiled using the EPIC array, preprocessed and normalised using the pipeline described in Section 3.10.2. Using genotype information, the dataset was separated into 634 European and 300 non-European samples. The mean age of individuals is 35.2 ± 12.8 years and the ratio of male to female individuals was 512:422.

Dataset Pai, generated to investigate DNA methylation differences associated SZ and bipolar disorder (Pai et al., 2019), contains 100 FANS sorted NeuN+ post-mortem PFC samples. The mean age of individuals was 47.6 ± 10.5 years, the ratio of males to females was 75:25. DNA methylation was profiled on the EPIC array and unnormalised βs were used.

Dataset BDR purified contains 107 samples of post-mortem PFC tissue, originating from the Brains for Dementia Research (BDR) cohort, the aim of which is to better understand how dementia affects the brain. Post-mortem PFC tissue from 28 individuals has been FANS sorted, utilising the antibodies, NeuN, SOX10 and IRF8, a microglial marker. resulting in 27 NeuN+ samples, 28 Sox10+ samples, 21 Double- samples, 3 IRF8+ (NeuN-/Sox10-/IRF8+) samples, 2 Triple- (NeuN-/Sox10-/IRF8-) samples, and 26 Total, unsorted samples, after the removal of poor quality samples (using the pipeline described in Section 3.10.2, without normalisation). DNA methylation was profiled on the EPIC array and unnormalised βs were used. The mean age of individuals was 80.8 ± 9.16 years, the ratio of males to females was 46:61.

Of note,all datasets had DNA methylation quantified using the Illumina array platform, which uses a bisulfite conversion step that does not distinguish between hydroxy- and methylated cytosines, and therefore what is referred to as DNA methylation is technically the sum of 5-methylcytosine and 5-hydroxymethylation. Further details on these datasets, including the GEO accession numbers for publicly available datasets, and information on which datasets were generated internally by the Complex Disease Epigenomics group can be found in **Table 6.3**.

3.4 Validation of Cetygo using simulated input data

To assess whether Cetygo indexed prediction accuracy, its performance was profiled across manufactured scenarios where cellular composition is fixed and known. To simulate the DNA methylation profile of an input sample, denoted $\tilde{M}_{SIM}$, the following equation was applied:

$$\tilde{M}_{SIM} = \sum_{i=1}^{n} P_i \tilde{M}_{CT.TEST_i} \tag{3.3}$$

where

- P_i is the assigned cell type proportion of model cell type i within vector $\underline{P}$
- i is the index for cell types within the tissue, where $i \in [1, n]$ and n is the total number of cell types
- $\tilde{M}_{CT.TEST_i}$ is the cell type specific DNA methylation profile of cell type i at the k DNA methylation sites included in model matrix $\tilde{M}_{CT}$, derived from testing data, i.e. samples not included in model generation. If the number of testing samples for each cell type was greater than one, the mean DNA methylation profile was used.

3.4.1 Simulation validation using Model 6CellBlood

To simulate bulk blood data, **Dataset Reinius** testing data was used to generate the cell type specific DNA methylation model matrix ($\tilde{M}_{CT}$ in equation 3.3), which contained a DNA methylation profile for six blood cell types, B cells, CD4$^+$ T cells, CD8$^+$ T cells, granulocytes, monocytes and NKs.

3.4.1.1 Cetygo can distinguish ‘noise’ from true cellular heterogeneity

One fundamental property that Cetygo requires to be of use as a deconvolution error metric is the ability to index the accuracy of the estimated cellular proportions. To test this, samples were simulated with increasing levels of noise (random DNA methylation signals at increasing proportions that did not correlate with cell type specific signal).

In reality, this simulated noise could represent technical or biological noise, or even the absence of an abundant cell type from the reference data.

To simulate blood samples with increasing noise, equation 3.3 was utilised. Noise was treated as an additional cell type, and as such was included as an additional column to the cell type specific DNA methylation matrix, $\tilde{M}_{CT.TEST}$. The noise DNA methylation profile was generated by randomising the DNA methylation sites of a randomly selected reference sample within the testing data of **Dataset Reinius**. The proportion of noise in the samples being simulated was increased systematically across samples, from 0 to 0.95 in steps of 0.05. The remaining fraction of the blood sample was assigned to the average blood cell type proportions found by Reinius et al., 2012 across whole blood (**Table 3.5**). The relative proportions of the non-noise cell types were kept constant so as to only see the effect of noise, rather than any variability that might arise due to the cell types being predicted, as some cell types may be more likely to have accurate prediction than others (explored in Section 3.4.1.3). A total of 20 samples were simulated.

The cellular proportions of each simulated sample was predicted using **Model 6CellBlood**. A strong linear relationship was observed between Cetygo and the proportion of noise (**Figure 3.5A**), as well as between Cetygo and prediction accuracy, quantified as the RMSE of assigned and predicted proportions across all cell types (Cor = 0.997, **Figure 3.6**). This demonstrates that Cetygo meets the aims of indexing how accurate model predictions are.

Panel **B** of **Figure 3.5** contains the predicted cell type proportions of each simulated sample, with colours representing the different cell types, and panel **C** contains the true proportions, with noise in grey. Panel **B** shows that, despite the constraint that the predicted proportions sum to 1, the total sum of predicted proportion across a sample decreases as the proportion of noise increases. This might be expected as the deconvolution model should not be able to accurately predict the proportions of noise as its DNA methylation profile is not within the reference dataset. It also suggests that the sum of the estimated proportions may also be an indicator of an inaccurate estimate resulting from poor quality data.

However, while correlating highly (Cor = 0.977), the difference between 1 and

Table 3.5: **Results from Reinius et al., 2012 showing the average cell type proportions of whole blood across six samples.** Adapted from Reinius et al., 2012 to include only B cells, CD4$^+$ T cells, CD8$^+$ T cells, monocytes, granulocytes and NKs. Percentages to not sum to 100 due to a lack of purify in flow cytometry.

Individual	CD4T	CD8T	Mono	Bcell	NK	Gran
1	13.4	11.6	5.8	2.1	4.0	55.0
2	14.0	3.9	10.1	3.0	3.4	62.0
3	11.2	4.1	1.3	1.9	1.0	73.3
4	11.3	5.6	2.6	4.3	1.6	68.6
5	11.3	3.6	5.3	1.6	0.7	75.9
6	19.3	8.0	7.3	5.2	3.9	54.7
mean	13.4	6.13	5.40	3.10	2.43	64.9
SD	3.12	3.13	3.17	1.44	1.5	9.19

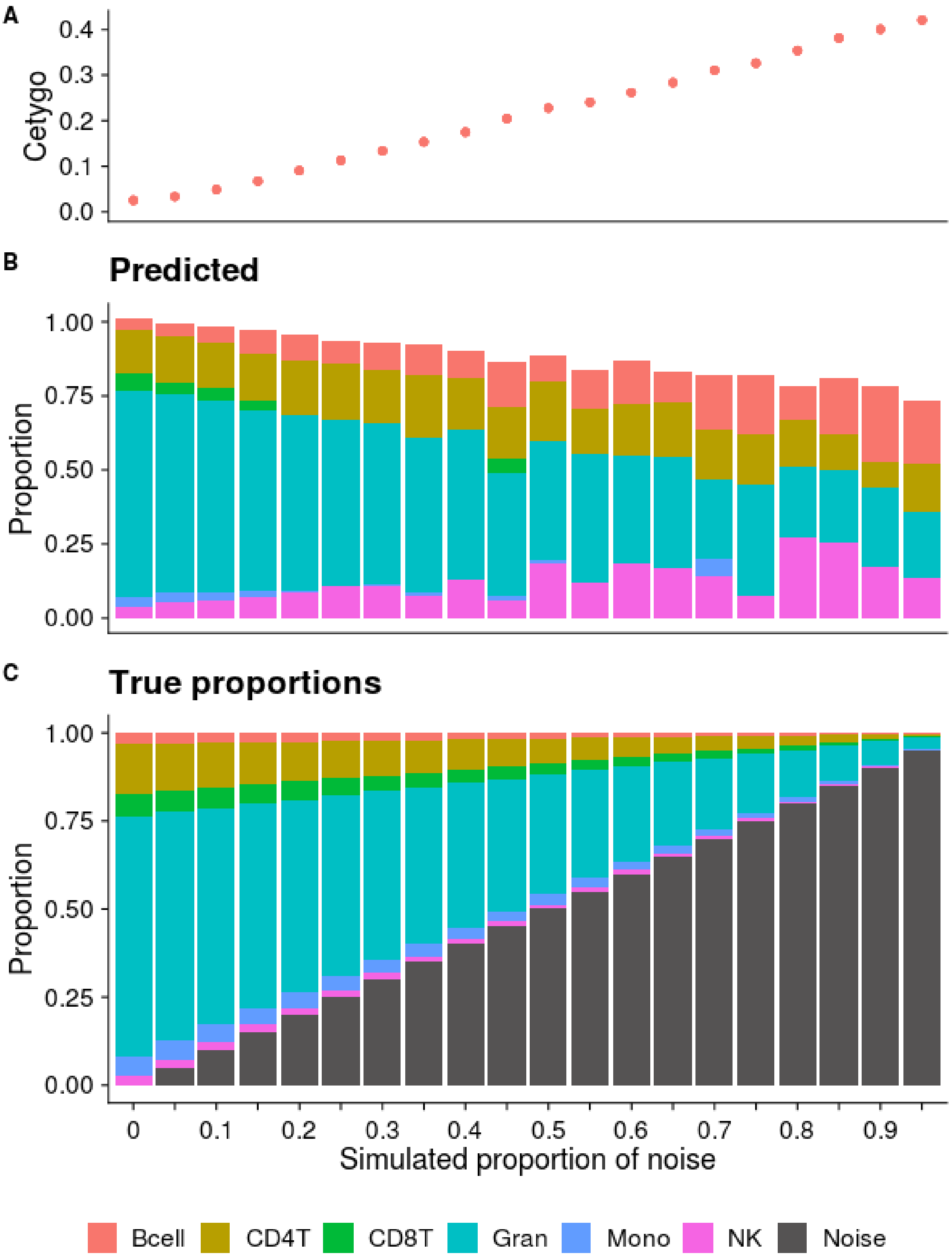


Figure 3.5: **A summary of Cetygo, estimated and actual cell type proportions of simulated blood samples with increasing noise.** Cetygo and prediction accuracy decrease as simulated noise increases. The figure shows the A) Cetygo, B) cell type proportions predicted by **Model 6CellBlood**, and C) true simulated cell type proportions across 20 simulated samples with the simulated proportion of noise ranging from 0 to 0.95. The cell type in plots B and C include B cells (salmon), CD4^{+} T cells (sand), CD8^{+} T cells (green), granulocytes (teal), monocytes (purple), and Noise (dark grey).

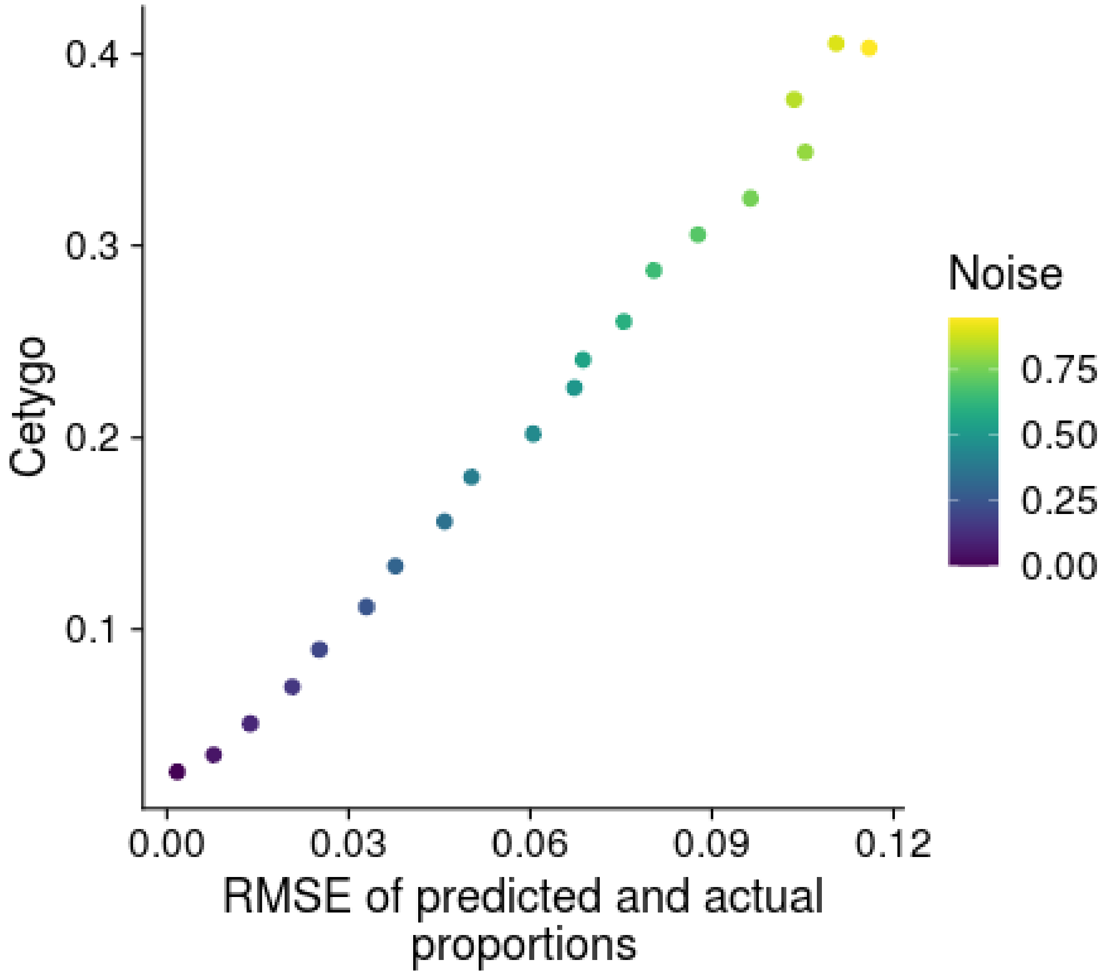


Figure 3.6: **A comparison between accuracy and Cetygo across samples simulated comprised of increasing noise** Cetygo and prediction accuracy correlate strongly across samples simulated with noise. Accuracy was quantified by calculating the RMSE between the predicted and actual cell type proportions across all cell types for each simulated sample. Here, points are coloured by the proportion of noise.

the sum of predicted proportions is not equivalent to the proportion of noise, and so, even though Houseman's algorithm allows for solutions where the sum of predicted proportions is less than one, the results will not necessarily be accurate. This highlights the importance of a deconvolution error metric in the context of deconvoluting samples that may not be as 'clean' as the testing data used to valid the model.

3.4.1.2 Cetygo increases as the quantity of missing data increases

Another scenario in which deconvolution accuracy may be reduced is when the sample to be deconvoluted does not contain DNA methylation values for all DNA methylation sites included in the model. This might be driven by DNA methylation quantification method or QC removing of poorly profiled DNA methylation sites. It would be expected that higher proportions of missingness would result in lower prediction accuracy and higher Cetygo. To assess how well Cetygo could index this potential reduced deconvolution accuracy, simulations were utilised; the 600 DNA methylation sites within **Model 6CellBlood** were simulated for one sample, again utilising equation 3.3, with cell type proportions ($\underline{P}$) set as the average cellular composition of whole blood data, according to Reinius et al., 2012 (the proportions of which can be found in **Table 3.5**). One set of cellular proportions was used so as not to confound results with any variability in prediction accuracy that might arise due to the cell types being predicted. Next, a proportion of said simulated DNA methylation sites were systematically repeatedly removed to generate samples with missing data, with the proportion of missingness ranging from 0-0.95 in steps of 0.05. For each proportion of missingness, the random removal of DNA methylation values was repeated 100 times so that the variability of Cetygo could be assessed, resulting in 2000 test cases. Finally, **Model 6CellBlood** was applied to estimate cellular composition, and Cetygo calculated.

As hypothesised, Cetygo increases as the proportion of missingness data increases (**Figure 3.7**). Interestingly, the variance of Cetygo seems to decrease with the proportions of missingness, suggesting that there may be some DNA methylation sites or combinations of sites that allow **Model 6CellBlood** to predict better than others. This makes intuitive sense as if, for example, the sites randomly removed all related to the prediction of a specific cell type then it would be assumed that the prediction of a sample including that cell type would be less accurate than a scenario in which the randomly removed samples had an even distribution across all cell type predictive sites. As missingness increases and more model sites are removed, the chances of the sites removed overlapping more between randomised removals, and the likelihood of more essential sites or groups of sites being removed in the same sample would be higher. Here, therefore, the variance may actually be capturing some bias in cell types more easily predicted given the remaining sites.

The relationship between prediction accuracy and Cetygo across missingness of model sites is not linear. **Figure 3.8** compares the accuracy, measured as RMSE across between the actual and predicted composition of the simulated samples, with the calculated Cetygo, colouring points by the proportion of missingness in the simulated data. Here, it can be seen that the variance of accuracy increases as Cetygo increases, which suggests that the specific sites removed matter for prediction accuracy, too. As such, Cetygo can be said to capture the likelihood of reliable prediction across this dataset, which would be beneficial in a scenario where true cell type proportions were not known. Given the comparatively low RMSE across predictions (maximum RMSE was 4 fold higher in **Figure 3.4.1.1**), overall accuracy does not seem largely affected by missingness, suggesting that a smaller number of DNA methylation sites might be sufficient for accurate deconvolution.

Most epidemiological DNA methylation studies are profiled using the array, and as such, deconvolution models are most commonly developed with and applied to array data. Despite this, differences between reference and input data can arise due to platform disparity, i.e. the platform in which the model was developed contains more sites than the input data array or not limiting model feature selection to sites also present in the input data, and QC, in which less reliable an poorly performing probes are removed from the data. Since missingness of predictive sites in input data has the potential to reduce accuracy, samples with $>5\%$ missingness are commonly removed, and deconvolution within QC R packages *minfi* and *wateRmelon* generate a new model for every dataset to which they are to be applied. To that end, the lack of direct correlation between Cetygo

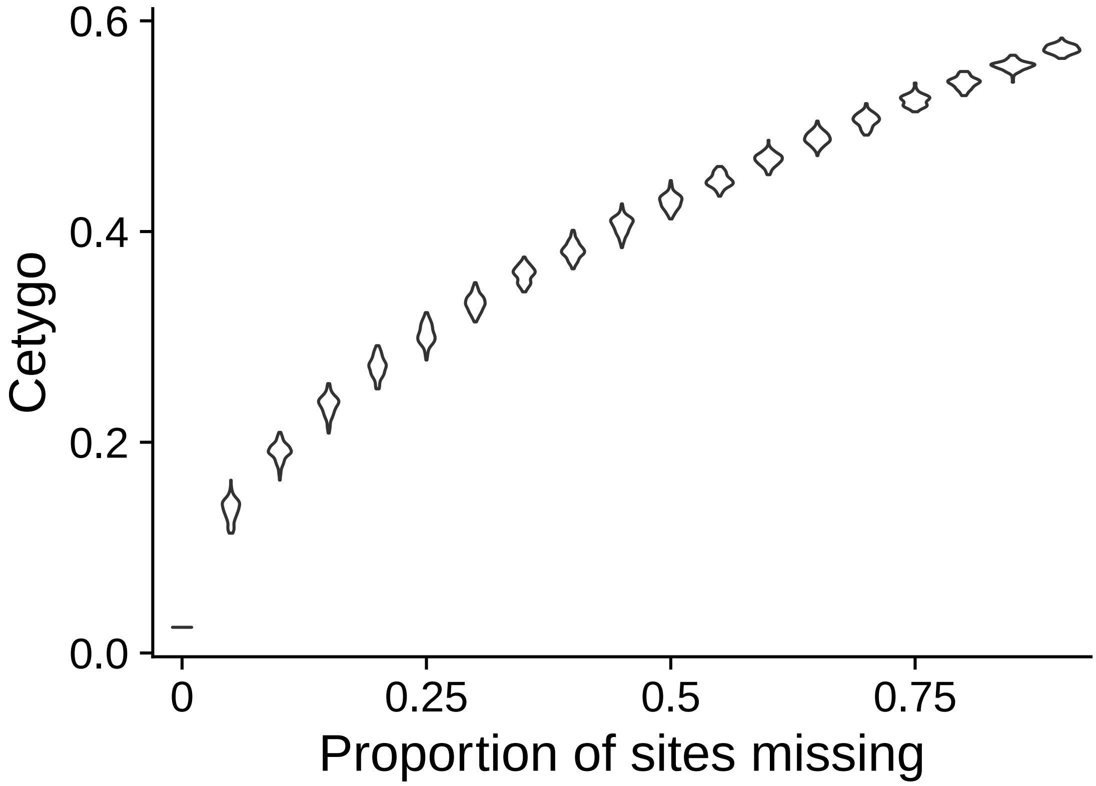


Figure 3.7: **a violin plot of Cetygo with increasing proportion of predictive model CpGs missing in simulated data.** Cetygo increases as the proportion of predictive model CpGs missing in the simulated samples increases. DNA methylation was simulated for a sample across the 600 CpGs in **Model 6CellBlood** with average bulk blood proportions (see **Table 3.5**). Increasing proportions of missingness were randomly assigned to sites at increasing levels (100 times per proportion missing).

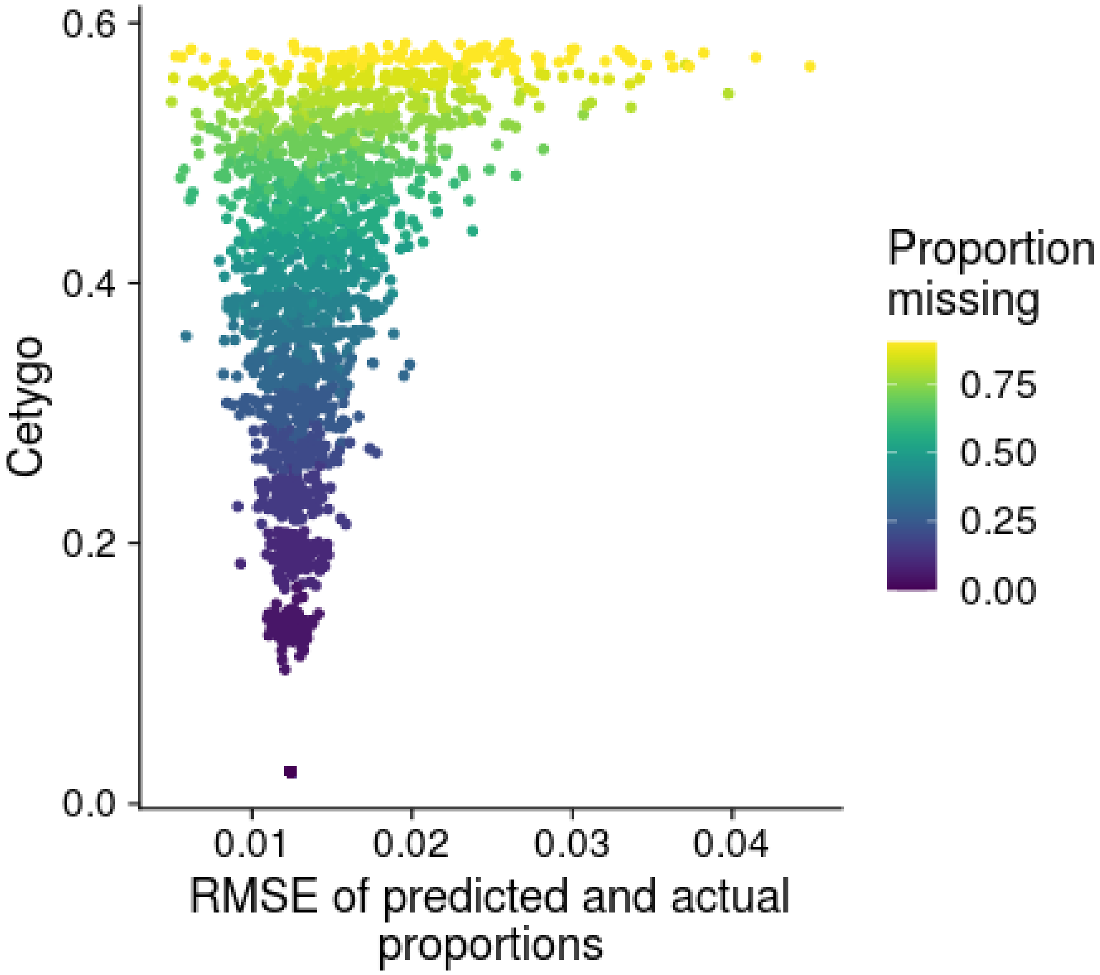


Figure 3.8: **A comparison of accuracy and Cetygo across samples simulated with an increasing proportion of predictive model CpGs missing.** The variance of prediction accuracy increases with the proportion of predictive model CpGs missing in the simulated data. Accuracy is measured here as the RMSE between the true simulated cell type proportions and the predicted values. As the variance of prediction accuracy increases, so does Cetygo.

and accuracy observed with increasing DNA methylation site missingness will not be detrimental to it's performance in conjunction with *minfi* or *wateRmelon*.

3.4.1.3 Cetygo can be used to assess completeness and cell type biases of models

Reference based cellular deconvolution requires a dataset comprised of the DNA methylation profiles of the main cell types that make up the tissue to be deconvoluted. If all cell types are not included in the reference dataset, it could be said that the deconvolution model will be insufficient for the tissue. This would occur when applying a model to a dataset in which the DNA methylation profiles of the input tissue is not completely represented by the cell types in the reference dataset, breaking the model assumption in equation 1.1. Therefore, it would be expected that when applying a model without all relevant cell types for the input tissue, the prediction accuracy would be decreased and ideally, Cetygo should reflect this.

To test this hypothesis, this scenario was simulated by generating blood deconvolution models that iteratively removed an increasing number of cell types from the training data, excluding 1-3 of the six purified blood cell types prior to generating the models. Unique models were generated using different subsets of cell types: 20 models were generated using a combination of 3 cell types, 15 models were generated using 4 cell types, and 6 models were generated using 5 cell types, resulting in 41 new "incomplete" models. **Model 6CellBlood** was utilised as a benchmark, being the most complete blood model available.

To test these models, the full range of possible whole blood DNA methylation profiles comprised of the six cell types were simulated, to permit a more equitable comparison not biased by specific, more abundant cell types. As such, the testing data simulated contained a systematically derived range of cell compositions; for each cell type, the possible proportions were selected from 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 or 1. A sample was simulated for every possible combination for which the sum of the proportions for 6 cell types was equal to 1, resulting in 2823 simulated profiles. The 42 deconvolution models were applied to each simulated sample such that Cetygo could

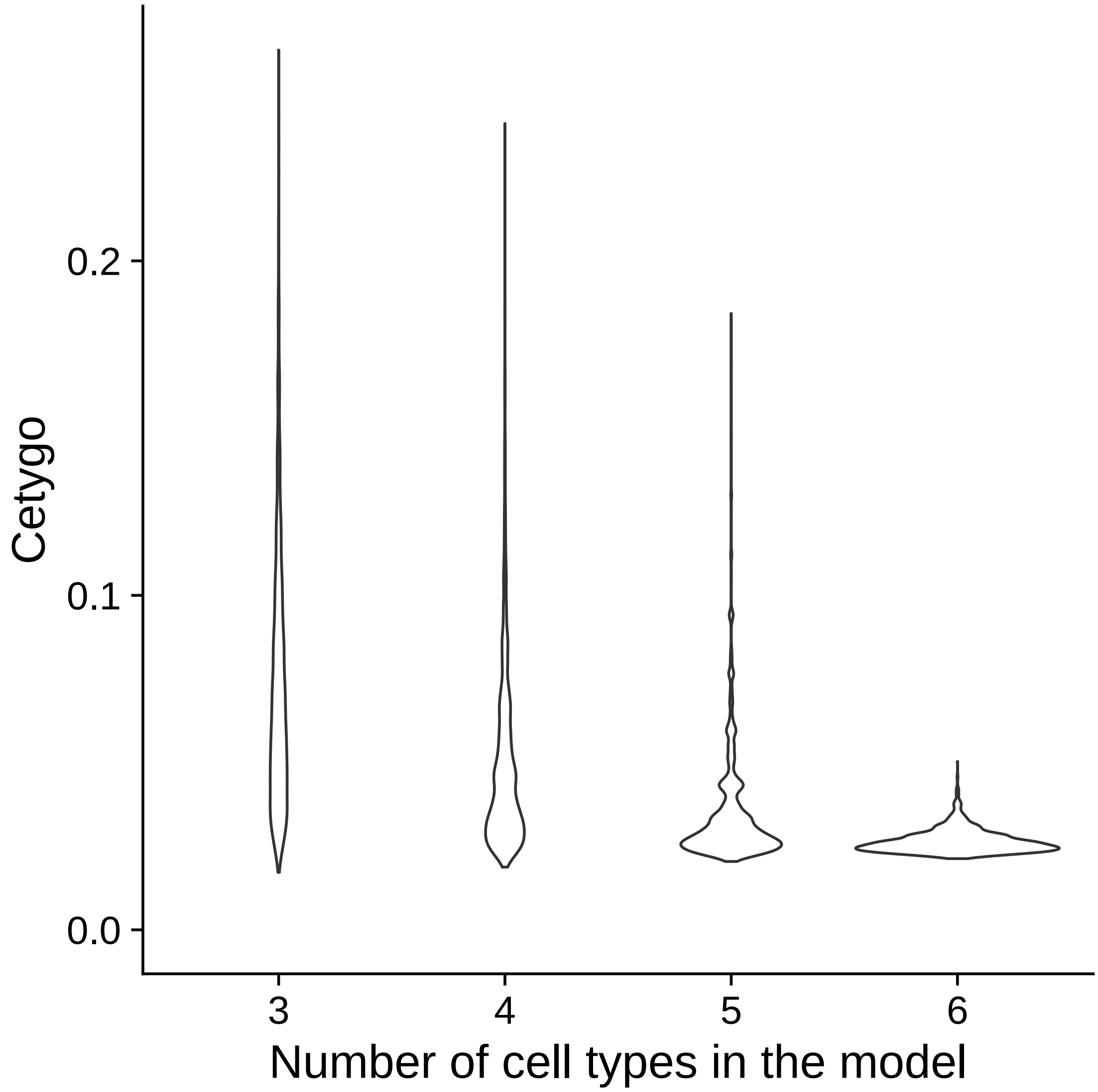


Figure 3.9: **A violin plot of Cetygo across blood deconvolution models with an increasing number of cell types.** Cetygo decreases as blood deconvolution models include more cell types. 42 unique models were generated, with 20, 15, 6 and 1 models including 3, 4, 5 or 6 blood cell types, respectively. All models were tested with the same 2823 samples, simulated from **Dataset Reinius** testing data.

be calculated.

The average and range of Cetygo decreased as more cell types were included in the models, with a higher number of predictions having a lower Cetygo with increased model completeness (3 cell type models: range - 0.0172:0.263, mean - 0.0731, standard deviation (SD) - 0.0378; 4 cell type models: range - 0.0187:0.241, mean - 0.497, SD - 0.0273; 5 cell type models: range - 0.0204:0.184, mean - 0.349, SD - 0.0161; 6 cell type model: range - 0.0213:0.05, mean - 0.0265, SD - 0.00353, **Figure 3.9**).

It might be the case that the DNA methylation sites included in the deconvolution model may be more informative for some cell types than others. As such, it was hypothesised that the variability in Cetygo was driven by bias towards specific cell types in the deconvolution models. This was first investigated across predictions from **Model 6CellPFC** by comparing the absolute difference between predicted and simulated proportions for each cell type across all 2823 simulated whole blood samples (**Figure 3.10**). Prediction accuracy was shown to vary by blood cell type, with the median absolute difference between actual and predicted cell proportions for B cells and CD4$^+$ T cells comparatively low (0.00358 and 0.005, respectively), and almost tenfold higher in all other cell types (CD8$^+$ T cells = 0.0278, granulocytes = 0.0398, monocytes = 0.0311, and NK cells = 0.0313).

Improved accuracy of some cell types compared to others may be driven by the uniqueness of said cell types within the model matrix. Model matrices have been generated to contain 100 distinct DNA methylation sites per cell type (see **Figure 3.36**), however, the HM450 array, with which the dataset DNA methylation was profiled, contains only 1.6% of all human CpGs. Given that the HM450 array was not generated with identifying cell type specific DNA methylation differences in mind, the sites selected for some cell types may more discriminative than for other cell types. To assess how effectively the model matrix of **Model 6CellBlood** distinguished each cell type a principal component analysis (PCA) was conducted (**Figure 3.11**).

Together, the first four principal components (PCs) explain 96.18% of the variance across the **Model 6CellBlood** model matrix. PCs 1 and 2 divide the cell types into three general groups, B cells, monocytes and granulocytes, and CD4$^+$ T cells, CD8$^+$ T

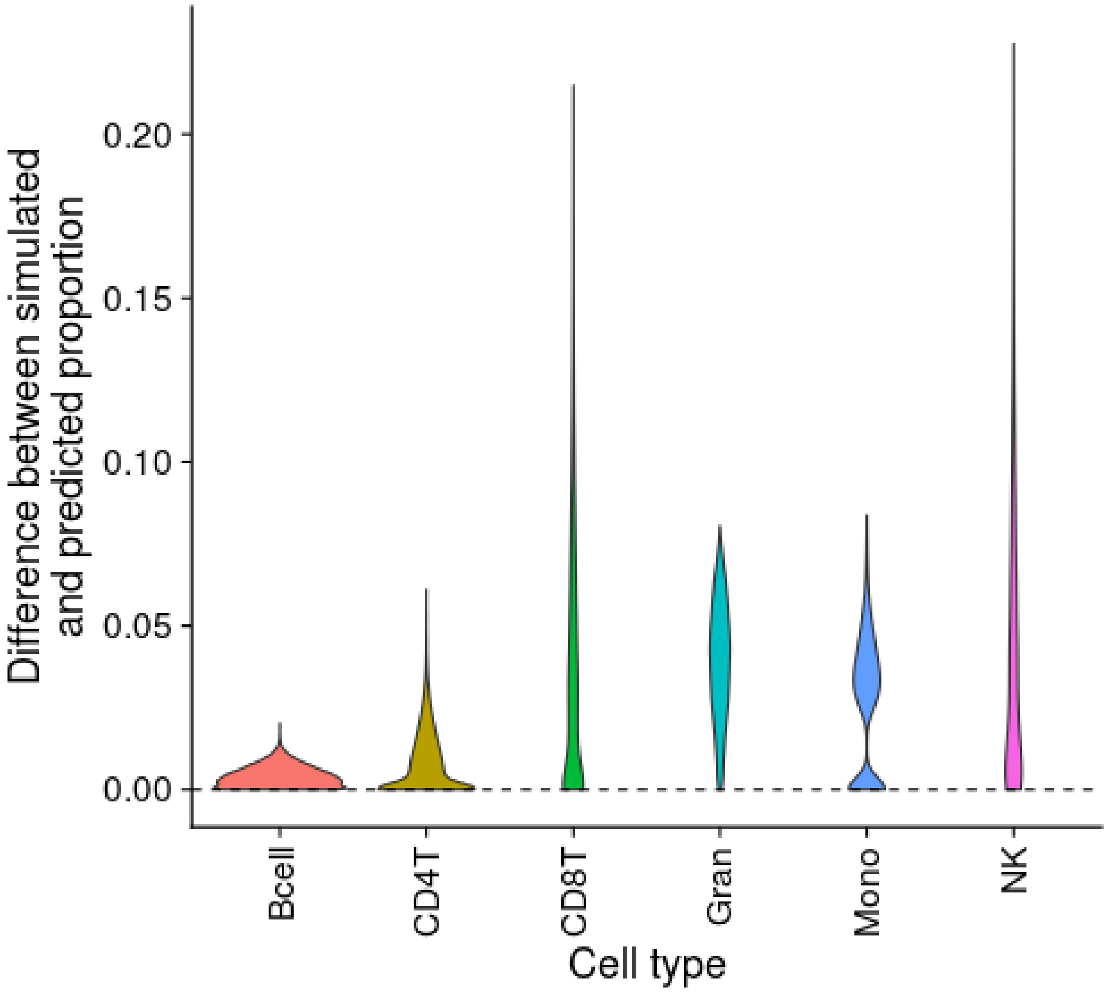


Figure 3.10: **A violin plot of the differences between predicted and actual cell type proportions predicted by Model 6CellBlood plotted by cell type.** The median and range of differences between predicted and actual cell type proportions predicted by **Model 6CellBlood** is cell type specific. **Model 6CellBlood** is applied across 2823 samples, simulated from **Dataset Reinius** testing data, with each cell type proportion set between 0-1 at step size of 0.1 so that the total proportion sums to 1.

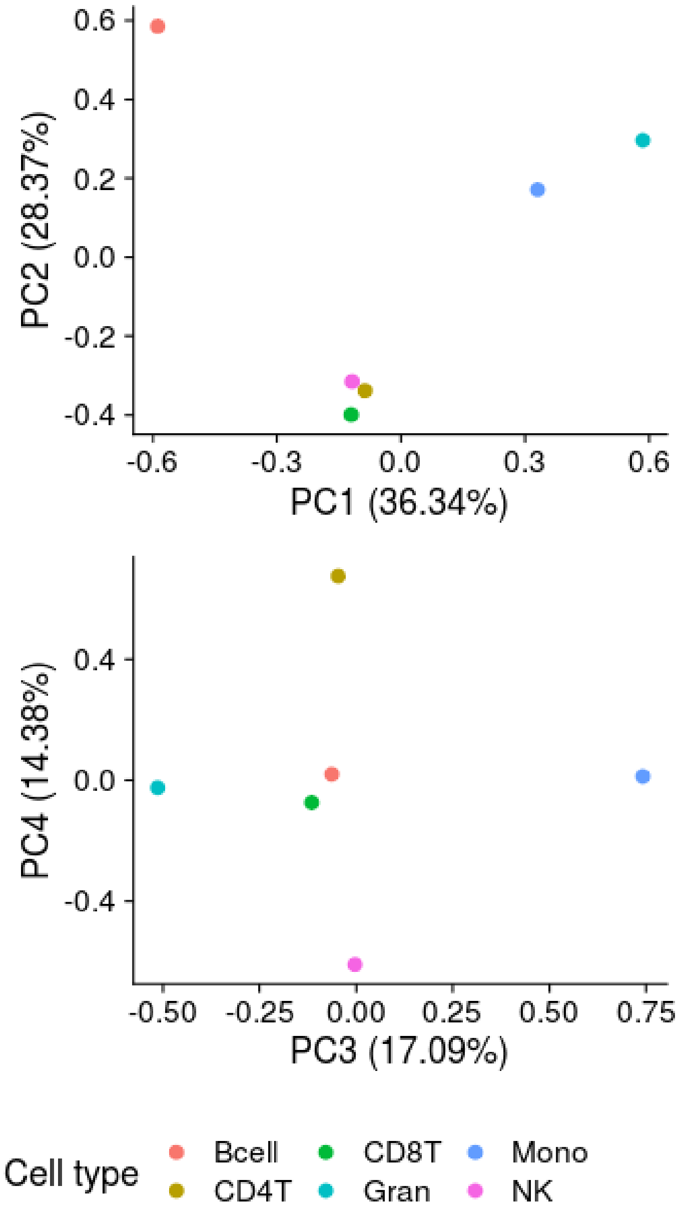


Figure 3.11: **Scatter plots of the first four principal components of the cell type specific sites in Model 6CellBlood.** The model matrix for **Model 6CellBlood** (M_{CT} in Section 3.1.1) contains a column of DNA methylation values for each cell type which were selected to distinguish cell types.

cells and NK cells. PC 3 divides monocytes and granulocytes from all other cell types, and PC 4 divides CD4$^+$ T cells and NK cells. According to PCs 1 and 2 B cells differ most from all other cell types, perhaps explaining their high prediction accuracy across simulated data in **Figure 3.10**. It is unclear, however, why CD4$^+$ T cells were predicted more accurately than the remaining cell types.

The large range in Cetygo observed in **Figure 3.9** across incomplete models (i.e. those containing less than 6 blood cell types) was most likely due to the cell type composition of simulated samples, as the omission of a cell type from a model would lead to bias in predictions. Specifically, predictions for simulated samples that only contained cell types included in the model should retain similar levels of accuracy (i.e where the proportion of the excluded cell types was set to 0), whereas the prediction should be less accurate when applied to simulated samples made up from cell types not represented in the model.

To verify this, the models containing five of the six blood cell types were investigated further. For each five cell type model, a new dataset was simulated containing increasing proportions of the missing cell type, from 0.1 to 1, in steps of 0.1. The remaining cellular composition was constructed from the other five cell types with proportions reflecting the average empirical blood proportions in whole blood (**Table 3.5**) so that the total cell type proportions summed to one. A total of 60 whole blood samples were simulated, with 10 samples per cell type, and each of the six 5 cell type models applied to their respective 10 samples.

As anticipated, Cetygo increased monotonically with the abundance of the missing cell type in the model (**Figure 3.12**) albeit with cell type specific gradients. B cells being omitted from the model had the most extreme effect in Cetygo, while CD8$^+$ T cells had the most subtle effect.

The predicted proportions of the simulated samples can be found in **Figures 3.13 - 3.18** Here, unlike with simulated samples containing noise (**Figure 3.4.1.1**), in which the sum of the predicted proportions decreased, the sum of predicted proportions remained close to one, even when the simulated sample is comprised almost entirely of the missing cell type. Instead, the missing cell type was predicted to be a different cell type; CD4$^+$

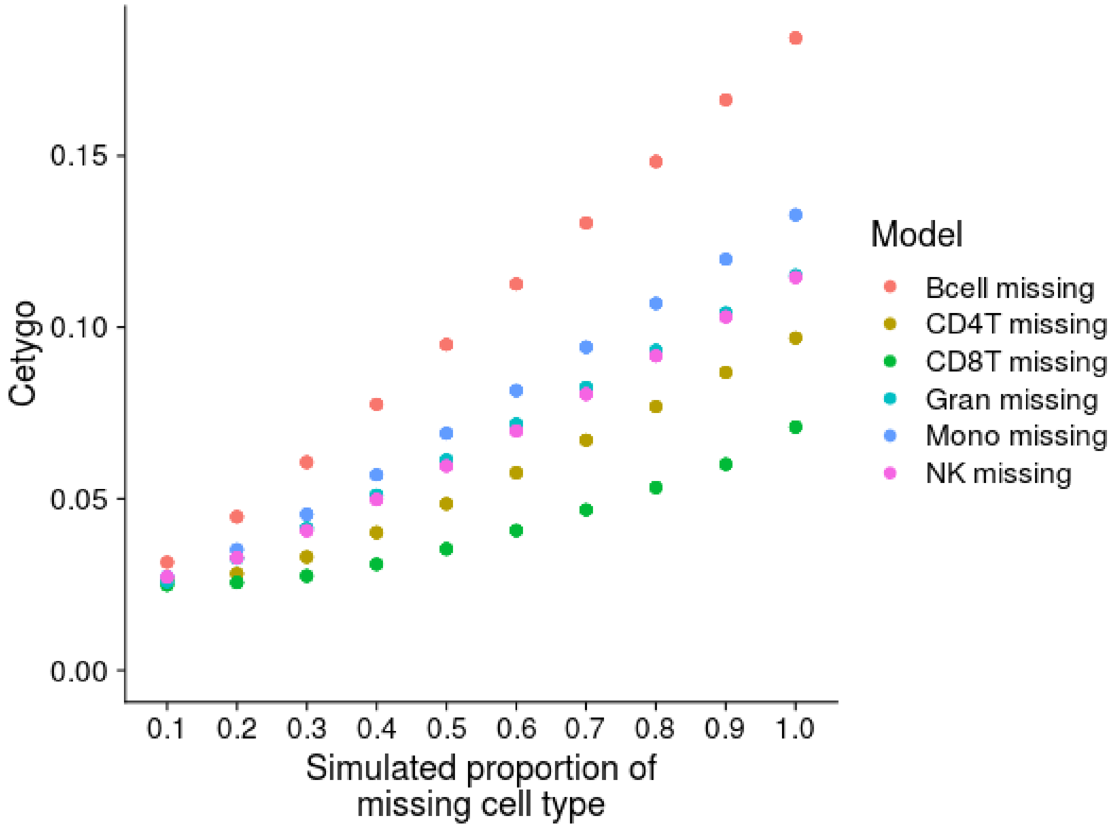


Figure 3.12: **A comparison of Cetygo and the proportion of unrepresented cells.** Cetygo increases as the proportion of unrepresented cells increases. Whole blood samples were simulated to reflect the average whole blood profile (as summarised by Reinius et al., 2012) with the exception of one each cell type that was set to have increasing proportions (from 0.1 to 1 in steps of 0.1), and then tested with a deconvolution model that excluded only that cell type. Cetygo points are coloured by the model used which are labelled by the cell type missing from the reference panel, with B cells missing in salmon, CD4$^+$ T cells missing in sand, CD8$^+$ T cells missing in green, granulocytes missing in teal, and monocytes missing in purple.

T cells and NK cells were predicted as CD8$^+$ T cells, and *vice versa*, and granulocytes and monocytes were predicted interchangeably, with monocytes also containing a small proportion as NK and B cells. B cells themselves were predicted as a equal combination of monocytes, NK and CD4$^+$ T cells, although B cells and monocytes were not similar according to **Figure 3.11**, which might explain the elevated Cetygo.

Findings demonstrate the utility of Cetygo to assess model completeness and applicability to samples containing cell types not included within the reference data. Overall, this Section supports the application of Cetygo as a metric for performing comparisons between deconvolution models, demonstrating its ability to distinguish those which are less complete and therefore less accurate. This is pertinent as, due to challenges in cell type purification, not all cell types within a tissue are always available for model generation, i.e. reference data for them does not exist.

3.4.2 Simulation validation using Model 2CellPFC

In the above Sections, the utility of Cetygo was explored using a blood based deconvolution model. While the Cetygo framework is expected to be applicable across tissues, given the cell type biases uncovered in deconvolution models, variability will likely exist. Here, analyses are repeated across a deconvolution model trained in different tissue, namely, PFC, to assess whether results are independent of the model and tissue applied.

The model used, **Model 2CellPFC**, was developed using the training data of **Dataset Guintivano**. The model reference data contains two distinct cell type populations, NeuN+ and NeuN-, sorted using the FANS protocol and NeuN staining, which identifies neuronal populations (See Section 4.1.1.1 for an overview of the purification methodology). The resulting model contains 100 predictive DNA methylation sites. Given that **Model 2CellPFC** contains only two distinct populations, in contrast to the 6 used in **Model 6CellBlood**, and a relatively low number of DNA methylation sites, the simulations investigating Cetygo and the proportion of site missingness (as done in Section 3.4.1.2) or model completeness (as done in Section 3.4.1.3) could not be replicated in using **Model 2CellPFC**.

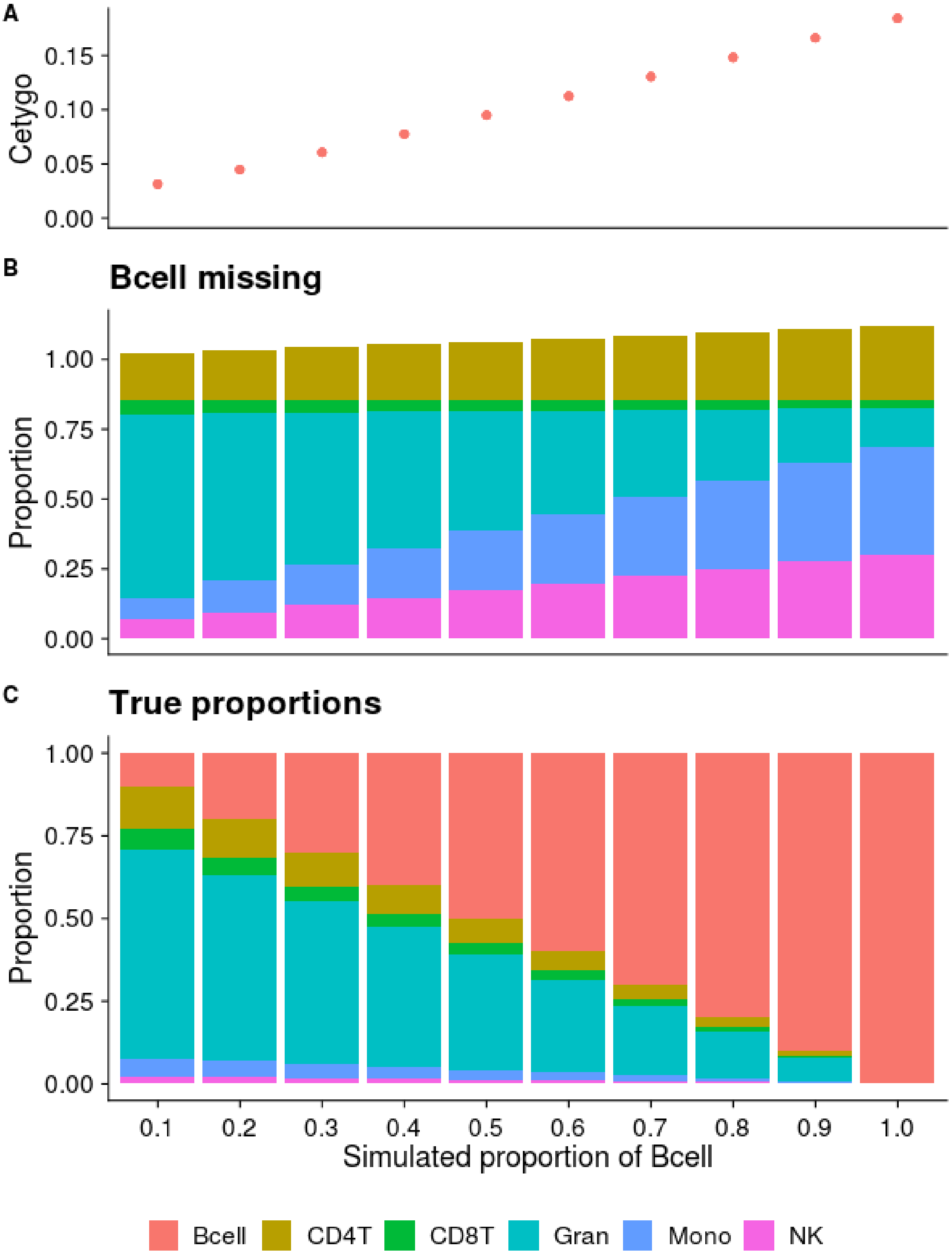


Figure 3.13: **A summary of Cetygo, estimated and actual cell type proportions of whole blood samples simulated to be comprised of increasing proportion of B cells, unrepresented in the deconvolution model used.** A) The Cetygo across simulated samples. B) Stacked bar plot of the predicted proportion of cell types by the model excluding B cells. C) Stacked bar plot of the true proportions of the simulated samples.

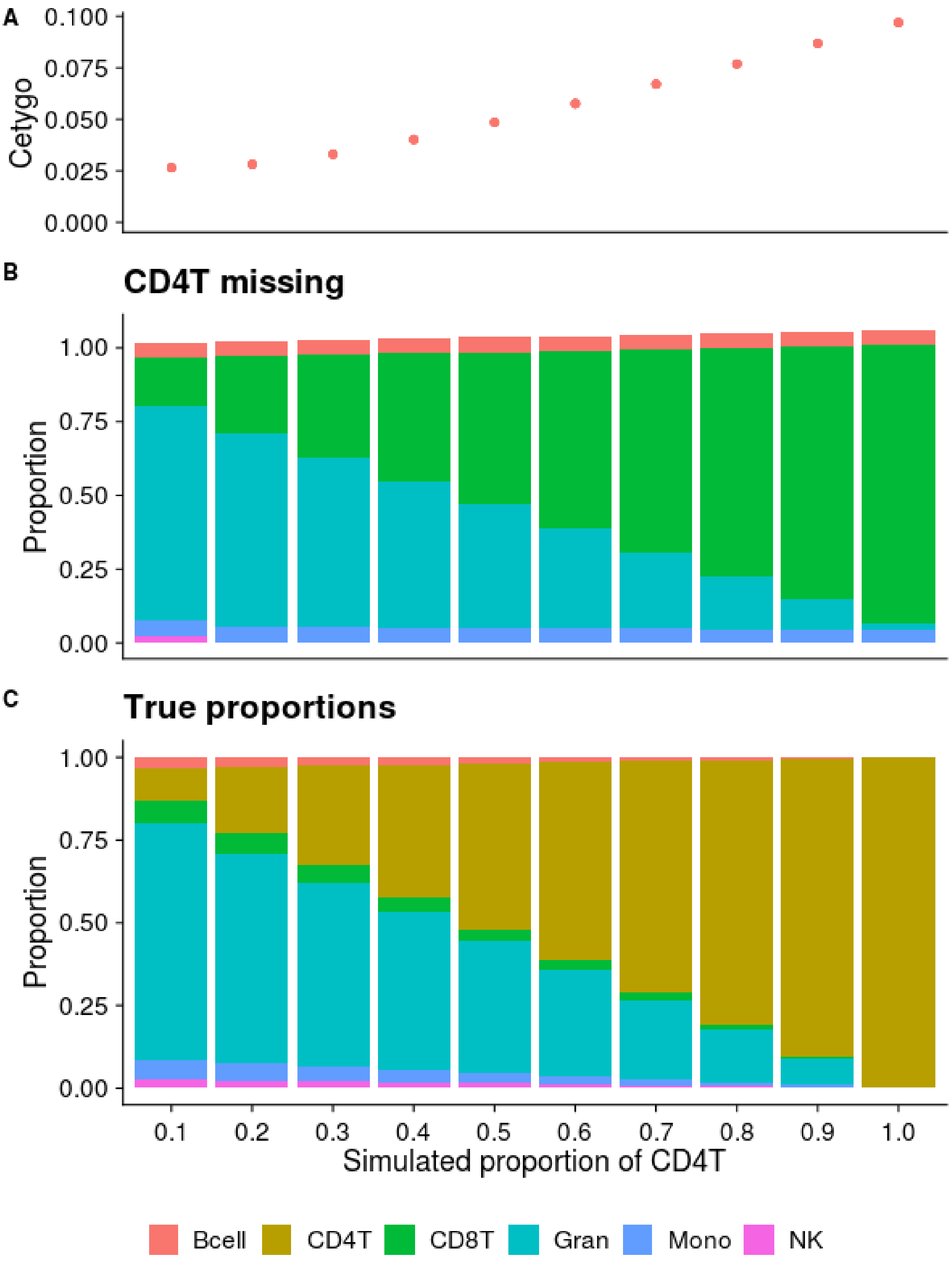


Figure 3.14: **A summary of Cetygo, estimated and actual cell type proportions of samples simulated to be comprised of increasing proportion of CD4$^+$ T cells, unrepresented in the deconvolution model used.** A) The Cetygo across simulated samples. B) Stacked bar plot of the predicted proportion of cell types by the model excluding CD4$^+$ T cells. C) Stacked bar plot of the true proportions of the simulated samples.

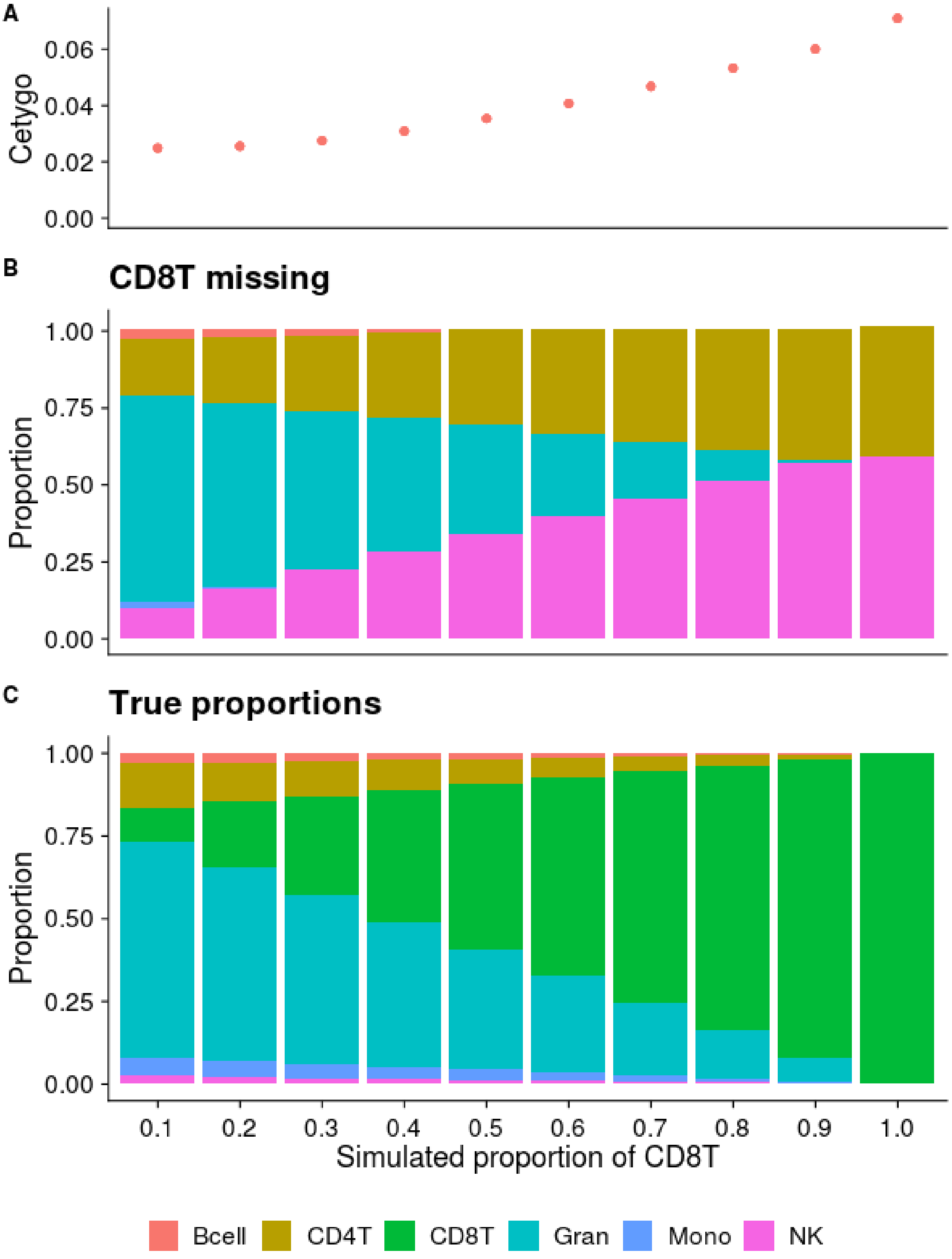


Figure 3.15: **A summary of Cetygo, estimated and actual cell type proportions of samples simulated to be comprised of increasing proportion of CD8$^+$ T cells, unrepresented in the deconvolution model used.** A) The Cetygo across simulated samples. B) Stacked bar plot of the predicted proportion of cell types by the model excluding CD8$^+$ T cells. C) Stacked bar plot of the true proportions of the simulated samples.

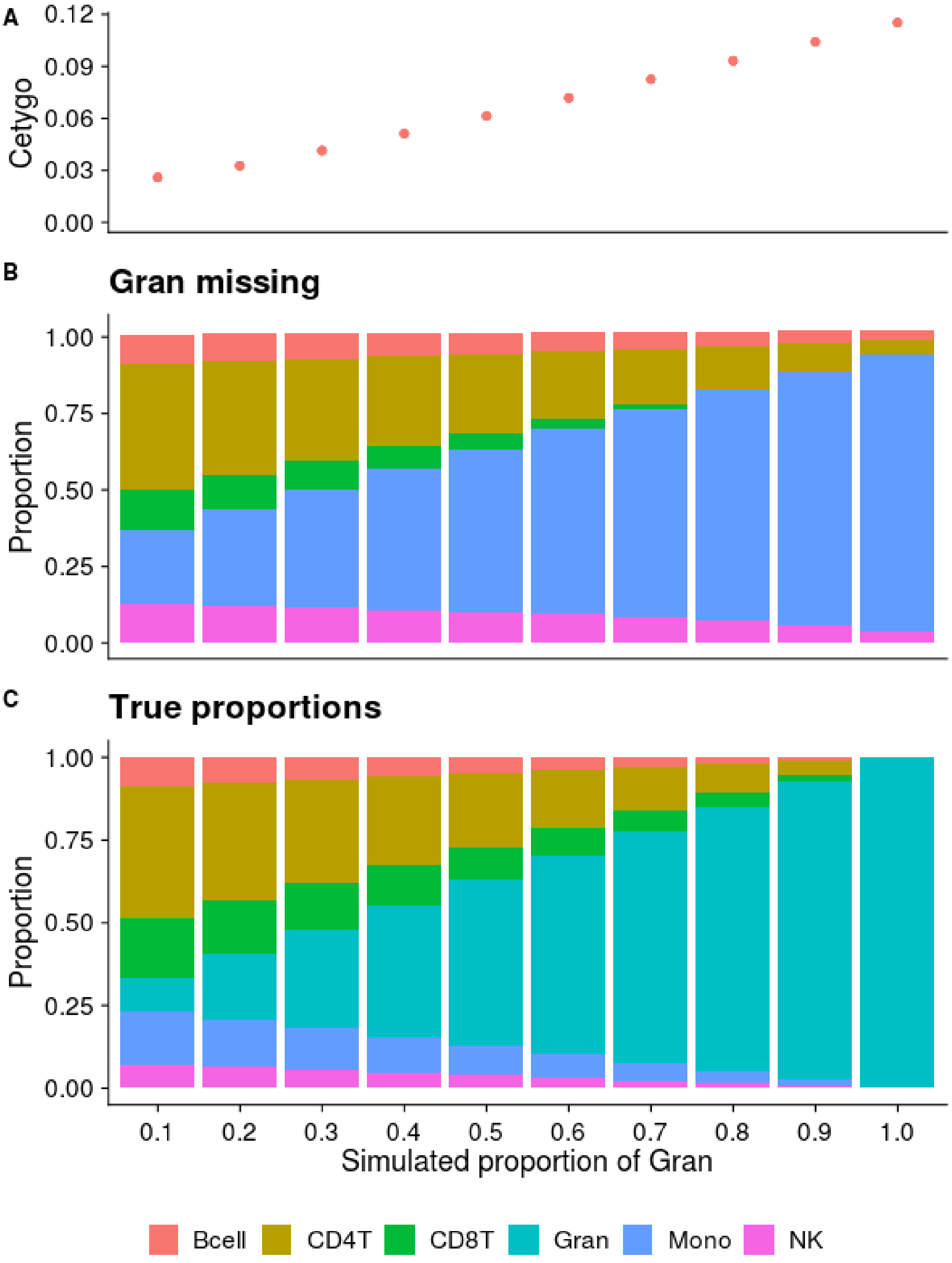


Figure 3.16: **A summary of Cetygo, estimated and actual cell type proportions of samples simulated to be comprised of increasing proportion of granulocytes, unrepresented in the deconvolution model used.** A) The Cetygo across simulated samples. B) Stacked bar plot of the predicted proportion of cell types by the model excluding granulocytes. C) Stacked bar plot of the true proportions of the simulated samples.

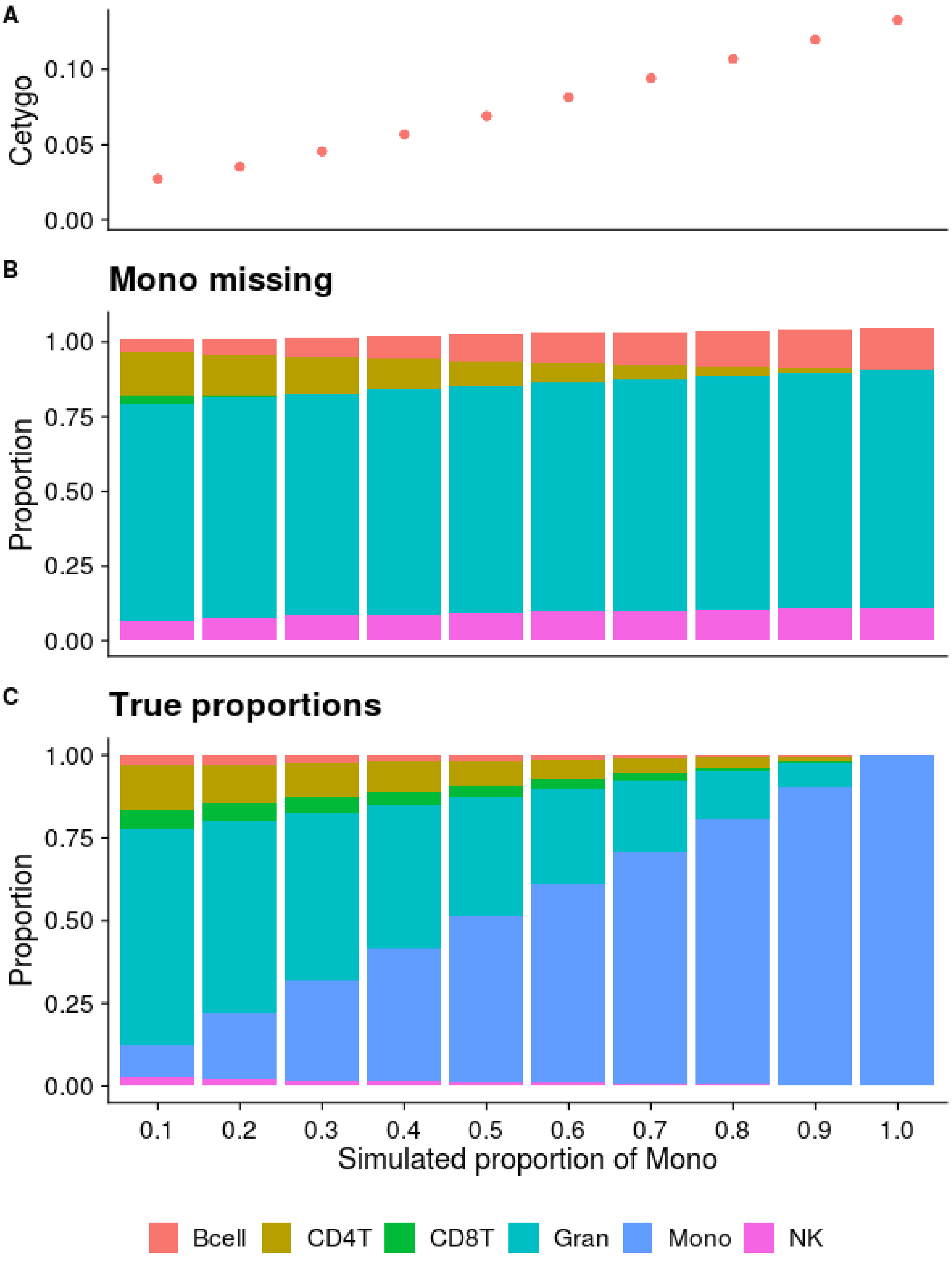


Figure 3.17: **A summary of Cetygo, estimated and actual cell type proportions of samples simulated to be comprised of increasing proportion of monocytes, unrepresented in the deconvolution model used.** A) The Cetygo across simulated samples. B) Stacked bar plot of the predicted proportion of cell types by the model excluding monocytes. C) Stacked bar plot of the true proportions of the simulated samples.

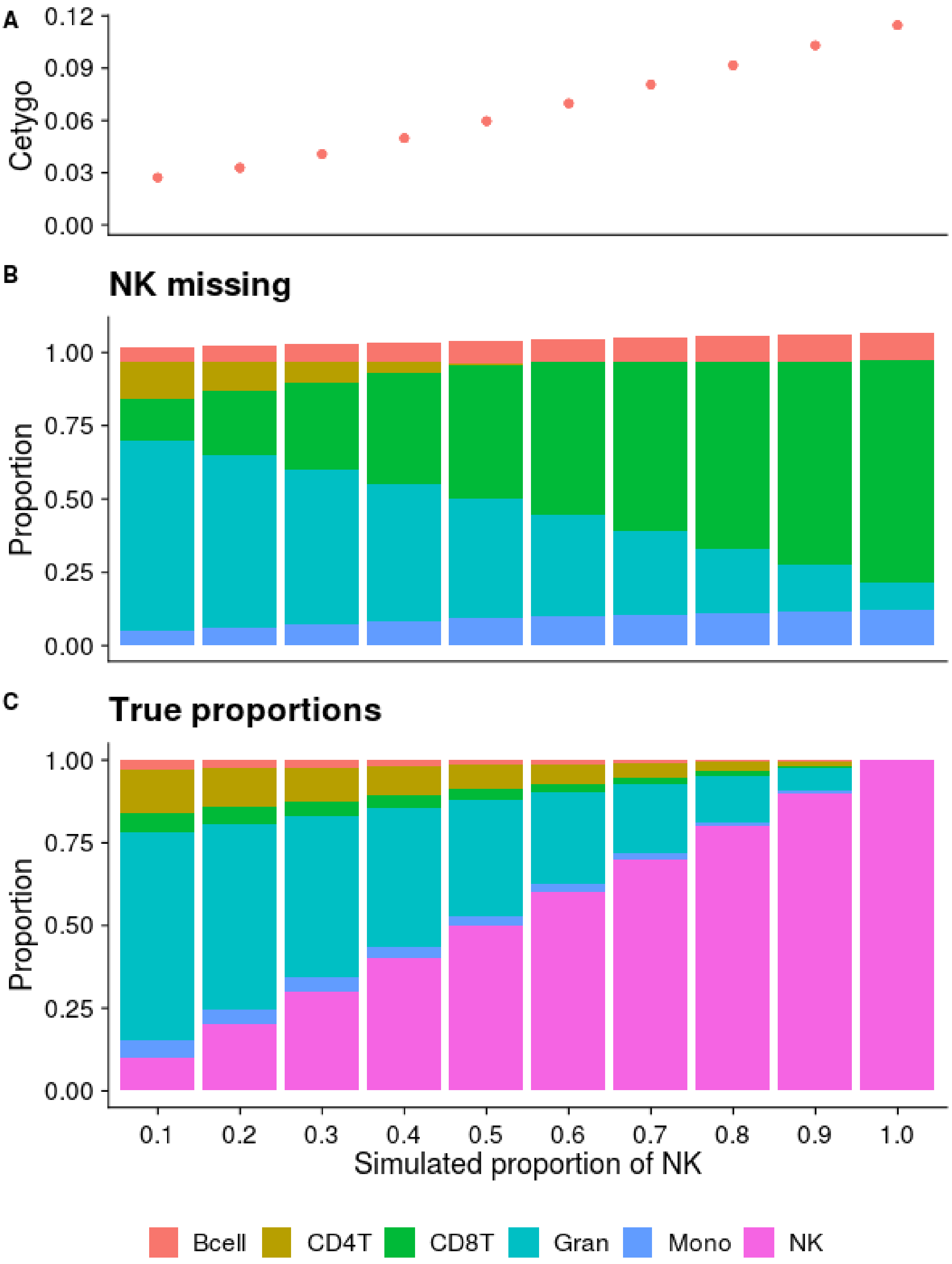


Figure 3.18: **A summary of Cetygo, estimated and actual cell type proportions of samples simulated to be comprised of increasing proportion of NK cells, unrepresented in the deconvolution model used.** A) The Cetygo across simulated samples. B) Stacked bar plot of the predicted proportion of cell types by the model excluding NK cells. C) Stacked bar plot of the true proportions of the simulated samples.

3.4.2.1 Cetygo also can distinguish 'noise' from true cellular heterogeneity using a PFC model

This section mirrors Section 3.4.1.1, in which 10 samples were simulated with increasing proportions of noise, with the remaining proportion assigned as 50:50 NeuN+:NeuN- (P_i in equation 3.3). To simulate the data, the testing data of **Dataset Guintivano** was used ($\tilde{M}_{CT.TEST_i}$ in equation 3.3).

As in blood, Cetygo increased linearly with increased proportion of noise (**Figure 3.19**). Accuracy, that is, the difference between predicted and actual proportions of NeuN+ and NeuN-, did not correlate with Cetygo in PFC prediction (Cor = 0.142), however, this may be due to the reduced number of cell types to be predicted in this model compared to blood. Again, it suggests that Cetygo is better capturing prediction reliability than other measures of prediction accuracy. It also demonstrates that the performance of Cetygo may be dependent on the model across which it is applied.

3.5 Testing the capability of Cetygo to distinguish incorrect deconvolution of a tissue type not matching the reference data

So far it has been shown that Cetygo indexes the quality of deconvolution predictions, capturing when data is noisy and when models produce inaccurate estimates due to incomplete reference panels. All results so far have utilised simulated data, which has the benefit of knowing the cellular composition, but may not be entirely representative of bulk tissue data. To that end, this Section investigates the utility of Cetygo to quantify deconvolution accuracy across a large set of DNA methylation cohorts collated into **Dataset GEO** (**Table** 7.4) containing 37 different tissue or cell types. All tissue annotations were obtained through a regular expression search for the term "Tissue" or "Source".

Model 6CellBlood and **Model 2CellPFC** are applied across across all tissue types with the expectation that Cetygo would be elevated for the model applied to

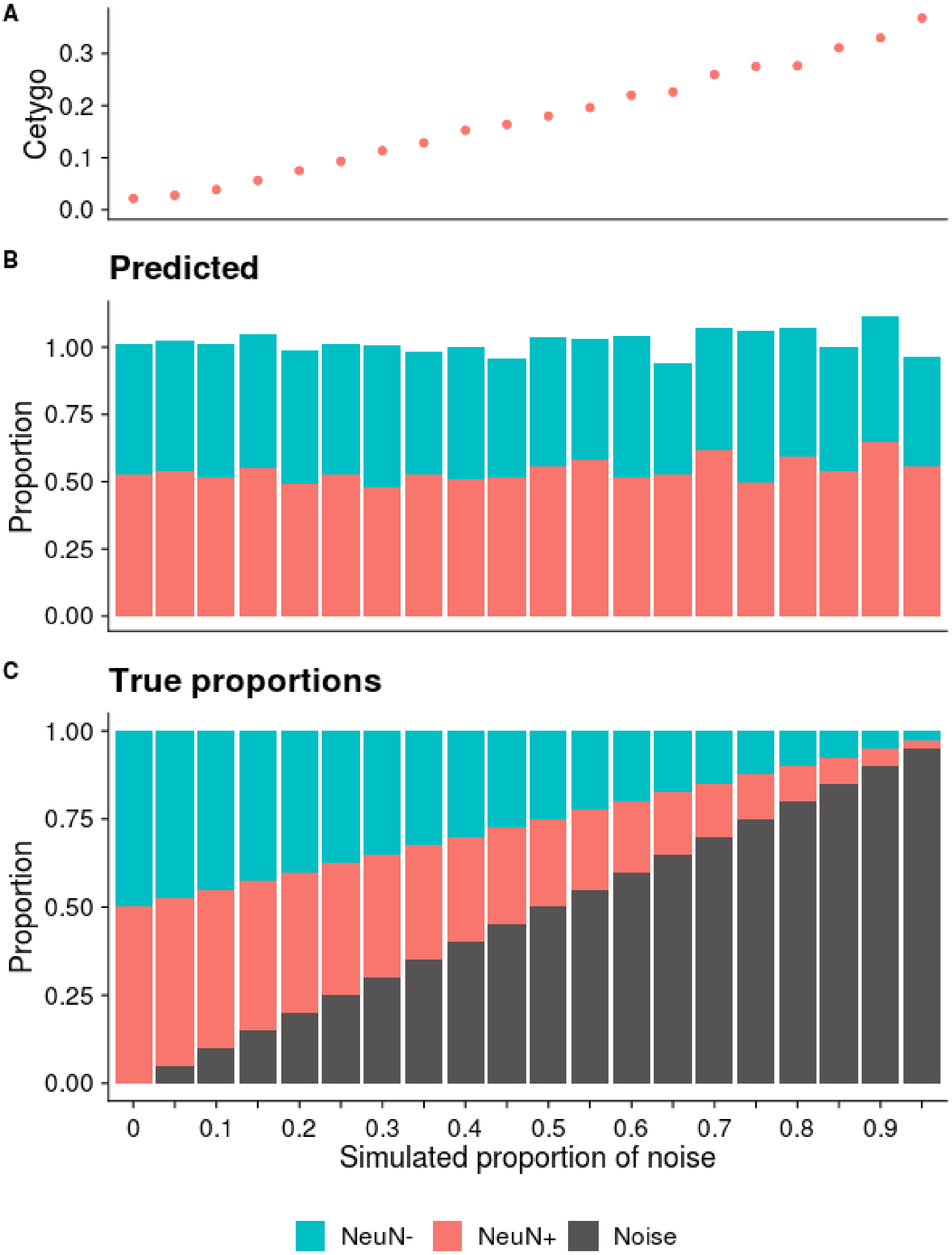


Figure 3.19: **A summary of Cetygo, estimated and actual cell type proportions of simulated PFC samples with increasing noise.** The figure shows the A) Cetygo, B) cell type proportions predicted by **Model 2CellPFC**, and C) true simulated cell type proportions across 20 simulated samples with the simulated proportion of noise ranging from 0 to 0.95. The cell types in the plot are NeuN+ (pink), NeuN- (teal) and noise (dark grey).

samples where the labelled sample type was not represented in the cellular reference data. In other words, Cetygo obtained from **Model 6CellBlood** should be higher for non-blood samples compared to blood samples and when obtained from **Model 2CellPFC** it should be higher for non-brain samples compared to brain samples. While, in reality, this is an extreme test scenario, there are situations when the origin of a sample might be questioned, for example sample swaps or sample mislabelling.

3.5.1 Cetygo is significantly higher in non-blood tissue compared to blood tissue using Model 6CellBlood

Predicted blood composition and Cetygo were calculated across **Dataset GEO** using **Model 6CellBlood**. Cetygo across blood and purified blood cell type samples ($n = 10607$) was significantly lower that than across all other tissues ($n = 6055$, $p < 5e\text{-}324$) (**Figure 3.20**). Blood Cetygo had a relatively small interquartile range (0.0415-0.0812) with a long tail of higher Cetygo. Based on the previous analysis, the subset of blood samples with higher Cetygo are likely inaccurately deconvoluted. The long tail may come from a number of sources, such as technical, including incorrect annotation, or biological differences, such as disease, in those samples and is not necessarily attributed to tissue differences. This further demonstrates Cetygo's potential utility as a QC metric for assessing sample quality in epidemiological studies prior to downstream analysis.

The median Cetygo was greater than 0.15 for all non-blood tissue types except induced pluripotent stem cells (iPSCs), which had median Cetygo 0.0723, although it is not known what tissue type the iPSCs were derived from. This would suggest the possibility of establishing a Cetygo threshold for 'good' deconvolution (explored in Section 3.6).

17.7% ($1614/9118$) of non-blood samples had Cetygo <0.1, including samples from adipose, buccal, and embryonic stem cells. This could have a number of causes, including possible sample contamination, even minimal contamination with blood on sample collection, could provide sufficient "blood-like" signal to mean the application of a blood deconvolution model has some meaning. It is also possible that the GEO tissue annotation retrieval was not perfect or specific enough (e.g. iPSCs will be induced to resemble another cell type, and that level of detail is not available). This is because the dataset was retrieved from GEO using an automated pipeline rather than being manually curated, which would not have been feasible given the large sample size. A possible caveat of applying Cetygo in this context is that it has been assumed that **Model 6CellBlood** contains enough predictive DNA methylation sites to differentiate between blood and other tissue types. However, the model was not created to assess tissue types outside of blood and only 600 DNA methylation sites were selected per cell type, meaning that the sites which might more effectively distinguish between blood and non-blood tissue may not be present in the model.

This analysis also suggests that while a low Cetygo is consistent with a sample being blood, it does not rule out that it is not another tissue.

3.5.2 Cetygo is significantly higher in non-brain tissue compared to brain tissue using Model 2CellPFC

Next, **Model 2CellPFC** was applied across **Dataset GEO**. The difference in Cetygo between brain (n = 1484) and non-brain samples (n = 17647) was highly significant (p = 8.39e-213).

The **Model 2CellPFC** Cetygo had more variation across brain samples (interquartile range = 0.0675:0.24) than in blood samples as predicted by **Model 6CellBlood** (Section 3.5.1, **Figure 3.21**), which may in part be because brain reference data is less likely to be applicable across all datasets, for example, due to the brain region used, the information for which was not available (as explored in Section 4.9.1). Higher Cetygo may also be driven by potentially lower data quality e.g. due to sample storage and processing which is less challenging in blood. The number of predictive DNA methylation sites in **Model 2CellPFC** is only 100, and as in blood, this may not be enough to distinguish between brain and non-brain cell types given that non-brain cell types are not included in the training data. Only rectal samples also had a median Cetygo below 0.1, although other tissues, such as chorion, blood and intestines, contained outlier samples with low Cetygo. It is important to note that the samples have not been through QC and that only 2.11% of non-brain samples had Cetygo <0.1 ($373/17647$).

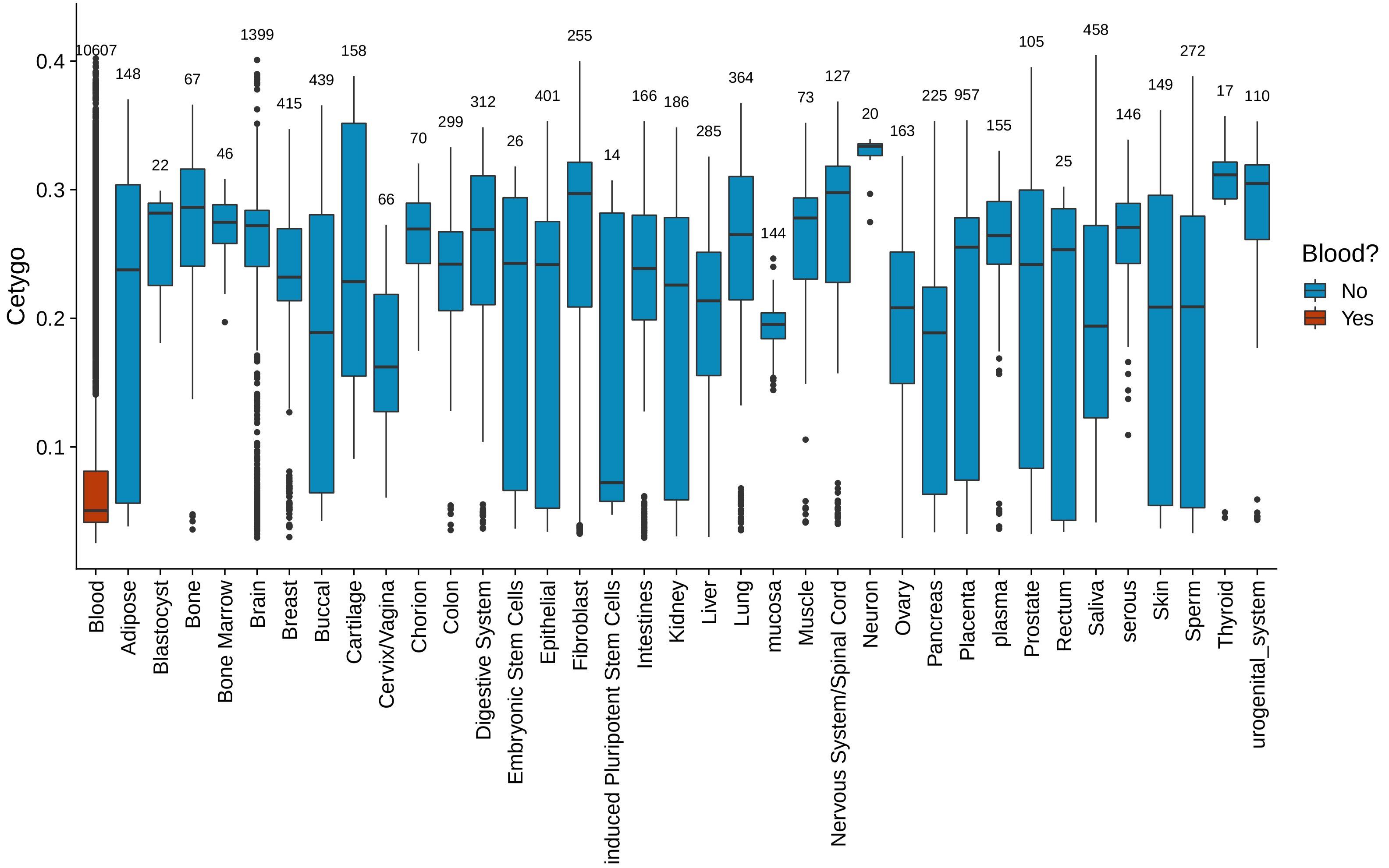


Figure 3.20: **A box plot of Cetygo from predictions by Model 6CellBlood varied tissue type samples in Dataset GEO.** The sample size of each tissue is shown above its respective box plot, with the blood tissue box plot highlighted in red.

Overall, Cetygo demonstrates differences between samples where an inappropriate deconvolution model has been used.

3.6 Assigning a soft threshold for Cetygo

As demonstrated in the above Sections, Cetygo has utility in assessing the quality of deconvolution outcomes. So far, Cetygo's usage has been as a comparative metric, e.g. comparing between samples (Sections 3.4.1.1 and 3.4.1.2), deconvolution models (Section 3.4.1.3), or tissues (Section 3.5). Theoretically, Cetygo could also be utilised to test that there was no deconvolution bias associated with a phenotype of interest. To ensure that Cetygo is used in a consistent way, this Section aims to recommend a soft threshold for Cetygo below which deconvolution predictions are considered 'good'.

To determine a suitable value for the soft threshold, the Cetygo of predictions in the blood and brain samples in **Dataset GEO** by **Model 6CellBlood** and **Model 2CellPFC** (applying both models across both tissue types) were utilised, the box plot distributions found in **Figures 3.20** and **3.21**, showed that predictions across both tissue types had a large range in Cetygo across both models (brain predicted in **Model 2CellPFC**: 0.0224-0.396, blood predicted in **Model 2CellPFC**: 0.0438-0.385, brain predicted in **Model 6CellBlood**: 0.03-0.401, blood predicted in **Model 6CellBlood**: 0.0253-0.389). The relationship between each model Cetygo across both tissues is used here as an indication of a suitable threshold at which Cetygo defines samples to be well deconvoluted.

Figure 3.22 shows the relationship between Cetygo ascertained from **Model 6CellBlood** and **Model 2CellPFC**, with each point coloured by the sample tissue annotation (pink for blood, blue for brain). It can be noted that the bottom left corner of the plot is sparse, implying that no sample performs well in both models. Three clusters can be observed in the plot, one with low Cetygo across **Model 6CellBlood**, another with low Cetygo across **Model 2CellPFC**, and at higher Cetygo, a long smudge of points where the correlation of Cetygo between models is higher, with reducing variability as Cetygo increases on each axis. The two clusters are largely tissue specific, with 97.2% of the samples with Cetygo for **Model 6CellBlood** <0.1 being blood samples, and 78.4% of

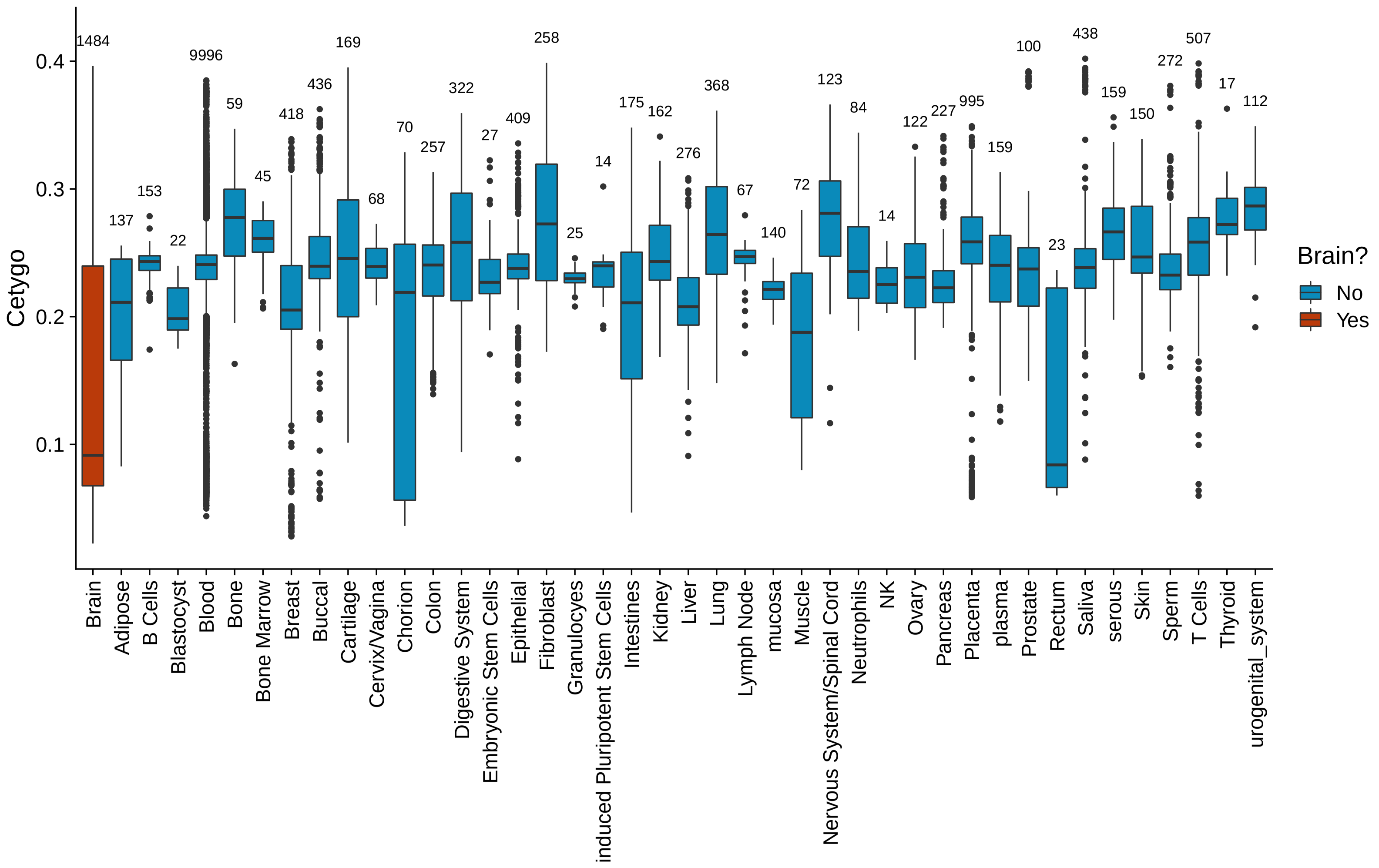


Figure 3.21: **A box plot of Cetygo from predictions by Model 2CellPFC varied tissue type samples in Dataset GEO.** The sample size of each tissue is shown above its respective box plot, with the blood tissue box plot highlighted in red.

samples with Cetygo for **Model 2CellPFC** <0.1 being brain samples (see **Table 3.7** for the number of samples in each section of the plot). However, the fact neither percentage are at 100% demonstrates that there are samples being deconvoluted accurately in a tissue not matching their GEO annotation, suggesting that said annotations are incorrect. Given that the clusters for both models are tightest at Cetygo below 0.1, it was chosen as the optimum soft threshold. Of note, the soft threshold appears to be the upper limit for the two Cetygo clusters across both models, suggesting that the range of Cetygo is consistent between models, regardless of the differing numbers of cell types and DNA methylation sites used. A higher threshold could have been selected, such as 0.175 which would signify the beginning of the strong correlation between model Cetygo's, however, the more stringent soft threshold was selected to encourage careful QC of deconvolution predictions. Using 0.1, no sample was said to perform well across both models, which biologically should never be the case. The soft threshold can be utilised when no easy comparison can be made between samples being deconvoluted.

3.7 Cetygo has utility as a metric for quantifying the effectiveness of cellular purification

DNA methylation profiles of purified tissue are being used more commonly in DNA methylation studies in order to avoid the issue of cellular heterogeneity. The tissue purification methods (such as FANS, described in Section 4.1.1.1) used are not always perfect, however, and may not therefore result in perfectly pure cell type populations. Where reference data exists for the tissue being purified, applying reference based deconvolution to purified samples can act as a QC metric for purification. As with all deconvolution predictions, their reliability is not guaranteed and so Cetygo can be utilised in conjunction with cell type estimates.

To demonstrate Cetygo's utility for aiding the assessment of purification quality, deconvolution was applied to three datasets: **Dataset E-Risk**, containing 173 purified blood samples, **Dataset Pai**, containing 100 NeuN+ samples purified from PFC tissue, and the subset of purified blood samples from **Dataset GEO**.

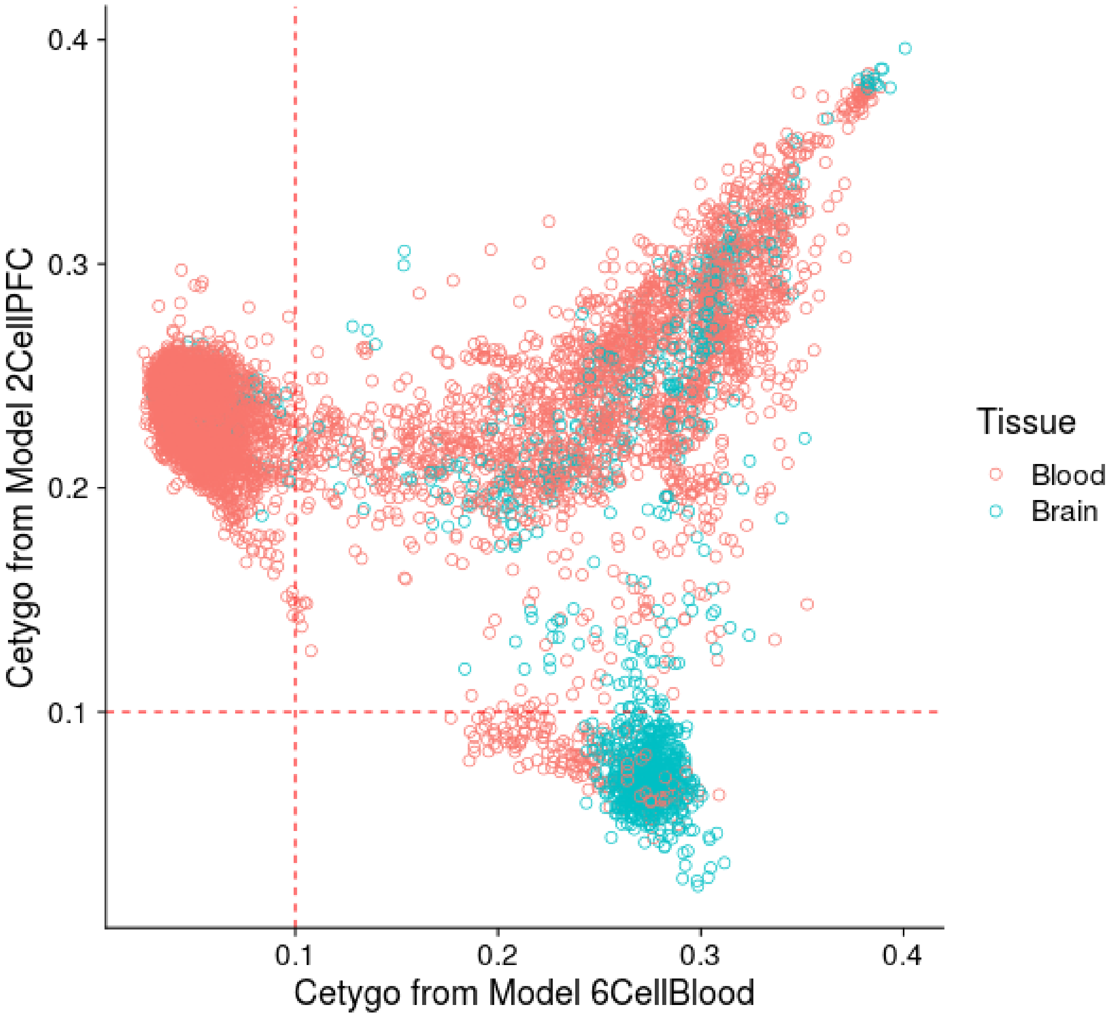


Figure 3.22: **A comparison of Cetygo ascertained from Model 6CellBlood and Model 2CellPFC across brain and blood samples within Dataset GEO.** Points are outlined by their tissue type, with blood in pink and brain in blue. Red dashed lines represent the selected soft threshold for Cetygo.

Table 3.7: **A summary of the number of blood and brain samples from Dataset GEO with Cetygo greater than or less than 0.1 for Model 2CellPFC and Model 6CellBlood.** This table contains the quantification of **Figure 3.22**, in which sections are defined by the bisection of the x and y axis by the Cetygo soft threshold 0.1

Blood Tissue	Blood Cetygo < 0.1	Blood Cetygo > 0.1
Brain Cetygo > 0.1	7757	2045
Brain Cetygo < 0.1	0	194

Brain Tissue	Blood Cetygo < 0.1	Blood Cetygo > 0.1
Brain Cetygo > 0.1	224	480
Brain Cetygo < 0.1	0	705

Cetygo applied to purified blood samples in Dataset E-Risk

Dataset E-Risk contains the DNA methylation profiles of purified blood samples sorted for the following cell types: B cells, CD4$^+$ T cells, CD8$^+$ T cells, monocytes, and granulocytes (n = 173). Cellular proportions were predicted using **Model 6CellBlood**.

Cetygo was below 0.1 for all purified cell populations (**Figure 3.23**), with all estimated cellular proportions >0.7. Interestingly, Cetygo correlated negatively with the predicted proportion of B cells (Cor = -0.61), granulocytes (Cor = -0.68) and monocytes (Cor = -0.91), which may be driven by technical noise in samples with higher Cetygo resulting in a DNA methylation profile less similar to the reference data in **Model 6CellBlood**.

Cetygo applied to purified NeuN+ samples in Dataset Pai

Dataset Pai contains the DNA methylation profiles of 100 FANS sorted NeuN+ PFC samples (see Section 4.1.1.1 for details on FANS sorting). Cellular proportions were predicted using **Model 2CellPFC**.

All samples have Cetygo less than the soft threshold, 0.1 (**Figure 3.24**), suggesting that the predicted proportions are reliable. The correlation between Cetygo and predicted proportion is very weak and negative (Cor = -0.18), suggesting that Cetygo does not index the efficacy of purification. Anecdotally, however, an elevated average Cetygo around samples with a predicted proportion of 0.8 can be observed, which may suggest that some factor across those samples (e.g. poor purification) may be limiting prediction quality. More data is required to assess further.

Cetygo applied to purified T cell samples within Dataset GEO

Within **Dataset GEO**, a subset of blood samples were classed as purified cell types.

To demonstrate Cetygo's potential utility as a metric for purified samples, an exemplar T cell dataset within **Dataset GEO** was used (**Figure 3.25**).

The first two samples seen in the plot have rough proportions matching those expected for bulk blood (see **Table 3.1**), with low Cetygo, and as such are likely unpurified

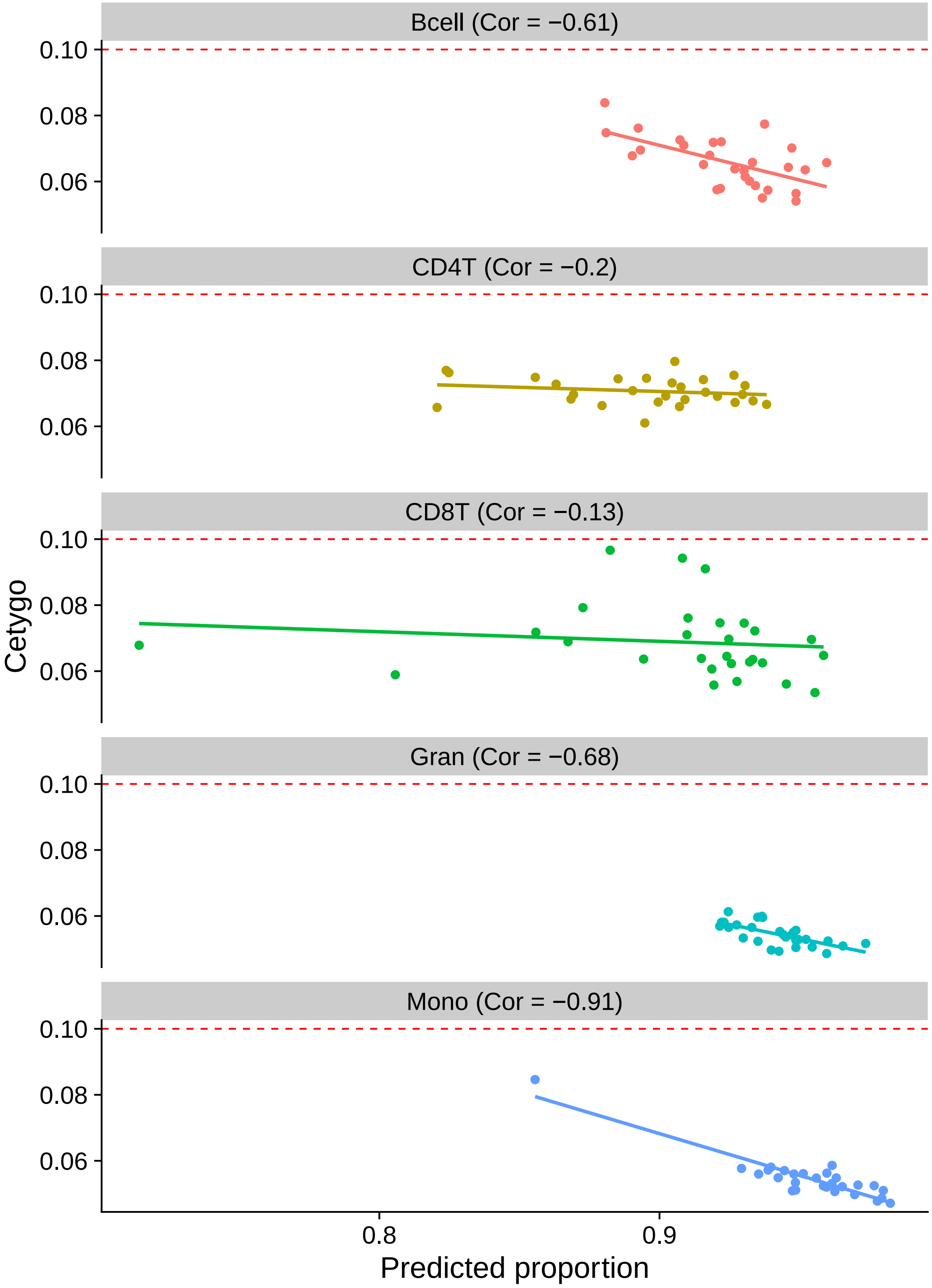


Figure 3.23: **A scatter plot between Cetygo and the predicted proportion of purified cell types in Dataset E-Risk.** The red dashed line on each plot represents the selected soft threshold for Cetygo.

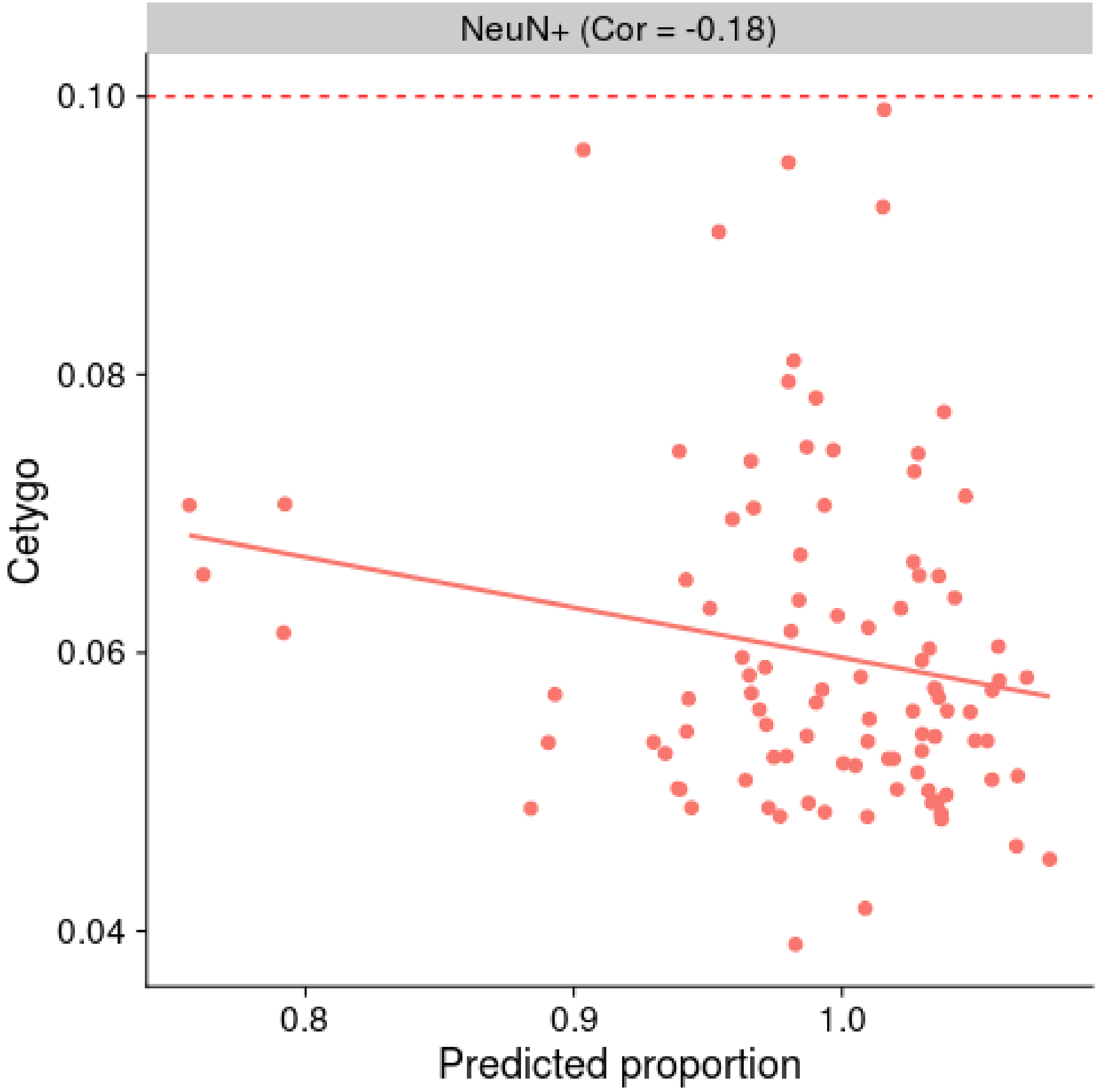


Figure 3.24: A scatter plot between Cetygo and the predicted proportion of purified NeuN+ cells in **Dataset Pai**.

blood samples. In contrast, three samples had Cetygo above 0.1, at ∼0.35 which is comparable to that of randomly shuffled DNA methylation data (**Figure 3.5**), suggesting that these samples were not accurately deconvoluted, potentially due to technical error.

This highlights the utility of Cetygo as a QC metric for cell type purified data, allowing us to do more than just quantify cell type proportions, but also understand how accurate those proportions might be. The use of Cetygo in this context would allow researchers to identify misannotated or sample swapped data, and differentiate such samples from those for which cell type purification has not worked.

3.8 Confirming model applicability across independent datasets using Cetygo

When applying a deconvolution model to an independent dataset for which the cellular composition is unknown, prediction accuracy cannot be guaranteed. As such, Cetygo (in conjunction with the pre-selected soft threshold) would be of utility to assess model applicability across independent data. Here, deconvolution models were applied to DNA methylation datasets where some phenotypic disparity exists between the samples used to generate the reference data and independent bulk tissue data data. Differences include: low quality DNA methylation data (i.e. technically noisy samples), batch effects, sex, age, or ethnicity. Cell type specific DNA methylation differences within the model are typically associated with large magnitudes of effect, and should, therefore, dwarf differences included by other factors, however, it is unknown how subtle variations may affect deconvolution accuracy. It may be expected then that, where the disparity between reference and input data resulted in DNA methylation differences between groups at the DNA methylation sites within the deconvolution model, Cetygo may reflect said deviations.

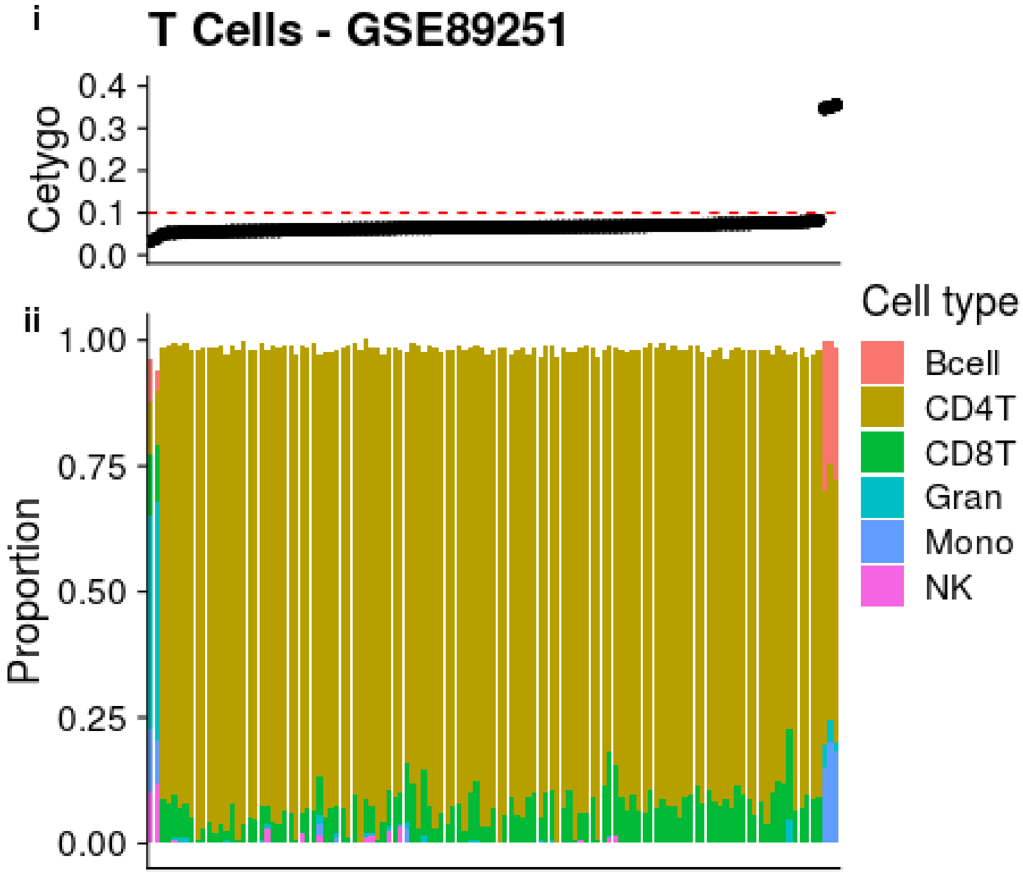


Figure 3.25: **A summary of Cetygo, estimated and actual cell type proportions of the dataset with GEO accession number GSE89251 in Dataset GEO samples.** Plots contain (i) Cetygo per sample, (ii) the predicted proportions per sample.

3.8.1 Samples with low median array intensity have elevated Cetygo

First, the relationship between Cetygo and median array intensity was investigated. Array intensity describes the fluorescent signal from the array which is used to quantify methylated and unmethylated probes (described in Section 1.3.4.2). Lower median intensities are indicative of a high noise to signal ratio and generate less sensitive DNA methylation profiles. This lack of accuracy in estimating DNA methylation is likely to influence the accuracy of estimating cellular composition. Typically, samples are filtered based on their median methylated or unmethylated intensity, the filtering threshold for which varies from study to study, but generally falling between 500-2000 depending on the clustering of samples. It was anticipated that at lower median intensities, where data quality is lower, a negative correlation would exist between Cetygo and intensity, as the presence of additional noise across the data would lead to a larger disparity between expected and actual DNA methylation profiles. This would most likely be followed by a plateau, signifying the point where the intensity is sufficient and no longer limiting the data quality.

The dataset used to compare array intensity across whole blood samples was **Dataset PPMI** (**Table 6.3**), from the PPMI cohort, containing 524 samples comprised of healthy controls, patients with early PD and subjects without evidence of dopaminergic deficiency. Cellular proportions were estimated using **Model 6CellBlood**, and Cetygo calculated.

The expected relationship between median intensity and Cetygo is not particularly strong across the full range of methylated intensities (**Figure 3.26A**), however it is anecdotally evident across unmethylated intensities, which were systematically lower than their methylated counterparts (**Figure 3.26B**), with a negative correlation between Cetygo and unmethylated intensity below ∼1300. This suggests that, in this dataset, median unmethylated intensity is no longer a limiting factor to data quality above 1300.

There exists variation across Cetygo that is not explained by median intensity alone, demonstrating that Cetygo provides information independent of intensity.

A caveat to this exploration is that it can be challenging to find poor quality

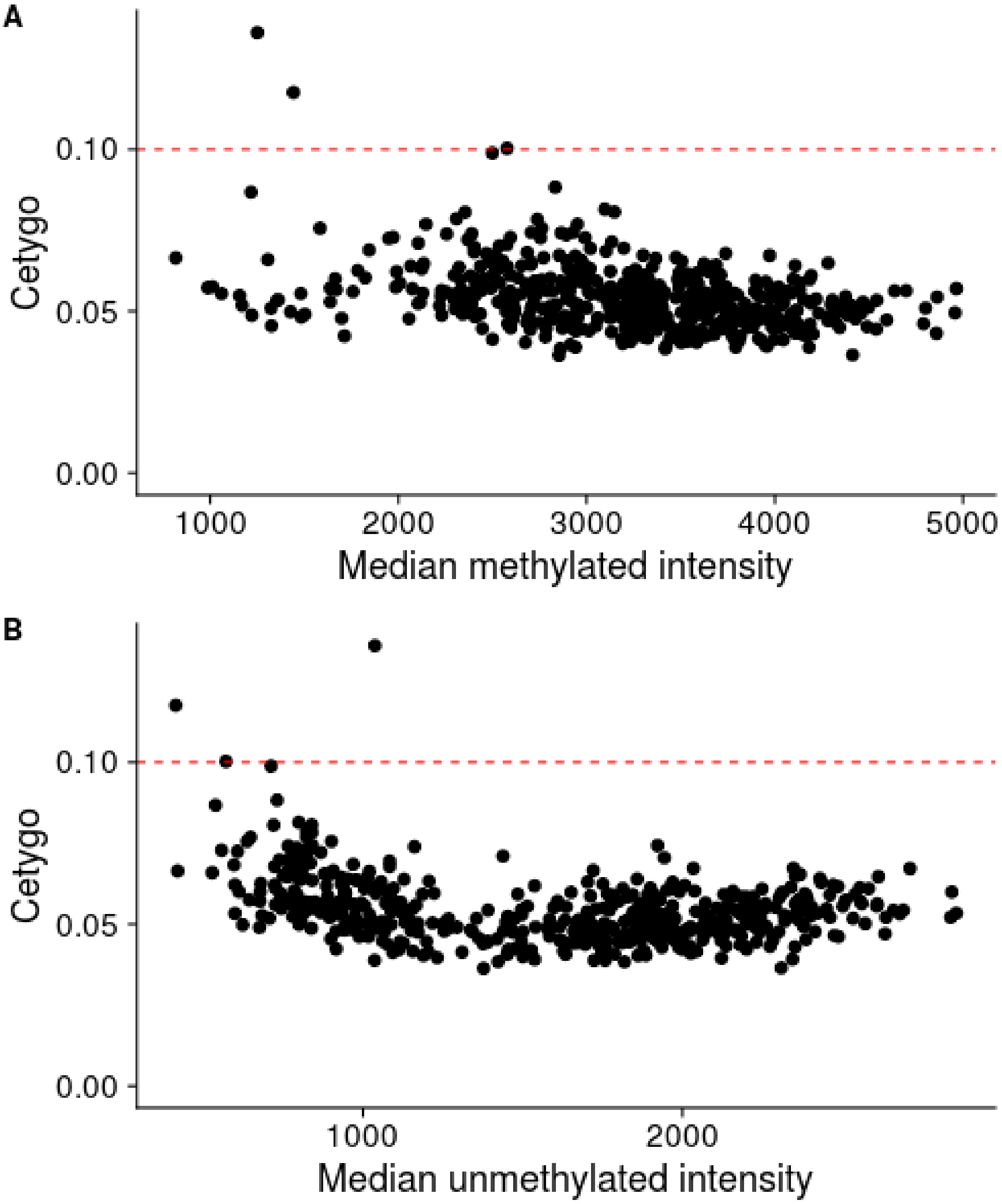


Figure 3.26: **A scatter plot between median array intensity, and Cetygo in blood samples from Dataset PPMI.** Plots show median A) methylated and B) unmethylated array intensity and Cetygo. Deconvolution was carried out using **Model 6CellBlood**. The red dashed line represents the selected soft threshold for Cetygo.

blood datasets to assess the relationship more thoroughly. In general, blood samples are easily obtained and come with fewer technical challenges compared to certain solid tissues, and as such, the relationship between median array intensity and Cetygo was next assessed across a purified PFC dataset, **Dataset BDR purified**. The data contains DNA methylation profiles for FANS purified nuclei: 27 NeuN+, 28 Sox10+ (NeuN-/Sox10+), 21 Double- (NeuN-/Sox10-), 3 IRF8+ (NeuN-/Sox10-/IRF8+), and 2 Triple- (NeuN-/Sox10-/IRF8-) samples, as well as 26 Total (unpurified nuclei). DNA methylation was quantified using EPIC (See **Figure 4.14** for the stain relationships). Cellular proportions were estimated using **Model 2CellPFC**, and Cetygo calculated.

The anticipated relationship between intensities and Cetygo can be seen more clearly in this dataset than the blood counterpart (**Figure 3.27**). The correlation between median methylated intensity and Cetygo is -0.6, with a highly significant linear fit (p = 8.37e-12) (**Figure 3.27A**). A similar relationship, although weaker, can be seen with the median unmethylated intensities, with Cor = -0.458, P = 7.12e-07 (**Figure 3.27B**). Cetygo varies across all cell types, however, IRF8+ is elevated for all samples. IRF8+ makes up a small proportion of the NeuN- population present in **Model 2CellPFC**, and so the elevated Cetygo here is more likely due to a lack of model granularity, than intensity alone (model granularity is explored further in Chapter 4).

Findings across blood and brain samples used here exhibit a weak relationship between median intensity and Cetygo. The intensity below which Cetygo is elevated differs between figures; the relationship is not consistent between datasets, which may be due to the tissue or deconvolution model used, or driven by technical factors of DNA methylation profiling.

3.8.2 A disparity in dataset origin effects, sex, age, and ethnicity between reference and input data can drive increased Cetygo

Cetygo is a measure of difference between the reference and input data to be deconvoluted, specifically between the input DNA methylation profile and the profile expected of a sample if predicted proportions were accurate (see Section 3.2). Possible differences between reference and input data include demographics, such as sex, age, and ethnicity

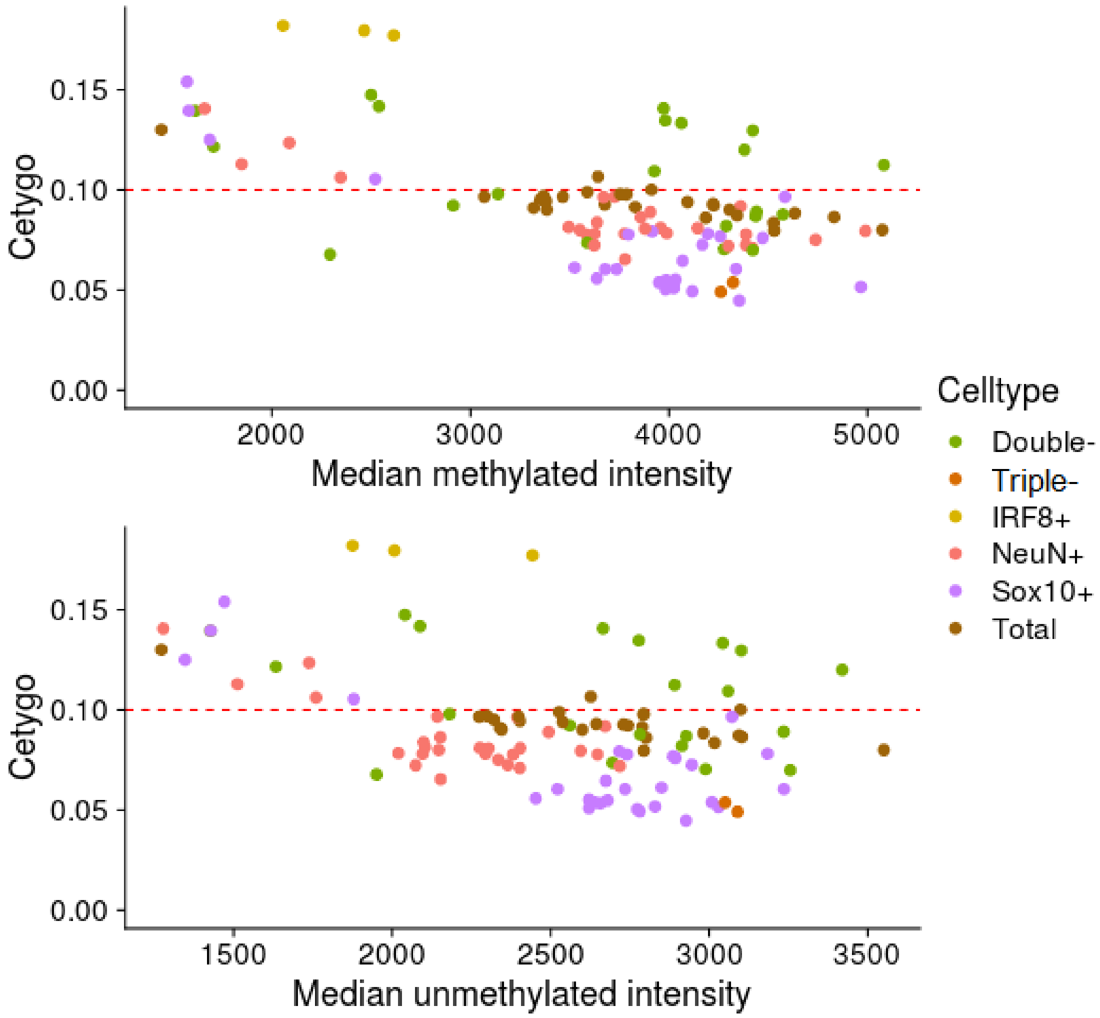


Figure 3.27: **A scatter plot between median array intensity, and Cetygo in purified PFC samples from Dataset BDR purified.** Plots show median A) methylated and B) unmethylated array intensity and Cetygo. Deconvolution was carried out using **Model 6CellBlood**. The red dashed line represents the selected soft threshold for Cetygo.. Cell populations were sorted using a FANS protocol (see Section 4.1.1.1), where NeuN+ were positively selected using NeuN, Sox10+ is comprised of NeuN-/Sox10+ nuclei, Double- is NeuN-/Sox10-, IRF8+ is NeuN-/Sox10-/IRF8+, is Triple-NeuN-/Sox10-/IRF8- and Total is comprised of unsorted nuclei.

(each of which are known to result in differential methylation (Horvath, 2013; Yusipov et al., 2020; Yousefi et al., 2015; Zhang et al., 2011)), and technical effects, including dataset origin (as seen in **Figure 3.28**). Dataset origin effects are defined as systematic technical variation that can result in noise across DNA methylation profiling that will be specific to the study in which samples were profiled. Systematic differences may be caused by factors such as operator bias or reagent batches used. Within a study, efforts are often made to minimise batch effects (i.e. the technical noise within a study), such as consistency in person running the arrays, parallelising sample profiling, and randomising samples across the array chips used, however, this will not have been done between independent projects. **Model 6CellBlood**, used in this Section, was generated using **Dataset Reinius** which contains healthy male samples with mean age 38 ± 13.6 years of Swedish ancestry.

To assess dataset origin, sex, and age differences on Cetygo, **Model 6CellBlood** was applied to four external cohorts: **Dataset EXTEND**, which contained 489 male and 686 female samples with age range 56.3 ± 11.7, **Dataset Understanding Society** containing 591 male and 643 female samples with age range 58 ± 15, **Dataset IoP**, containing 503 male and 296 female samples with age range 44 ± 11.3, and **Dataset EUGEI** with 512 male and 422 female samples with age range 35.2 ± 12.8 (**Table 6.3**). All datasets had been normalised prior to deconvolution.

Dataset origin and demographic effects can result in differential DNA methylation (Mill and Heijmans, 2013), and as such, here, the difference between Cetygo across the four aforementioned datasets was observed with the aim of generally comparing the resulting distributions.

For all four datasets, the vast majority of samples had Cetygo below the soft threshold of 0.1 (4/4143 samples had Cetygo >0.1), suggesting that blood deconvolution was reliable across datasets, excluding outliers (**Figure 3.28**).

The magnitude of difference in Cetygo between datasets was small (median Cetygo: **Dataset Understanding Society** - 0.061, **Dataset EXTEND** - 0.0524, **Dataset EUGEI** - 0.0555, **Dataset IoP** - 0.0458), but anecdotally, dataset specific distributions can be observed in the violin plots in **Figure 3.28**.

When applying deconvolution through R packages *minfi* or *wateRmelon*, it is standard procedure to normalise input datasets with training data to reduce the difference in distribution of DNA methylation between the reference and sample profiles. In doing so, a deconvolution model is generated for each new input dataset after normalisation, which may reduce the dataset origin effects observed. Normalisation of training and input data was not carried out here to allow the application of the same deconvolution model (**Model 6CellBlood**) across all datasets, minimising potential confounding with different DNA methylation sites used.

The cause of these dataset origin related differences could be numerous and could include technical effects, as well as sex, age or ethnicity of samples profiled, which are explored below.

When comparing across sex, Cetygo was significantly higher for female compared to male samples across the four datasets (p = 2.07e-31). This shows that predictions are more accurate if the training and testing data are matched for sex, as **Model 6CellBlood** was generated using only male samples (**Figure 3.29**). It is worth noting that 8 of the 600 predictive DNA methylation sites included in **Model 6CellBlood** were located on the X chromosome (**Table 3.9**), all of which had significantly different DNA methylation between male and female samples (**Figure 3.30** and **Table 3.8**), which may explain this disparity.

The exact age of **Dataset Reinius** (reference data of **Model 6CellBlood**) was not available, however, the mean and standard deviation were (38 ± 13.6 years). When assessing the impact of age on predictions, Cetygo was observed to differ significantly when comparing samples within a standard deviation of the reference data mean and those outside (p = 1.05e-05, two-sided t-test), however, this may be driven by dataset batch effects, as the datasets differ in age range (**Figure 3.31**). The correlation between age and Cetygo was low, at 0.211.

While significantly different, the mean differences in Cetygo between sexes and age across datasets was again very small, demonstrating the high sensitivity of Cetygo for identifying differences between training and testing data. It also suggests that mismatching sex and age may not be detrimental to prediction accuracy on the whole,

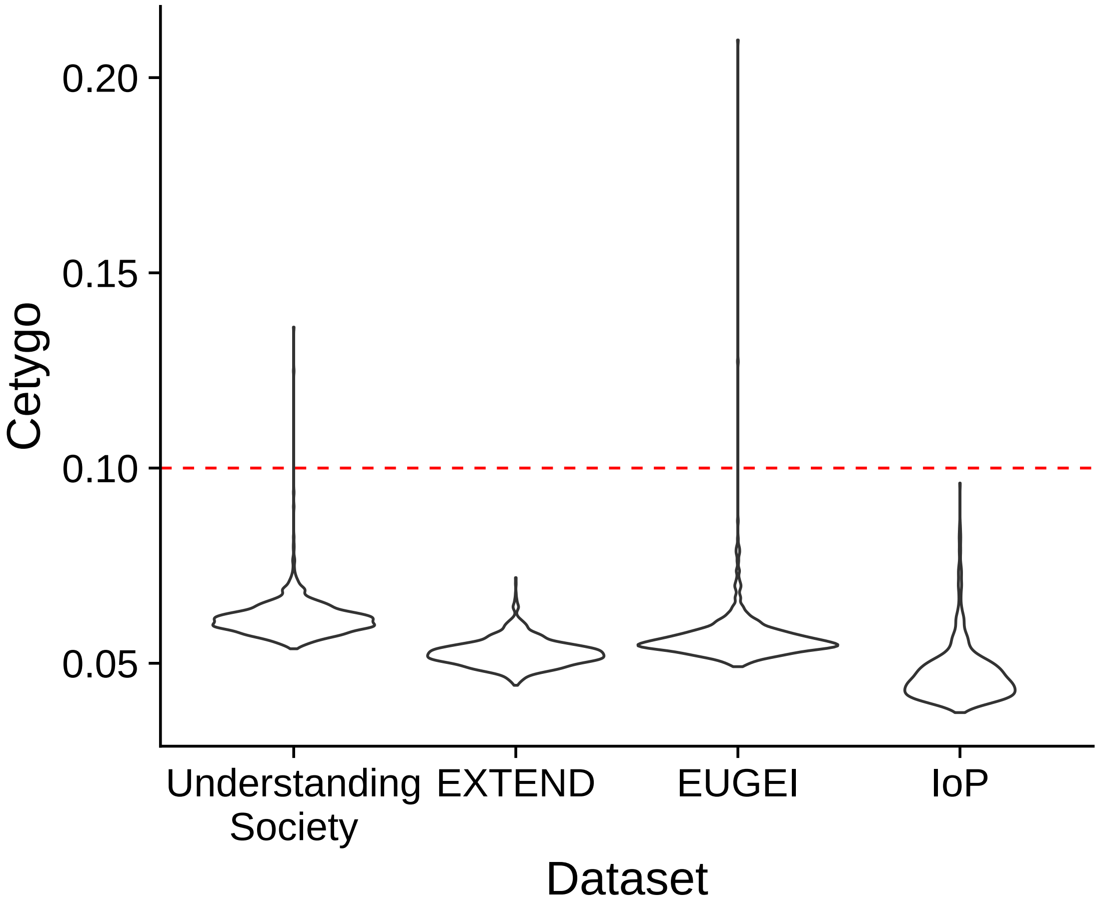


Figure 3.28: **A violin plot of Cetygo ascertained from Model 6CellBlood across blood samples in Dataset EXTEND, Dataset Understanding Society, Dataset EUGEI, and Dataset IoP.** The red dashed line represents the selected soft threshold for Cetygo.

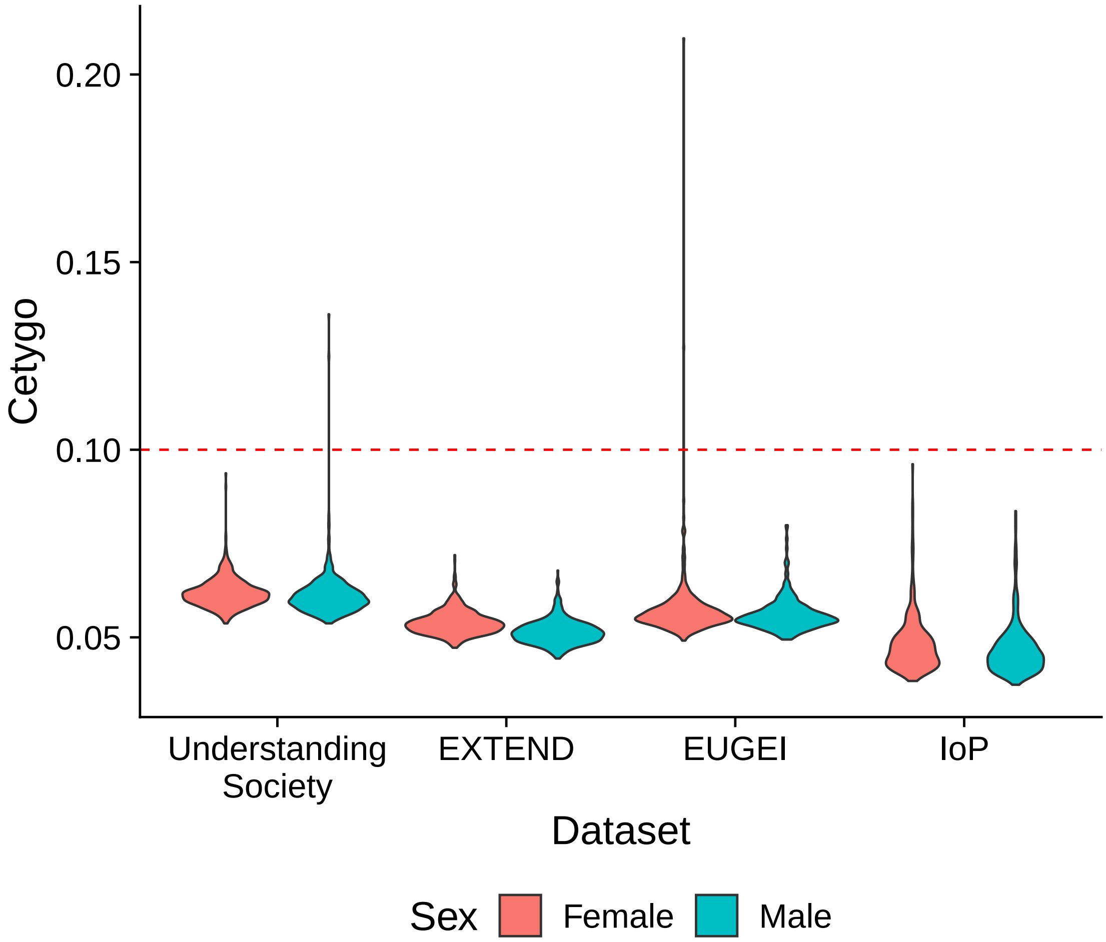


Figure 3.29: **A violin plot of Cetygo ascertained from Model 6CellBlood across blood samples in Dataset EXTEND, Dataset Understanding Society, Dataset EUGEI, and Dataset IoP divided by sex. Model 6CellBlood** was generated using male training data. The red dashed line represents the selected soft threshold for Cetygo.

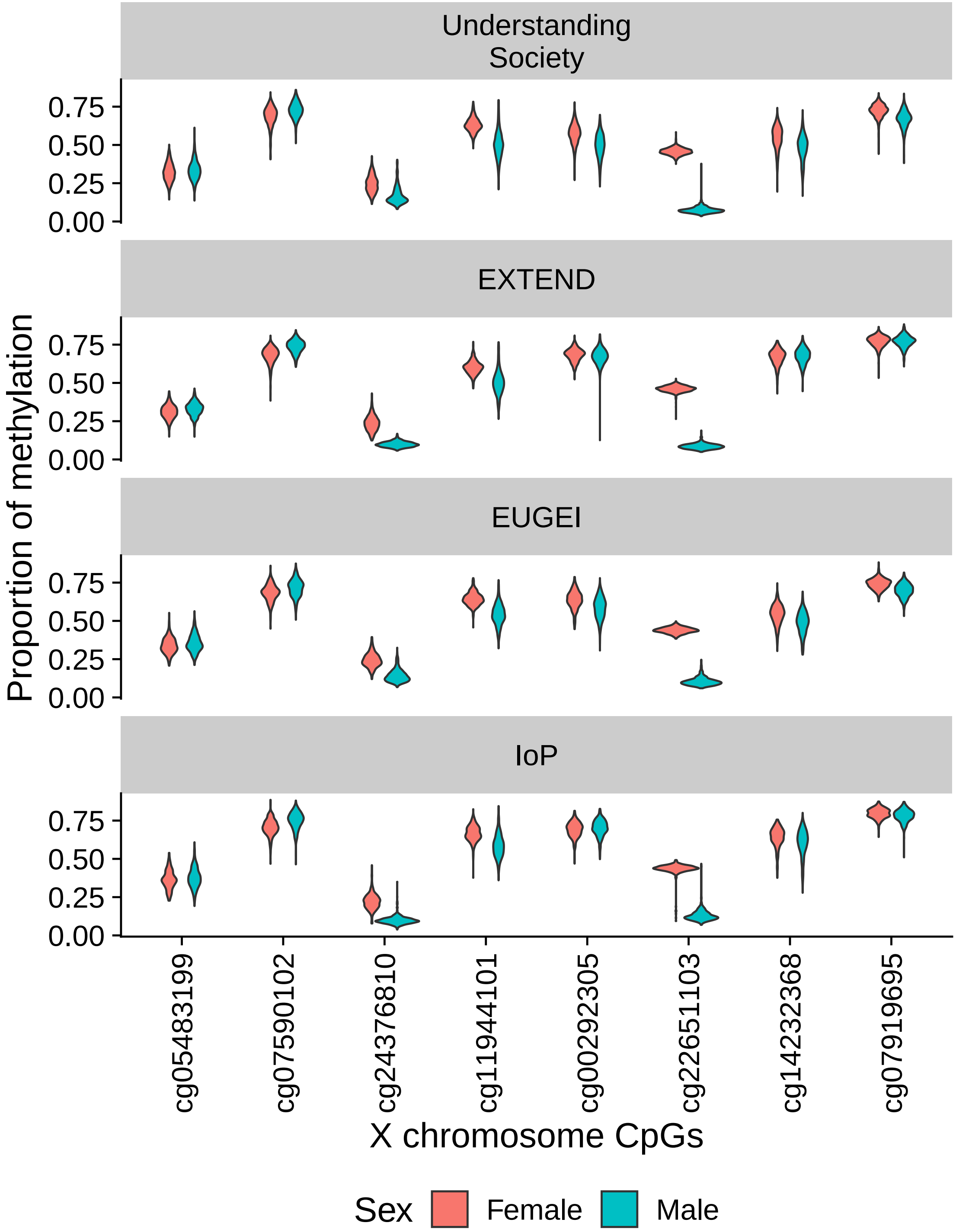


Figure 3.30: **A violin plot comparing the DNA methylation of X chromosome CpGs included in Model 6CellBlood between male and female samples**. Samples from **Dataset EXTEND**, **Dataset Understanding Society**, **Dataset EUGEI**, and **Dataset IoP** were utilised.

Table 3.8: **A table showing the statistical comparison between DNA methylation of X chromosome CpGs in Model 6CellBlood between male and female samples.** Samples from **Dataset EXTEND**, **Dataset Understanding Society**, **Dataset EUGEI**, and **Dataset IoP**, were utilised, as seen plotted in **Figure 3.30**.

CpG ID	P-value	Mean methylation		
			Female	Male
cg05483199	1.9e-28	0.328	0.349	
cg07590102	$<$ 5e-324	0.686	0.733	
cg24376810	$<$ 5e-324	0.239	0.126	
cg11944101	$<$ 5e-324	0.629	0.532	
cg00292305	4.4e-14	0.637	0.617	
cg22651103	$<$ 5e-324	0.451	0.0989	
cg14232368	1.2e-21	0.597	0.567	
cg07919695	1.3e-31	0.753	0.732	

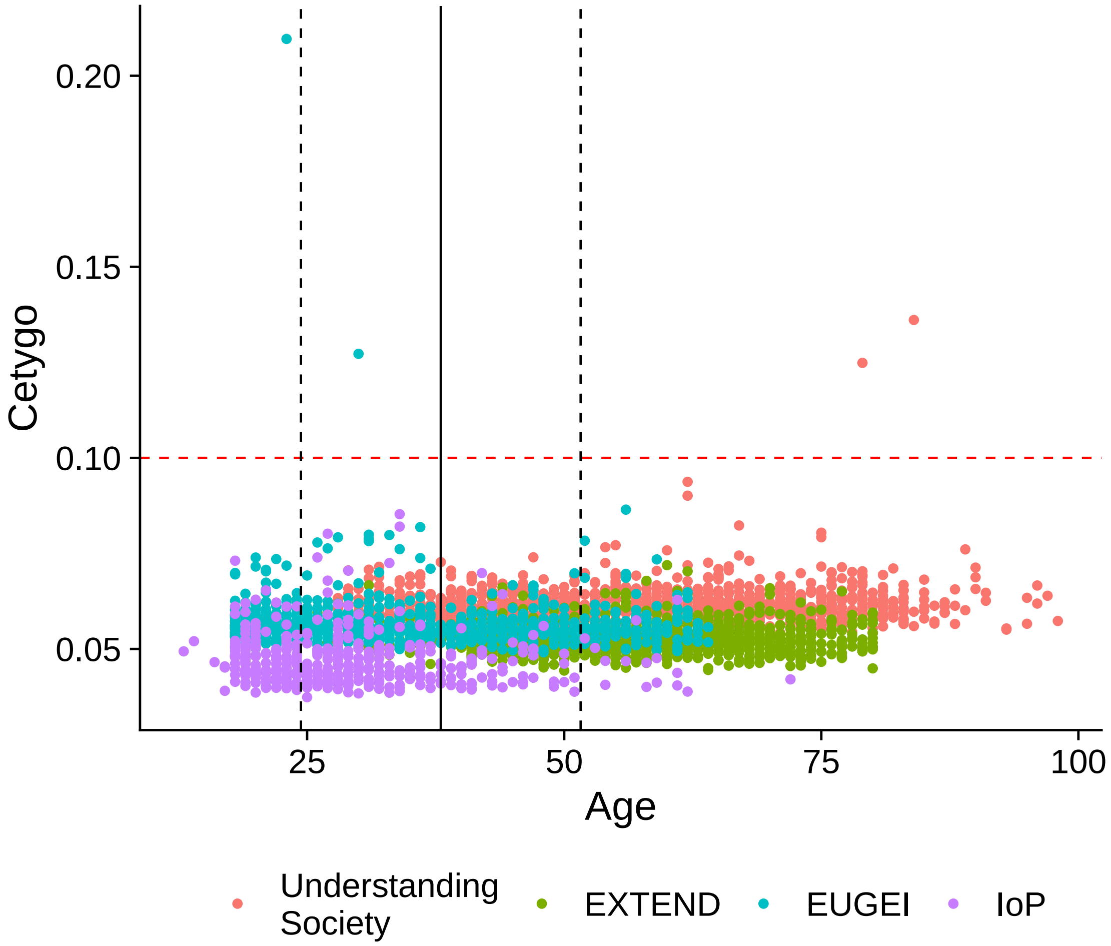


Figure 3.31: **A scatter plot of age and Cetygo ascertained from Model 6CellBlood across blood samples in Dataset EXTEND, Dataset Understanding Society, Dataset EUGEI, and Dataset IoP.** Ages were not available per sample for **Dataset Reinius**, with which **Model 6CellBlood** was trained, but the mean age is plotted by the black vertical line, with the standard deviations plotted using black dashed lines. The red dashed line represent the selected soft threshold for Cetygo.

at least in the age ranges tested.

The impact of ethnicity on predictions was investigated using **Dataset IoP**, containing 376 European and 424 non-European samples, and **Dataset EUGEI**, containing 634 European and 300 non-European samples (**Figure 3.32**). The deconvolution model, **Model 6CellBlood**, was generated in European (Swedish) samples. In both datasets, a significant difference was found between the Cetygo of European versus non-European samples (**Dataset IoP** P = 5.542e-29, **Dataset EUGEI** P = 7.2e-05, using a two-sided t-test), with higher Cetygo in non-Europeans. As in batch, sex and age differences, the mean differences between groups were small (<0.001 difference between means).

The findings of this Section demonstrate the sensitivity of Cetygo and its potential utility to distinguish model applicability across independent data to be deconvoluted.

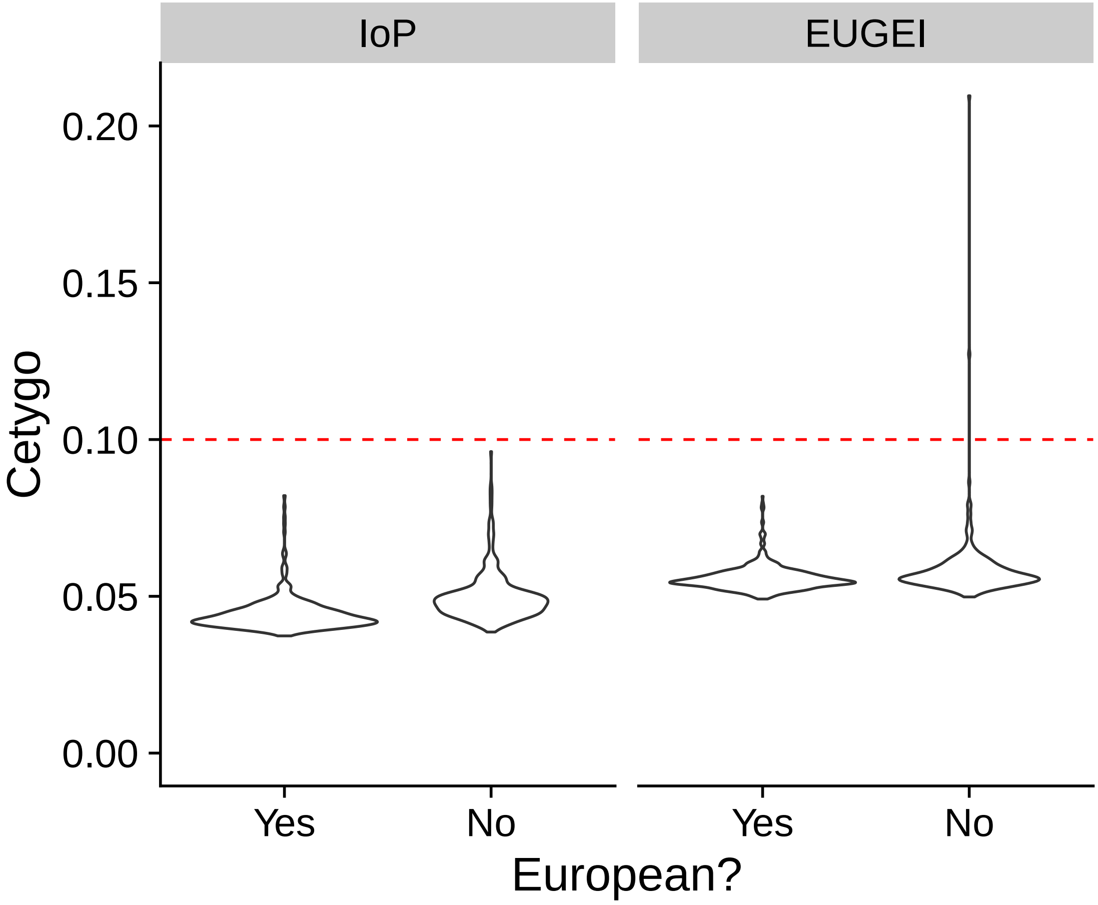


Figure 3.32: **A violin plot of ethnicity and Cetygo ascertained from Model 6Cell-Blood across blood samples in Dataset EUGEI, and Dataset IoP. Model 6Cell-Blood** was generated using Swedish (European) training data. The red dashed line represents the selected soft threshold for Cetygo.

3.9 Discussion

3.9.1 Overview of results

Reference based cellular deconvolution from DNA methylation data generated in bulk tissue is routinely carried out to estimate cellular composition in EWAS, but to date, no methods exist to assess the accuracy of prediction estimates. Deviations in the applicability of deconvolution models, either due to issues of reference data applicability or poor bulk sample quality, can lead to inaccurate proportion estimates. In this Chapter, a reference based deconvolution error metric, Cetygo, was established which quantifies the difference between the input DNA methylation profile and the expected DNA methylation profile if cell type estimates were accurate.

Cetygo was shown to be able to quantify deconvolution accuracy using simulated whole blood and brain tissue. To investigate the relationship between Cetygo and accuracy, first, Cetygo and accuracy were compared across data with increased simulated noise, which could represent technical or biological noise not relating to cellular composition. Second, samples containing increasing proportions of a blood cell type not included in a blood deconvolution model were simulated. Both simulations demonstrated that Cetygo increased linearly with prediction accuracy, demonstrating the utility of Cetygo for quantifying prediction quality. It is well established that differential DNA methylation patterns identified across highly heterogeneous tissue are liable to confounding and misinterpreted associations (Koestler et al., 2016; Adalsteinsson et al., 2012; Reinius et al., 2012; Koestler et al., 2012). As such, DNA methylation studies adjust for cellular composition when identifying differential DNA methylation, often using composition estimates from deconvolution models where empirical measurements are not available (Comes et al., 2020; Hannon et al., 2021a; Rovira et al., 2020, to name a few). With the accuracy of said composition estimates unknown, using them to adjust for cellular heterogeneity could further confound associations and reduce power to detect true biological differences. Cetygo allows researchers to gain insight into the validity of their deconvolution estimates, allowing them to make informed decisions as to whether reference based deconvolution is appropriate for inclusion as covariates in their analysis.

More likely, Cetygo will identify outlier samples that should be excluded from further analyses. Alternative deconvolution methods, i.e. reference-free deconvolution, may be more suitable where composition estimates are shown by Cetygo not to be reliable (e.g. the reference panel is incomplete), as they use data driven approaches to account for sources of variation in bulk tissue measurements of DNA methylation.

Deconvolution models vary in applicability to independent input data, and Cetygo can be used to compare between models. This was shown using blood deconvolution models that excluded a subset of available blood cell types, in which the mean and range of Cetygo was found to inversely correlate with the number of cell types in the model. With the constant development and optimisation of lab based techniques to characterise DNA methylation from purified nuclei (Policicchio et al., 2020a) or single cells (Karemaker and Vermeulen, 2018), more reference datasets will almost certainly become available in the near future. As such, reference datasets for the same tissue, containing different combinations of cell types might be available and Cetygo can be utilised to compare between models, allowing researchers to implement the most meaningful and accurate deconvolution approach across their dataset. Novel models can be utilised to reanalyse bulk data in which an inadequate deconvolution model may previously have been used to estimate cellular composition. Additional model cell types can provide additional granularity to these deconvolution approaches (which is explored in the context of brain deconvolution in Chapter 4). Being able to test the applicability of deconvolution models would not only improve certainty when applying a model, but also, where a model is shown to be widely applicable, reduce the need to generate novel reference datasets with very specific utility.

A central theme of this thesis is reproducibility, which includes open science practices. To that end, the code to calculate Cetygo has been made publicly available in https://github.com/ds420/CETYGO. To allow for easiest application by researchers, Cetygo has also been integrated into the most recent version of *wateRmelon*, which allows users to generate error as they deconvolute their samples with no additional labour.

3.9.2 Limitations and future work

In this Chapter, Cetygo was only assessed using Houseman's reference based deconvolution algorithm. Various other reference based deconvolution algorithms exist, such as CETS (Guintivano, Aryee and Kaminsky, 2013), CIBERSORT (Newman et al., 2015; Teschendorff et al., n.d.), EpiDISH (Newman et al., 2015; Teschendorff et al., n.d.), and IDOL (Koestler et al., 2016), which out-perform Houseman's algorithm in certain scenarios (Salas et al., 2018; McGregor et al., 2016). This Chapter only utilised Houseman's algorithm (the most commonly used method in epigenetic epidemiology) as proof of principal, however, Cetygo should be applicable to the other algorithms and future work should aim to assess its performance across these alternative approaches. Indeed Cetygo might be a useful method of comparing these algorithms against each other.

One caveat of this Chapter is that only blood and brain tissue was used to assess Cetygo, both of which are commonly used within epidemiological studies of DNA methylation. Deconvolution models for these two tissues differ in characteristics, with the blood model containing all common cell types, whereas the brain model, while containing all common cell types, only distinguishes them into two major groups which may subsequently contain residual cellular heterogeneity. Furthermore, while a bulk blood sample is comprised mostly of granulocytes (∼50-80%), brain tissue is on average made up of a similar proportion of neuronal and glial cells (Bartheld, Bahney and Herculano-Houzel, 2016), which may affect deconvolution, as it was shown in this Chapter that deconvolution models could have cell type biases. Despite this, Cetygo was shown to be applicable to models of both tissue types and, as such, the framework is expected to be applicable to other tissues for which cellular reference datasets are available, such as saliva (Middleton et al., 2020).

Reference based deconvolution is not the only type of prediction that genome wide DNA methylation profiles can be used to estimate. For example, epigenetic clocks, which estimate age (Hannum et al., 2013; Horvath, 2013; Shireby et al., 2020; Steg et al., 2021), and smoking status predictors (Bollepalli et al., 2019) have also been utilised when phenotypic data is not available. A similar framework could be utilised to assess the quality of other DNA methylation based predictions to allow users to explore their validity.

DNA methylation is not the only datatype in which cellular composition can confound research. Similar is seen in gene expression and chromatin accessibility data, for which deconvolution methods also exist (Li et al., 2020; Titus et al., 2017; Donovan et al., 2020). As such the Cetygo framework could be expanded to assess accuracy of cellular estimates across other types of genomic data.

3.9.3 Conclusion

This Chapter presents the first error metric for reference based deconvolution where the true cellular composition is unknown. Results demonstrate that Cetygo has utility in profiling deconvolution accuracy and model applicability. Cetygo has been made publicly available in the interest of open science, and has been integrated into the commonly used DNA methylation QC package *wateRmelon* for easy application during data cleaning. Greater clarity and reporting of deconvolution prediction quality will improve the transparency of deconvolution and the subsequent use of estimated cellular proportions to adjust for cellular composition within DNA methylation studies, improving overall reproducible science practices.

3.10 Additional methods

The main methods used in this Chapter are described where used. Here the additional methods are summarised.

3.10.1 Deconvolution model generation

All deconvolution models applied in this Chapter were generated using the Houseman algorithm, the stages of which are described in Section 3.1.1. The first stage of deconvolution model development is comprised of acquiring appropriate reference data. The reference data used is tissue specific, and are detailed in the following Sections (3.10.1.1 and 3.10.1.2). Generally, a reference dataset will contain the DNA methylation profiles of cell types within a tissue across multiple samples. To allow for the testing of Cetygo, a subset of samples were set aside from the reference dataset to allow for the simulation of data with varying proportions of each cell type (see Section 3.4 for validation using simulated data).

The second stage of deconvolution model generation involves selecting DNA methylation sites from the reference data which distinguish one cell type from all others, and then utilise a deconvolution algorithm to predict the composition of bulk samples using said sites. This thesis utilises *minfi*'s application of the Houseman algorithm, contained in the R function *estimateCellCounts()*. *estimateCellCounts()* contains two internal functions (also used in *wateRmelon*):

- **pickCompProbes()** - selects DNA methylation sites from the reference data that can distinguish each cell type. By default, 50 sites are selected for each cell type in each direction, that is, 50 sites in which DNA methylation in one cell type compared to all others, and 50 sites at which DNA methylation is lower. The default number of sites was used across this thesis so as to create a model akin to those that would most commonly be used in the QC pipeline to ensure that Cetygo was applicable to the norm.
- **projectCellType()** - uses the sites selected using *pickCompProbes()* to estimate the cellular proportions of a bulk sample. The output contains a proportion for

each cell type available in the reference data, and was adapted to additionally output Cetygo.

By using the two internal functions, rather than *estimateCellCounts()*, the normalisation step in which reference and input data are normalised together, that differs between *minfi* and *wateRmelon*, was bypassed. Neither of the two reference datasets used in this Chapter were normalised, so as to create one model per tissue type that could be applied across all datasets used to assess Cetygo, rather than a new model for each dataset the model was to be applied to (as in *minfi* and *wateRmelon*), which might make resulting Cetygo less comparable.

3.10.1.1 Generating Model 6CellBlood

Model 6CellBlood is the blood deconvolution model utilised in this Chapter for the testing of Cetygo. It was generated using data from **Dataset Reinius**, the blood reference data used for blood deconvolution within *minfi* (been previously validated for use in blood deconvolution in Reinius et al., 2012). The dataset contains 36 DNA methylation profiles across six individuals and six purified cell populations: B cells, CD4$^+$ T cells, CD8$^+$ T cells, granulocytes, monocytes and NKs. The individuals were all male, Swedish, and middle aged (mean age 38, SD = 13.6). The populations were purified from whole blood using MACS (see Section 3.1.2 for details on MACS purification) and DNA methylation was quantified using the HM450 array. The data is used in raw betas format, i.e. unnormalised.

To allow for testing Cetygo comparisons using simulated data, purified reference samples were required for testing. Data is commonly split 70/30 or 50/50 to make up training and testing data, however, since **Dataset Reinius** only contained 6 individuals (with each cell type measured once per individual) those ratios may be detrimental to the deconvolution model generated. To that end, the optimum number of individuals in the training data was calculated: Models were generated using *pickCompProbes()* and a subset of the reference data, with 6 models generated using 5 of the 6 individuals in the reference dataset, 15 models generated using 4 individuals, 20 generated using 3 individuals and 15 generated using 2 individuals, resulting in a total of 56 models. Bulk data was simulated for each model from individuals not included in the training data, with simulated samples containing every possible proportion of each cell type in steps of 0.1 that sum to 1, resulting in 2823 samples (see Section 3.4.1.3 for full details on data simulation). Each model was applied to its respective simulated dataset using *projectCellTypes()* and RMSE between the predicted and simulated cell type proportions was calculated across samples. Results were plotted by the number of individuals included in the training data (**Figure 3.33**). The RMSE for 2, 3, and 4 individuals were comparable, with the lowest RMSE in the 6 models generated using 5 of the 6 individuals available in the reference data, and so samples from five individuals were randomly selected to make up the training data for **Model 6CellBlood**.

To assess that the random assignment of training and testing across **Dataset Reinius** was not biased in some way, PCA was carried out within each cell type and across all samples. **Figure 3.34** contains PC 1 and 2 of each cell type across the top 1000 most variable DNA methylation sites (calculated using SD), with point fill signifying train test status. No strong clusters are seen within any of the plots suggesting that the division is not likely to be biased. **Figure 3.35**, which contains the first two PCs across all samples within **Dataset Reinius**, coloured by cell type, further confirms that the randomisation has been sufficient, with all testing samples being contained neatly within their cell type specific clusters.

Prior to model generation, DNA methylation sites were subset to contain those in both HM450 and EPIC array platforms. Models have previously been found to be applicable across either data type regardless of which the model was generated using DNA methylation sites present in both arrays (Fortin et al., 2014). The sites were also filtered for the four datasets which had undergone QC (the steps for which are described in Section 3.10.2): **Dataset EXTEND**, **Dataset Understanding Society**, **Dataset IoP**, and **Dataset EUGEI**, as QC can result in the removal of sites of low quality.

Using *pickCompProbes()*, **Model 6CellBlood** was generated. The model contains 600 DNA methylation sites (using the default of 50 sites selected in each direction per cell type) which can be seen in **Figure 3.36**. The chromosomal locations of the sites included can be seen in **Table 3.9**.

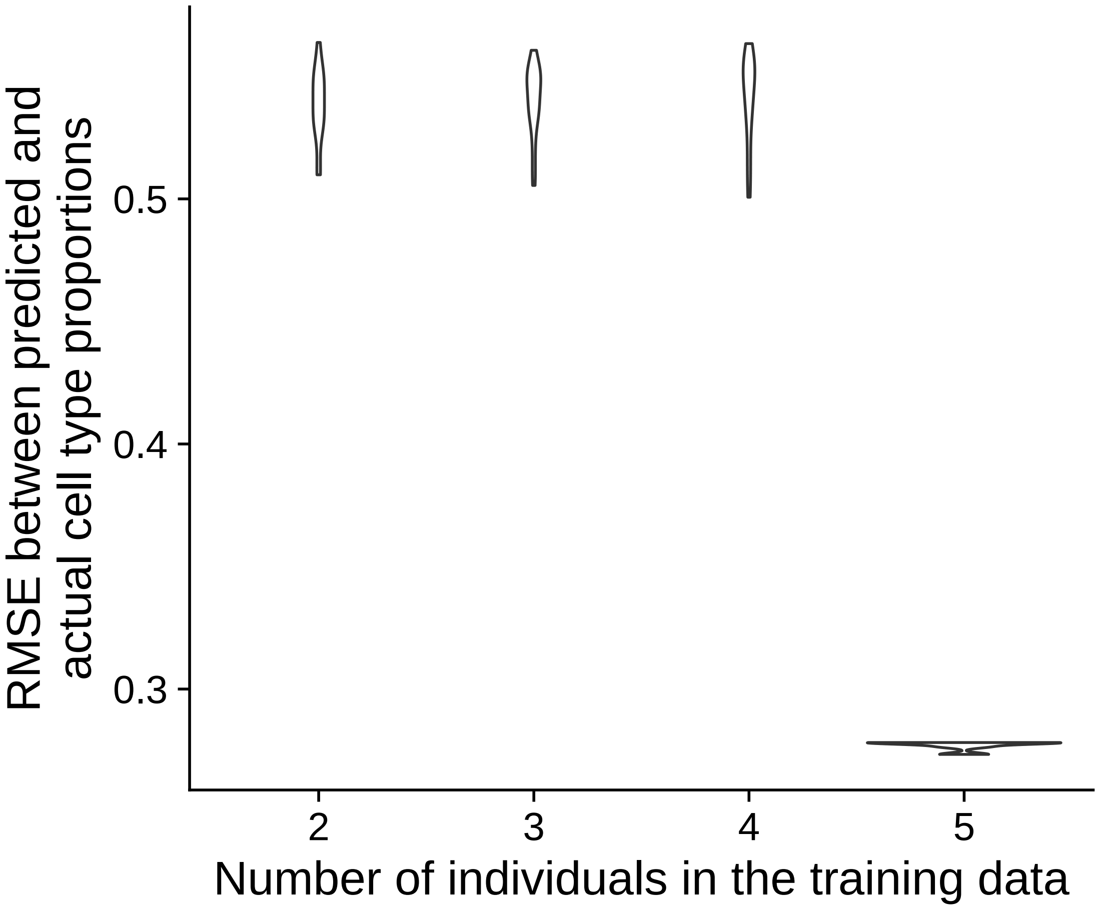


Figure 3.33: **A violin plot of the RMSE of predicted and simulated cell type proportions calculated using deconvolution models generated with subsets of reference data Dataset Reinius.** A total of 56 models were applied, using each possible combination of 2, 3, 4, and 5 individuals within **Dataset Reinius**. 2823 samples were simulated for each model from the data not included in the training data using the method described in Section 3.4.1.3.

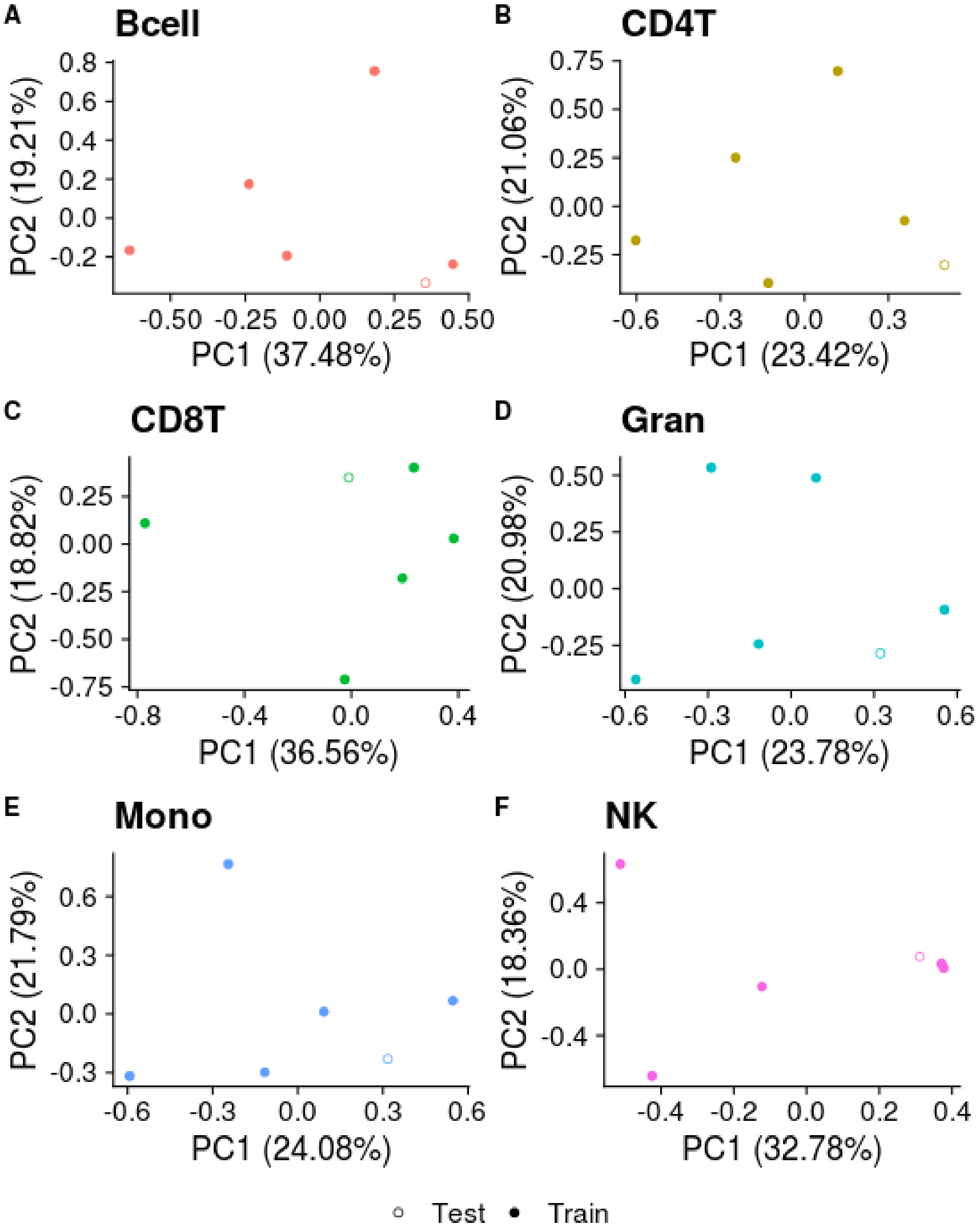


Figure 3.34: **A scatter plot of the first two principal components across each cell type within Dataset Reinius.** PCA was carried out on the top 1000 most variable DNA methylation sites according to SD. Plots show A) B cells, B) CD4^{+} T cells, C) CD8^{+} T cells, D) granulocytes, E) monocytes, and F) NKs, with shape corresponding to train or test group status.

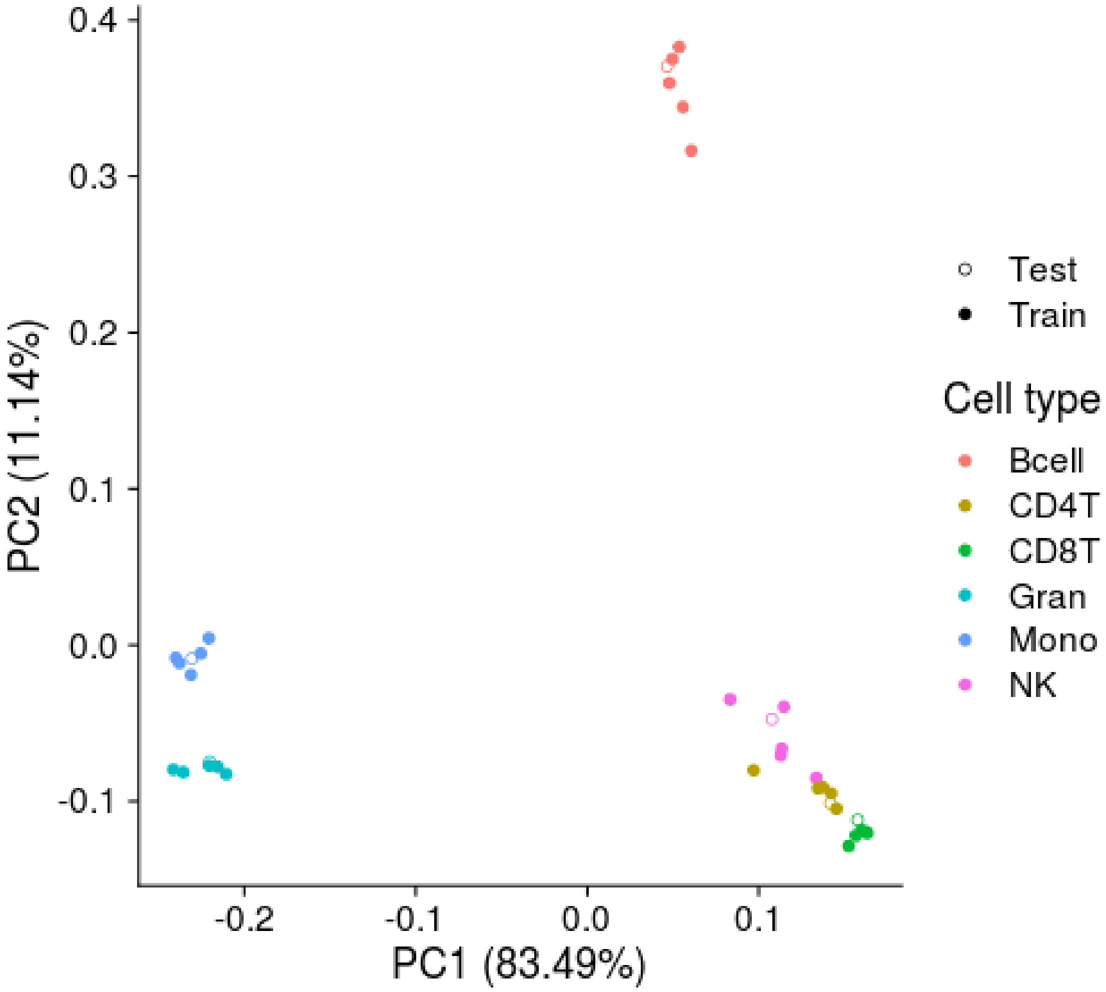


Figure 3.35: **A scatter plot of the first two principal components across Dataset Reinius.** PCA was carried out on the top 1000 most variable DNA methylation sites according to SD. Shapes are coloured by cell type: B cells (salmon), CD4^{+} T cells (sand), CD8^{+} T cells (green), granulocytes (teal), monocytes (purple), and NKs (pink), with shape corresponding to train or test group status.

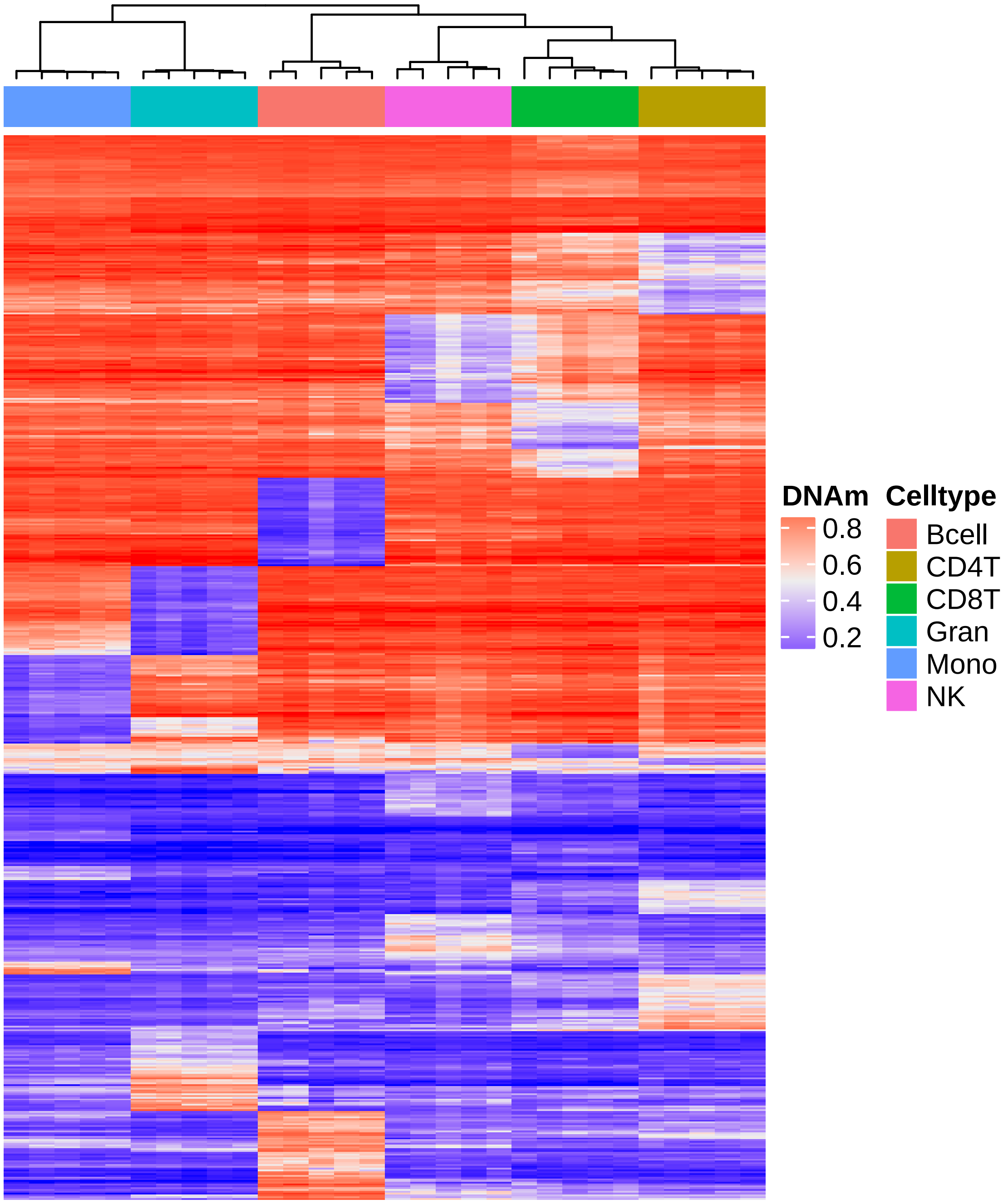


Figure 3.36: **A heatmap of the predictive DNA methylation sites included in Model 6CellBlood.** 600 CpGs are visualised across the 30 training samples of **Dataset Reinius**, clustering by cell type.

Table 3.9: **A table showing the chromosomal locations of the DNA methylation sites included in Model 6CellBlood and Model 2CellPFC.**

Chr	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**X**	**Y**
Reinius	53	60	39	17	18	43	32	28	10	23	47	42	15	21	15	32	33	5	30	8	5	16	8	0
Guintivano	13	7	3	6	7	5	5	4	0	4	10	7	4	6	1	3	6	0	4	3	0	2	0	0

3.10.1.2 Generating Model 2CellPFC

Model 2CellPFC is the PFC deconvolution model utilised in this Chapter to test Cetygo. The reference data utilised to generate the model is **Dataset Guintivano**, which has been previously validated as a reference dataset used for cortex deconvolution within *minfi* in conjunction with the Houseman algorithm and across CETS (Guintivano, Aryee and Kaminsky, 2013) (see Section 1.4 and **Table 1.1** for available deconvolution algorithms). The dataset contains 58 DNA methylation profiles across 29 individuals and two purified nuclear populations: NeuN+ and NeuN-, which represent neuronal and non-neuronal cell populations, respectively, and were sorted using FANS (Section 4.1.1.1). The mean age of individuals was 32.1 ± 15.9, the ratio of males to females was 14:15, and the ratio of African to Caucasian individuals was 6:23. DNA methylation was quantified using the HM450 array and is used in raw betas format, that is, the data was unnormalised.

To allow for Cetygo testing using simulated data, **Dataset Guintivano** was divided into training and testing, using a 70:30 split. Given that individuals within the dataset vary in ethnicity and sex, samples were stratified across both to minimise the potential bias in training data, resulting in 30 training samples, with 24:6 ratio of Caucasian:African samples, and 15:15 male:female samples (when the sex mislabeled sample is correctly assigned to the male group (**Figure 3.37**)), with age 34.9 ± 17.3, and testing data with a 22:6 ratio of Caucasian:African samples and 12:16 male:female samples, with age 29.1 ± 13.9.

To assess that the random assignment of training and testing samples across **Dataset Guintivano** was not biased in some way, PCA was carried out within each cell type and across all samples. **Figure 3.37** contains PC 1 and 2 of each cell type across the top 1000 most variable DNA methylation sites (calculated using SD), with the colour of the point signifying train test status. Clusters seen across PC 1 dividing samples by sex (with one misannotated individual) and PC2 dividing samples by ethnicity, both of which have been stratified for. Train test assignment is distributed evenly across the clusters.

Prior to model generation, DNA methylation sites were subset to those present on both HM450 and EPIC array platforms. Using *pickCompProbes()*, **Model 2CellPFC** was generated. The model contains 100 DNA methylation sites (using the default of 50

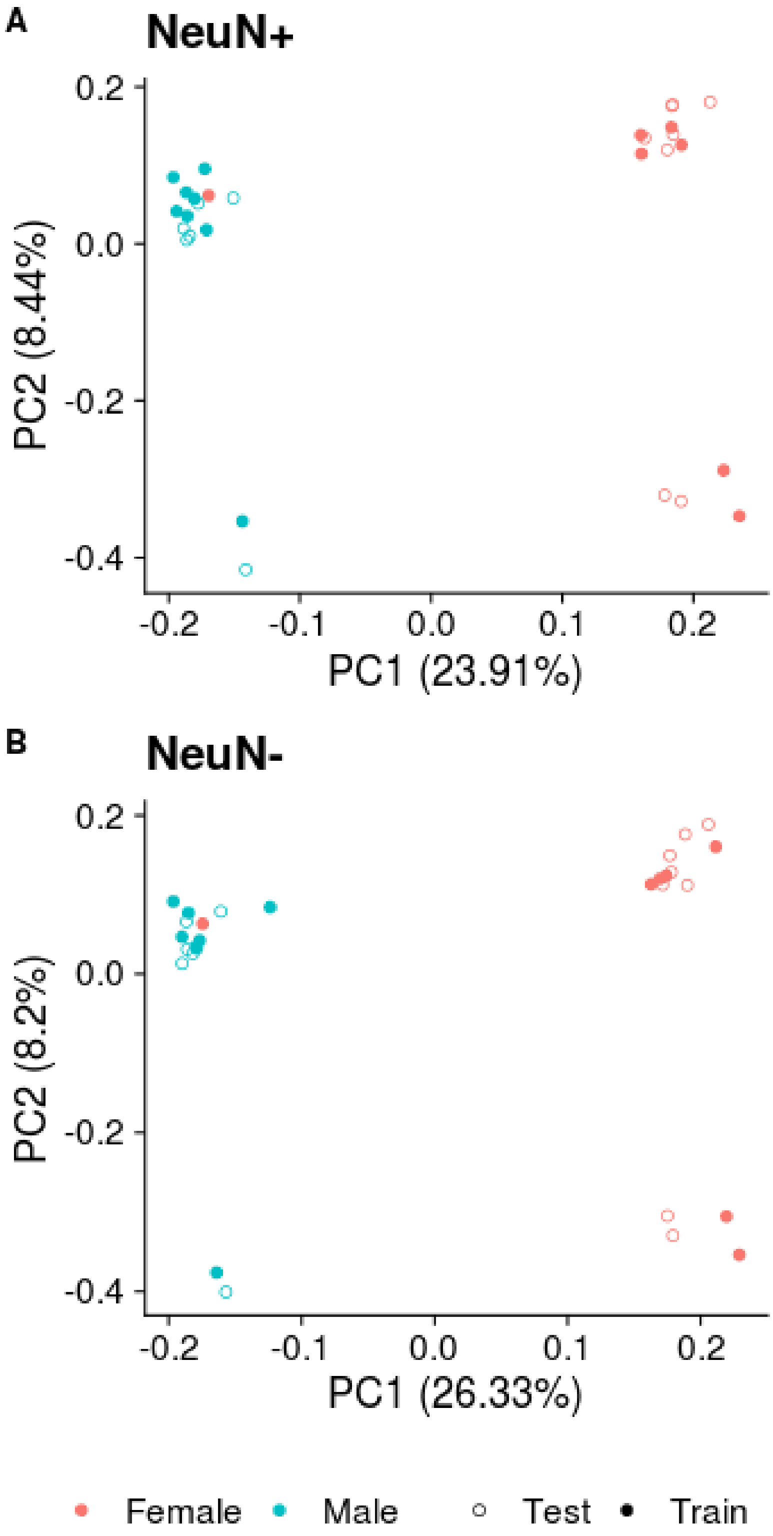


Figure 3.37: **A scatter plot of the first two principal components across Dataset Guintivano.** PCA was carried out on the top 1000 most variable DNA methylation sites according to SD. Shapes are coloured by sex: female (pink) and male (teal), with shape corresponding to train or test group status across A) NeuN+ samples, B) NeuN- samples.

sites selected in each direction per cell type, which with two cell types, are each others inverse) which can be seen in **Figure 3.38**. The chromosomal locations of the sites included can be seen in **Table 3.9**.

3.10.2 Quality control of Datasets EXTEND, Understanding Society, IoP, and EUGEI

Datasets EXTEND, **Understanding Society**, **IoP**, and **EUGEI** all underwent QC prior to use in this Chapter. The following steps were applied by Gemma Shireby across **Datasets EXTEND**, **Understanding Society** and by Eilis Hannon across **Datasets IoP**, and **EUGEI** using the R packages *wateRmelon* (Pidsley et al., 2013) and *bigmelon* (Gorrie-Stone et al., 2019):

1. Checking signal intensities of methylated and unmethylated samples and excluding those that performed poorly
2. Excluding samples with a bisulfite conversion rate of <80%
3. Ensuring that fully methylated control samples were in the correct location where applicable
4. Confirming reported sex using multidimensional scaling of CpG sites on the X and Y chromosomes separately
5. Checking for sample mismatches and duplicates, and ensuring correct genetic identity using the 59 single nucleotide polymorphisms (SNPs) on the Illumina Infinium Methylation EPIC BeadChip (EPIC) or HM450 array (depending on the platform used)
6. Exclude samples with >1% of probes with a detection P value >0.05 and probes with >1% of samples with detection P value >0.05 using *pfilter()* function in *wateRmelon*
7. Exclude samples that are dramatically altered by normalisation (compared to the raw data)
8. Remove outliers with *outlyx()*, a data-driven outlier detection tool that uses

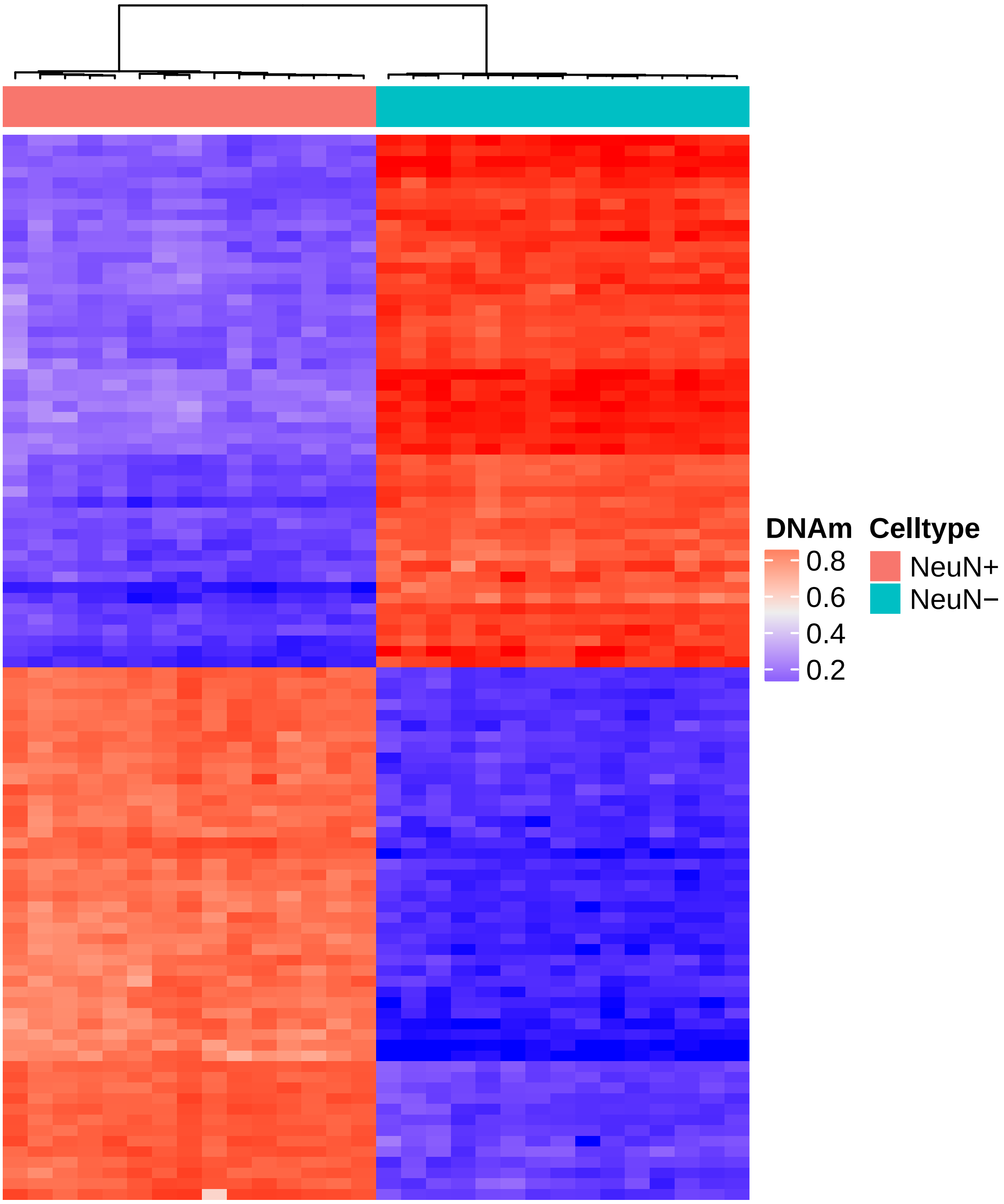


Figure 3.38: **A heatmap of the predictive DNA methylation sites included in Model 2CellPFC.** 100 CpGs are visualised across the 30 training samples of **Dataset Guintivano**, clustering by cell type.

dimensional reduction techniques to identify outliers according to two separate tests

9. Remove of cross-hybridising and SNP probes (Chen et al., 2013)
10. Normalise the data using the *dasen()* function in *wateRmelon* and *bigmelon*

3.10.3 Determining ethnicity of Dataset IoP and Dataset EUGEI

Genotype data from **Dataset IoP** and **Dataset EUGEI** was merged with genotype data from the 1000 Genomes Project (Pennisi, 2010) and PCA used to ascertain ethnicity (Eupopean or non-European).

3.10.4 Data analysis

All analysis was carried out in R version 3.5.2. The packages used can be seen in **Table 6.1**.

4. Developing a DNA methylation reference based deconvolution model to quantify the relative abundance of three distinct neural cell types in the human cortex from bulk tissue DNA methylation data

4.1 Introduction

Post-mortem brain tissue is commonly utilised in epigenetic studies of neuropsychiatic and neurodegenerative disorders (Lunnon et al., 2014; Smith et al., 2021; Policicchio et al., 2020b; Mill et al., 2008; Numata et al., 2014; Wockner et al., 2014; Pidsley et al., 2014), with the aim of identifying differential DNA methylation levels associated with the onset of disease or neuropathology. For pragmatic reasons, most existing studies have utilised "bulk" brain tissue, which is comprised of a number of different cell types (the main brain cell types are described in Section 3.1.2, **Figure 3.3** and **Table 3.3**), a characteristic referred to as cellular heterogeneity. Given the cell type specific nature of DNA methylation profiles (see Section 1.2.4), a shift in the composition of these cell types associated with the trait of interest (or the activation of genes associated with a specific cell type) will induce differences in the DNA methylation profile derived from bulk brain tissue that can lead to false positive (or negative) associations between measured DNA methylation and the disease/trait being studied.

This is potentially highly problematic for research into neuropsychiatric and neurodegenerative disorders, many of which are characterised by a shift in cellular proportions; for example, neuronal cell loss and an elevation in levels of active microglial cells is associated with Alzheimer's disease (AD) progression (Prinz and Priller, 2017; Matigian et al., 2007). Other examples include the reduced hippocampal volume observed in females associated with depression (Nifosì et al., 2010), the reduction in neuronal (but not oligodendroglial) density in cortical layers 1 and 5 associated with bipolar and depression (Cotter, Hudson and Landau, 2005), and the reduction of oligodendroglial cells observed in schizophrenia (SZ), bipolar, and depression patients (Uranova et al., 2004). In these cases, differences in DNA methylation observed between patients and controls may simply reflect the underlying changes in cell proportion in the piece of brain tissue being studied.

As such, adjusting for cellular composition is essential in analyses of bulk post-mortem brain tissue. Quantifying cellular composition empirically is not always feasible (see Section 1.3.3). As an alternative, computational deconvolution algorithms are used to provide estimates of specific cell proportions from DNA methylation data generated on bulk tissue.

4.1.1 Cellular deconvolution of brain derived DNA methylation profiles

The standard method for estimating the cellular composition of cortical samples from bulk DNA methylation data uses existing reference data comprised of the DNA methylation profiles of two purified neural cell types (described in Section 4.1.1.2) and a reference based deconvolution algorithm. As described in Section 3.1.1, reference based deconvolution algorithms rely on the following relationship:

$$\tilde{M}_{BULK} = \sum_{i=1}^{n} \hat{P}_i \tilde{M}_{CT_i}$$

where:

- $\tilde{M}_{BULK}$ is a matrix with one column containing the genome wide DNA methylation values for a bulk tissue sample, each row of which contains the DNA methylation value at the subset of k DNA methylation sites selected to distinguish the cell types within the model
- i is the index for the purified cell populations within the reference data, where $i \in [1, n]$ and n is the total number of unique cell types profiled

- $\hat{P}_i$ is ith value of vector $\underline{P}$ containing cell proportions for the n cell types
- $\tilde{M}_{CT_i}$ is the ith column within matrix $\tilde{M}_{CT}$, containing the DNA methylation values for cell type i at the subset of k DNA methylation sites selected to distinguish the cell types within the model

to be solved for $\underline{\hat{P}}$, with the constraint that the cell type estimated cellular proportions should sum to close to one, i.e. $\sum_{i=1}^{n} \hat{P}_i \leq 1$, and proportions each exceed one.

Generating the model requires two main stages (as described in Section 3.1.1 and 3.10.1). First, acquiring the reference DNA methylation dataset, $\tilde{M}_{CT}$, that contains the predominant cell types within the tissue to be deconvoluted (in this case brain tissue), and second, to use a deconvolution algorithm to generate a model by selecting DNA methylation sites from the reference data that can distinguish each cell type, which can subsequently be used to estimate cellular composition, $\underline{\hat{P}}$, from a bulk (i.e. cellularly heterogeneous) tissue sample.

4.1.1.1 Nuclear purification of brain tissue using fluorescent activated nuclear sorting (FANS)

As stated above, reference based deconvolution requires a reference dataset, $\tilde{M}_{CT}$, comprising the DNA methylation profiles of purified cell types within brain tissue. One common method to purify neural cell populations for DNA methylation profiling is fluorescence-activated nuclear sorting (FANS). FANS is a specialised flow cytometry method (Ibrahim and Van Den Engh, 2007) in which a heterogeneous mixture of cellular nuclei are fluorescently tagged and sorted based on light scattering and fluorescent emission detection (Adan et al., 2017) (**Figure 4.1**). The sorting of post-mortem brain tissue generally purifies nuclei rather than cells because the method requires cells to be in single cell suspension, which is not easily applicable to post-mortem brain tissue (depending on the method of tissue preservation) (Martin et al., 2017).

Several cell type specific antibodies exist for the fluorescent immunotagging of nuclei from different brain specific cellular populations (some of which have been listed in **Table 4.1**). For the purification of neuronal populations, an anti-NeuN antibody (Matevossian

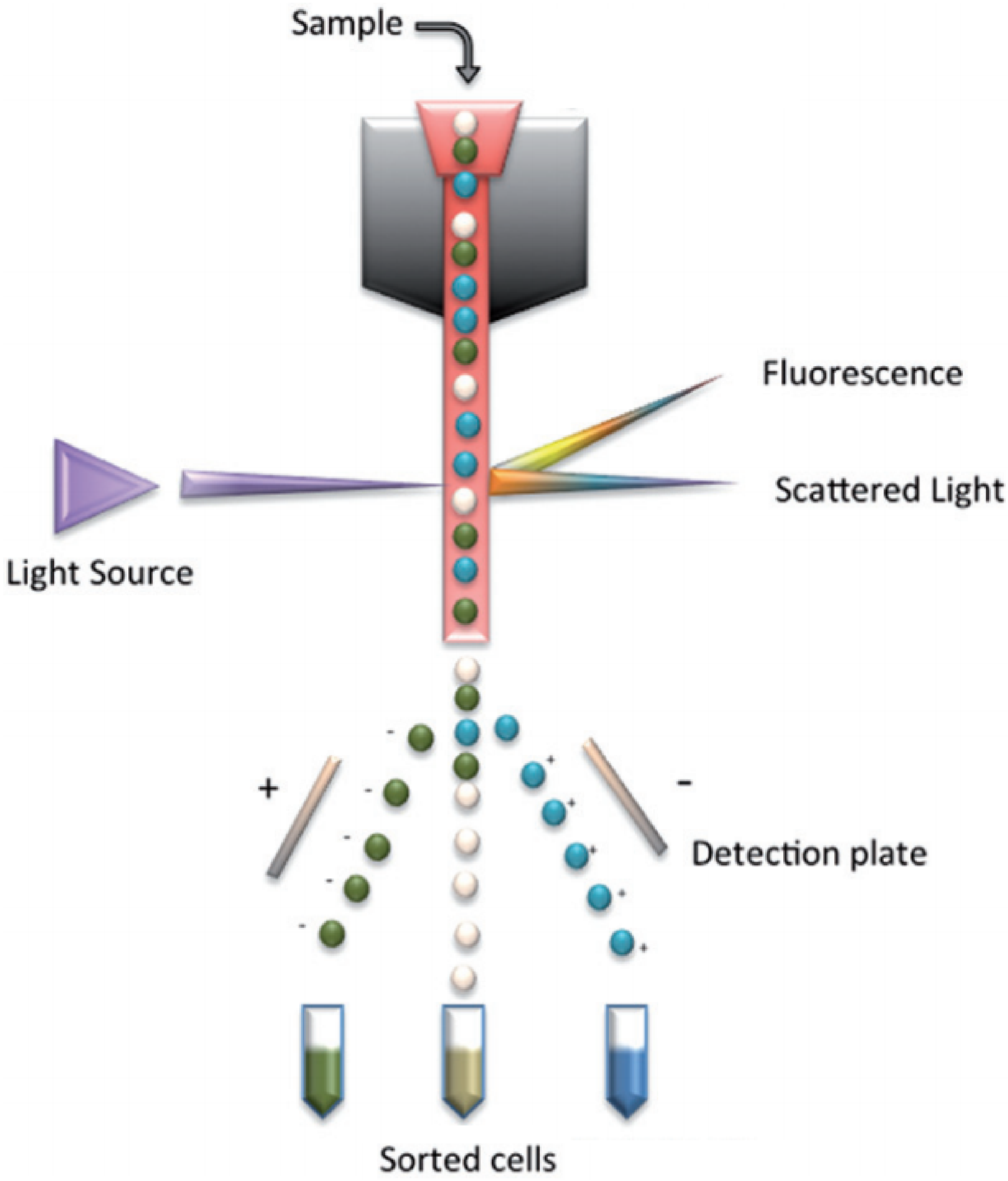


Figure 4.1: A schematic representation of sorting cells by droplet deflection, taken from Adan et al., 2017.

and Akbarian, 2008) (subsequently referred to as NeuN staining) is commonly utilised (Guintivano, Aryee and Kaminsky, 2013; Gasparoni et al., 2018; Tulloch et al., 2018; Bundo et al., 2020), purifying populations into NeuN+ (neuronal) and NeuN- (glial) populations. For the purification of nuclei from multiple cell types across brain tissue, additional markers can be used in combination, for example, the protocol by Policicchio et al., 2020a describes the use of NeuN, SOX10 and IRF8 to purify tissue into four distinct populations: NeuN+ (neuronal), NeuN-/SOX10+ (oligodendrocytes), NeuN-/SOX10-/IRF8+ (microglial), and NeuN-/SOX10-/IRF8- (other, including astrocytes) (Policicchio et al., 2020a). Similarly, Kozlenkov et al., 2015 utilised multiple markers to further purify the NeuN+ population, resulting in three distinct groups: NeuN+/SOX6+ (GABAergic neurons), NeuN+/SOX6- (glutamatergic neurons), NeuN- (glial) (Kozlenkov et al., 2015). Mussa et al., 2021 utilised three markers, purifying tissue into NeuN+ (neuronal), PAX6+/NeuN- (astrocytes),and OLIG2+/NeuN- (oligodendrocytes) (Mussa et al., 2021). Mendizabal et al., 2019 utilised two stains, NeuN+ (neuronal), and OLIG2+ (oligodendrocytes) to purify brain tissue into two populations (Mendizabal et al., 2019).

Some studies utilise the negatively selected nuclei as well as the positive, which will capture all other cell types, for example, the NeuN- population should consist of non-neuronal populations. These negatively selected populations will contain residual heterogeneity and as such it may be beneficial, when used for reference data, for brain tissue to be purified further using additional cell type specific stains. It is not always feasible, however, to apply multiple stains across brain tissue and obtain a high enough number of nuclei for subsequent DNA methylation profiling, especially if a cell type has low abundance within the tissue. This can be seen in **Dataset BDR purified** (used previously in Section 3.8.1), which was profiled using the aforementioned protocol by Policicchio et al., 2020a, utilising NeuN, SOX10, and IRF8 stains. However, of the 28 individuals from whom tissue was purified, only 3 DNA methylation profiles of NeuN-/Sox10-/IRF8+, and 2 of NeuN-/Sox10-/IRF8- were obtained due to the low yield of DNA from these specific populations. This may be due to the possibility that microglia make up a very small proportion of the sample being purified. This demonstrates that generating a reference dataset containing the complete set of purified brain cell types can be challenging and most likely proportional to the cellular abundance of the cell types to be included. To generate a deconvolution model, a sufficient number of samples from each cell type are required so as to identify DNA methylation sites with significantly differential DNA methylation between cell types. Given that DNA methylation differences between cell types are large and stable, the number of samples needed for each reference dataset is relatively small (e.g. a well-utilised blood cell reference dataset contains only six samples per cell type (Reinius et al., 2012)).

4.1.1.2 Current standard for reference based cellular deconvolution in PFC

The brain is one of the rarer non-blood tissues for which a reference dataset for deconvolution exists. The majority of brain DNA methylation studies perform reference based cellular deconvolution using data generated by Guintivano, Aryee and Kaminsky, 2013 (referred to here as **Dataset Guintivano** (see **Table 6.3**)). Briefly, it contains 58 samples obtained by FANS sorting post-mortem prefrontal cortex (PFC) tissue from 29 individuals into NeuN+ (neuronal) and NeuN- (glial) samples. DNA methylation profiling was then carried out using the Human Methylation 450K BeadChip (HM450) array.

To generate a deconvolution model to estimate cellular composition of bulk PFC tissue, one of two deconvolution algorithms is commonly used:

- **cell epigenotype specific (CETS)** - Guintivano, Aryee and Kaminsky, 2013, who generated **Dataset Guintivano**, also developed a novel deconvolution algorithm CETS. The algorithm selects the top 10,000 differentially methylated DNA methylation sites between the NeuN+ and NeuN- samples and utilises a linear slope model to predict the cellular composition of independent bulk tissue samples.
- **Houseman's algorithm** - Houseman's algorithm can be applied through commonly used array quality control (QC) packages including *minfi* and *wateRmelon* (Houseman et al., 2012; Jaffe and Irizarry, 2014). The default version of the algorithm used, described in Section 3.10.1, selects the 100 most significant DNA methylation sites that distinguish NeuN+ and NeuN- samples, with 50 sites where NeuN+ samples are hypermethylated, and 50 where they are hypomethylated in comparison to NeuN- samples. The algorithm uses constrained projection (CP)quadratic

Table 4.1: **A table summarising nuclear antibodies used for cell type specific fluorescent immunotagging.** Cell type specific proteins are targeted for each cell using an anti-protein antibody. Here the name of the protein, used as shorthand for the name of the anti-protein antibody, is listed along with the cell type that said protein is specific to in the brain.

Protein	Cell type purified
NeuN	neurons
SOX6	GABAergic neurons
SOX10	oligodendrocytes
OLIG2	oligodendrocytes
IRF8	microglia
PAX6	astrocytes

programming (QP) to predict the composition of independent samples.

4.1.1.3 Limitations of the current PFC deconvolution methodologies

Dataset Guintivano is currently the predominant reference dataset utilised for the deconvolution of different cell types in the human cortex. The primary limitation to the dataset is that it can only distinguish two cell types, NeuN+ and NeuN-, both of which will be comprised of sub cell types (see Section 3.1.2.2. The NeuN- population, which is comprised of non-neuronal cell types, will be comprised of microglia, oligodendrocytes and astrocytes. DNA methylation profiles of each of these glial cell types will differ (as observed using single-cell DNA methylation profiling in the mouse brain (Liu et al., 2021)), and as such, the DNA methylation profile of the NeuN- samples are arguably still a bulk tissue. Substantial variation in the proportions of brain cell types has been observed both within and between individuals (Rizzardi et al., 2019a). If the proportions of glial subtypes in a sample to be deconvoluted differs from the NeuN- samples within reference data, the reference data may not be applicable to said samples, as the assumption of $\tilde{M}_{BULK} = \sum_{i=1}^{n} \hat{P}_i \tilde{M}_{CT_i}$ (equation 4.1.1) will not hold. As such the resulting estimated cellular proportion may not be accurate.

Since the DNA methylation profile of a bulk sample is the weighted sum of the DNA methylation profiles of each cell type within the tissue, the larger the proportion of glial cells within the tissue, the more that the heterogeneity of the NeuN- fraction may matter. On average, across the central nervous system there have been found to be an equal number of neurons as glial cells (Azevedo et al., 2009; Bartheld, Bahney and Herculano-Houzel, 2016). As such, a reference dataset containing a further division of the NeuN- population for PFC reference data is likely to be beneficial to deconvolution accuracy. Furthermore, the generation of a model including more specific cell types may be especially important for identifying differential DNA methylation associated with a trait of interest that is anticipated to affect cells other than neurons.

4.1.1.4 Wider applicability of PFC deconvolution models

The generation of reference datasets for deconvolution is an expensive experimental process that requires specialised equipment, optimisation, and highly skilled researchers. Deconvolution models generated in a specific reference dataset may not be optimal or valid for all study designs due to demographic disparities between reference data and input data.

Disparities between reference and input data that could reduce the accuracy of brain deconvolution by a PFC reference based model include the following:

4.1.1.4.1 Brain regions

The brain consists of a number of discrete regions which are distinct in cognitive functions and behaviors, with differring DNA methylation profiles between regions (Ladd-Acosta et al., 2007; Hannon et al., 2015a; Liu et al., 2021). DNA methylation differences can occur within a cell type, as shown by Rizzardi et al., 2019a, who observed differential DNA methylation across FANS sorted NeuN+ between the dorsolateral prefrontal cortex, anterior cingulate cortex, hippocampus, and nucleus accumbens (Rizzardi et al., 2019a). As such, the disparity between DNA methylation in purified populations may mean that a reference dataset generated in one brain region may not be applicable to other regions.

4.1.1.4.2 Neurodegenerative disorders - e.g. Alzheimer's disease

AD is a neurodegenerative disorder, the characterisation of which includes neuronal loss and the accumulation of neurofibrillary tangles (Gómez-Isla et al., 1997). AD severity has been associated with differential DNA methylation in bulk brain samples across multiple brain regions (Smith et al., 2021). Using single cell RNA-seq methods, transcriptional changes were observed across all major brain cell types across AD pathology (Mathys et al., 2019), and as such, given that DNA methylation is cell type specific and is associated with gene expression, it is likely that DNA methylation differences would also be present within cell types. Alterations to cell type specific DNA methylation profiles due to disease may reduce the similarity between reference and sample cell type DNA methylation profiles, decreasing deconvolution accuracy.

4.1.1.4.3 Applicability across very young samples

There are dramatic changes in DNA methylation across fetal brain development (Spiers et al., 2015) and many cells will not yet be mature and not be characterised by the expression of key cell type specific marker genes. For example, not all neurons express NeuN in early development, with only deep neurons within the cortical plate being marked at 19–22 post conception weeks (pcw) (Sarnat, Nochlin and Born, 1998). To that end, a reference based deconvolution model generated using a reference dataset of adult individuals may not be applicable to use for the deconvolution of tissue from fetal or very young donors.

4.1.1.4.4 Applicability across neuronal iPSC and SH-SY5Y cell models

Cell models are commonly used to investigate direct effects of stimuli in specific cell types. If applicable, deconvolution could be utilised to assess the purity of cell models. Common brain cell lines include the SH-SY5Y cell line and neuronal induced pluripotent stem cells (iPSCs); SH-SY5Y cells are derived from a cancer cell line, and can be differentiated into dopamanergic neurons, given specific media conditions (Xie, Hu and Li, 2010). Following differentiation they become morphologically similar to primary neurons (Påhlman et al., 1984). They are commonly utilised in the research of Parkinson's disease (PD) and SZ (Xie, Hu and Li, 2010; Bray, Kapur and Price, 2012), as these disorders involve the dopamine pathways. iPSC cell lines can be derived from any cell from an individual or patient (Dolmetsch and Geschwind, 2011). Reprogramming factors are used to revert said cells to stem cells and then growth factors can be used to mature cells down specific lineages, including neuronal lineages.

Brain cell lines, SH-SY5Y and neuronal iPSCs may differ in DNA methylation profile as cell type specific epigenetic marks are not always fully erased (Kim et al., 2011). Furthermore, the usually short length of culture may mean that cells are biologically young (Steg et al., 2021) compared to the reference populations, which, as described above, may also result in differential DNA methylation and subsequent inaccurate deconvolution using an adult reference dataset.

4.1.2 Chapter aims

Reference based deconvolution can be utilised to estimate the cellular composition of heterogeneous tissues from DNA methylation profiles generated on bulk samples. These estimates can be used to adjust for cellular composition in DNA methylation studies performed on bulk tissue which might otherwise be confounded by cellular heterogeneity between individuals. The currently used reference based deconvolution method for PFC tissue predicts only two cellular proportions, neuronal and glial (Guintivano, Aryee and Kaminsky, 2013). The glial population in the reference data is, however, highly heterogeneous and is comprised primarily of microglia, oligodendrocytes and astrocytes. The proportions of glial subtypes differ from sample to sample, and as such a model that can further divide this heterogeneous population would be highly beneficial in better accounting for cellular heterogeneity in brain in DNA methylation studies.

The main objective of this Chapter was to generate and test a novel three cell type deconvolution model, entitled **Model 3CellPFC**, for PFC tissue which could predict the proportions of neurons, oligodendrocytes and other neural cells (primarily microglia and astrocytes) in an independent input dataset comprised of bulk tissue DNA methylation data.

The specific Chapter aims are to:

1. establish a three cell type reference based deconvolution model for DNA methylation samples using the DNA methylation profiles of FANS purified post-mortem PFC tissue
2. assess model accuracy using simulated data

3. compare the prediction accuracy and Cetygo (Chapter 3) of predictions between **Model 3CellPFC** to the two cell type model applied to the same datasets.
4. characterise the behaviour of **Model 3CellPFC** when estimating the cellular composition of purified samples more refined than the model training data.
5. investigate cell type bias in prediction accuracy in **Model 3CellPFC** using simulated data
6. assess the wider applicability of **Model 3CellPFC** to samples that differ from the reference datasets including:
    - non-PFC brain regions
    - samples across AD progression
    - fetal and very young postnatal samples
    - neuronal cell lines
7. characterise the cellular composition information gained between **Model 3CellPFC** and the two cell type deconvolution model across brain regions and AD progression

4.2 Model 3CellPFC: generating a three cell type deconvolution model for post-mortem PFC tissue

The dataset utilised for the generation of a novel three cell type reference based deconvolution model for PFC, **Model 3CellPFC**, is **Dataset CortexFANS**, comprised of 112 NeuN+ samples, 107 Sox10+ (NeuN-/Sox10+) samples, and 98 Double- (NeuN-/Sox10-) samples, purified using FANS (see **Figure 4.2**) by the Complex Disease Epigenomics Group at the University of Exeter. The dataset also contained 113 samples of bulk nuclei, referred to as "Total" from the same individuals, which were excluded from model training but utilised in testing. DNA methylation profiling was carried out using the Infinium Methylation EPIC BeadChip (EPIC) array. Prior to model generation, firstly, the quality of the reference data needed to be assured. Secondly, in order to be able to address the model validation aims set out in Section 4.1.2, the purified samples were assigned to either reference (i.e. training) or testing data, with only those in the reference group used to generate the model.

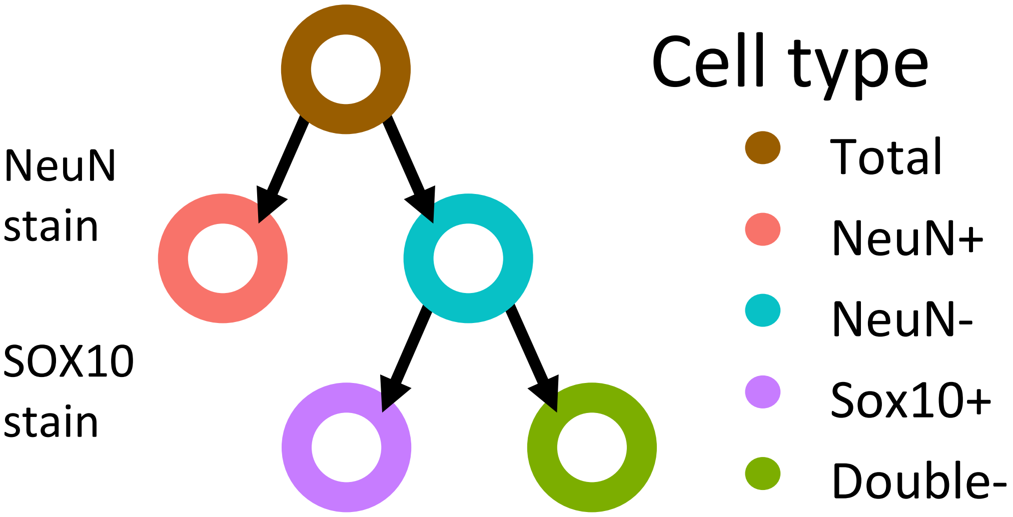


Figure 4.2: **A diagram of the cellular hierarchy in FANS sorting of PFC tissue using the protocol developed by Policicchio et al., 2020a using the antibodies to NeuN and Sox10.** NeuN+ populations are enriched for neuronal nuclei, and Sox10+ are enriched for oligodendrocyte nuclei.

4.2.0.1 Ensuring the quality of reference dataset Dataset CortexFANS

It is important for the DNA methylation profiles of the purified nuclei to be of high quality so as to be fully representative of the cell types within bulk samples to be deconvoluted, and therefore result in a more sensitive model.

A potential driver of low quality is inefficient sample purification, for example if the nuclei selected for a sample were more heterogeneous than intended, rather than being comprised of a singular population (in the case of positively selected populations).

DNA methylation patterns are highly cell type specific and as such, cell type should be the largest driver of variance across the dataset, allowing for the confirmation of cell type group assignment and purification.

A principal component analysis (PCA) was carried out across the top 1000 most variable DNA methylation sites, as ranked by standard deviation (SD), across the data set to assess cell type clustering (see **Figure 4.3A**). Three general cell type clusters could be observed (one per cell type within the dataset), with a subset of samples of each cell type not clustering to their group but instead being spread between the Sox10+ and NeuN- clusters along principal component (PC) 1. These samples were assumed to be poorly purified, and potentially comprised of a mix of cell type nuclei. As such, samples with -6<PC1<7 and PC2<0 were removed, along with two presumed misannotated samples, one NeuN+ and one Double-, that clustered in the Sox10+ population. A total of 29 samples were removed (see **Figure 4.3B**) from the final dataset.

4.2.0.2 Dividing into training and testing

For deconvolution model generation and validation, **Dataset CortexFANS** was to be divided into training and testing cohorts. The testing cohort allows for the verification of model accuracy in samples in which the cell type proportions are already known and can be utilised to generate simulated data of mixed cellular proportions for further testing.

Dataset CortexFANS was originally generated as part of an ongoing project aiming to analyse DNA methylation patterns between SZ patients and controls. Since potential cell type specific differential DNA methylation patterns have been previously observed between SZ patients and matched controls within post-mortem brain tissue

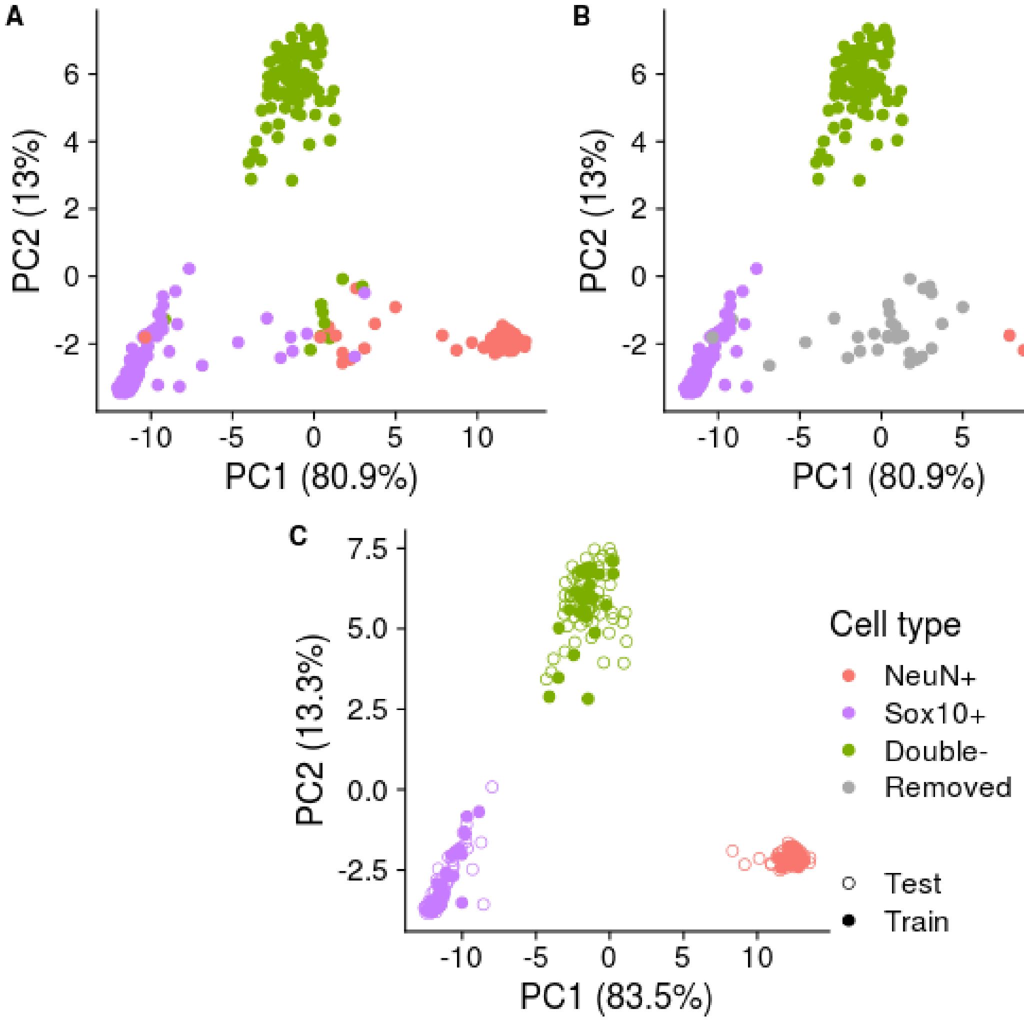


Figure 4.3: **Scatter plots of the first two principal components across Dataset CortexFANS.** PCA was carried out on the top 1000 most variable DNA methylation sites according to SD. The scatter plots plot the first two principal components containing A) all samples, labelled cell type: NeuN+ (salmon), Sox10+ (purple), and Double- (green); B) all samples, with samples not closely clustering to their cell type, highlighted in grey, and C) with grey samples from B removed, and PCs recalculated for the 1000 most variable DNA methylation sites across the remaining samples, with point shape representing membership of the training (filled circle) and testing (circle only) sub-datasets.

(Mendizabal et al., 2019; Pidsley et al., 2014), all 139 SZ samples were assigned to the testing data, as it was unknown what effect said differences might have on deconvolution accuracy. The remaining samples were randomly evenly assigned to training or testing, resulting in training data comprised of 29 NeuN+, 28 Sox10+ and 24 Double- samples, and a testing dataset containing 72 NeuN+, 70 Sox10+ and 65 Double- samples (the first two PCs of which can be seen in **Figure 4.3C**). The training data contained higher number of samples per cell type than used to generate **Model 6CellBlood** (n = 5 per cell type) and was therefore deemed sufficient in size, especially given the large magnitude of DNA methylation differences observed across different cell types.

4.2.0.3 Model generation

The **Dataset CortexFANS** training dataset was utilised for model generation. The R function *pickCompProbes()* (from the R package *minfi*) was applied to generate the model, which uses ANOVA to select DNA methylation sites that distinguish each cell type from all other cell types. The default number of sites selected was 50 per cell type per direction, i.e. 50 sites for which the DNA methylation is higher than in other cell types, and 50 in which it is lower. Default values were used here as they are most commonly used and would represent the most likely way a model would be generated using this reference dataset by other researchers. As such, **Model 3CellPFC** contained 300 DNA methylation sites, the DNA methylation profiles of which across all reference dataset samples can be seen in **Figure 4.4**. Cell type differences were distinct and the hierarchical cluster in the plot shows the samples clustered by cell type.

4.3 Model 2CellPFC: the baseline PFC deconvolution algorithm

In order to assess the potential utility of **Model 3CellPFC** to the wider field, it will be compared to the current standard PFC deconvolution method, which utilises **Dataset Guintivano**. Deconvolution models using **Dataset Guintivano** can be applied through CETS (Guintivano, Aryee and Kaminsky, 2013) or Houseman's algorithms (Houseman

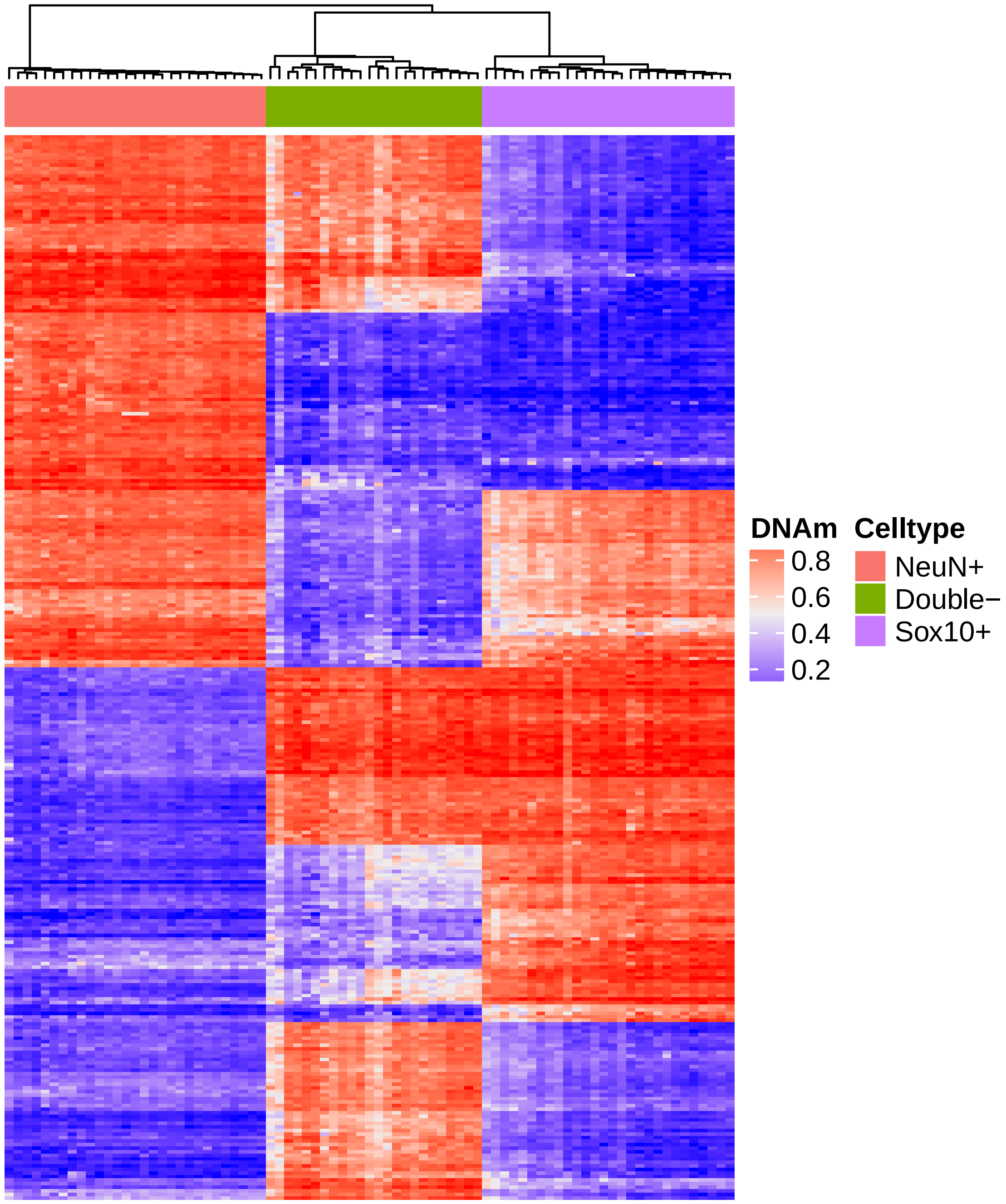


Figure 4.4: **A heatmap of the predictive DNA methylation sites included in Model 3CellPFC**. Each row represents a site, with a total of 300 sites in the model. Each column represents a sample in the training dataset. Purified cell type profiles are in three distinct groups. DNA methylation (DNAm) at a site is red when most methylated and blue when unmethylated.

et al., 2012; Jaffe and Irizarry, 2014). Here, to minimise the differences between models that may influence accuracy outside of the reference data used, **Dataset Guintivano** is divided into training (n=30, 15 NeuN+ and 15 NeuN-) and testing (n=28, 14 NeuN+ and 14 NeuN-) and Houseman's algorithm is applied through functions within the *minfi* R package (see Section 3.10.1.2 for the full description of model generation). The resulting model, entitled **Model 2CellPFC**, contains 100 DNA methylation sites for the prediction of NeuN+ and NeuN- populations in input samples.

4.4 Overview of the datasets utilised and tools to assess Model 3CellPFC

To compare the predictions and applicability of **Model 2CellPFC** and **Model 3CellPFC**, the Cetygo algorithm was utilised. Cetygo is a reference based deconvolution error metric, established in Chapter 3. In general, its use is comparative with lower values signifying higher quality prediction, although a soft threshold limit of 0.1 was established, below which deconvolution accuracy is deemed accurate.

In order to thoroughly test the performance of **Model 3CellPFC**, multiple brain DNA methylation datasets were used to address the aims set out in Section 4.1.2. This section provides a summary of the datasets used in this analysis. For each dataset, the raw unnormalised betas were used (defined in Section 1.3.4.2). DNA methylation is quantified in all datasets using bisulfite conversion without oxidisation, and so what is referred to as DNA methylation is technically the sum of 5-methylcytosine and 5-hydroxymethylation. An overview of all datasets used in this thesis can be found in **Table 6.3**, which includes the Gene Expression Omnibus (GEO) accession numbers for publicly available datasets, as well as stating which datasets were generated by the Complex Disease Epigenomics Group at the University of Exeter.

Dataset Guintivano testing data, a subset of **Dataset Guintivano** (Guintivano, Aryee and Kaminsky, 2013), generated for the deconvolution of PFC), contains 28 samples derived from post-mortem FANS PFC tissue stratified by NeuN staining: 14 NeuN+ samples, and 14 NeuN- samples. DNA methylation profiling was carried out using the HM450 array. The mean age of individuals was 29.1 ± 13.9 years, the ratio of males to females was 12:16, and the ratio of African to Caucasian individuals was 6:22.

Dataset CortexFANS testing data, a subset of **Dataset CortexFANS** (generated to investigate cell type specific DNA methylation patterns associated with SZ), contains 320 samples from post mortem PFC tissue. The testing dataset contains 207 FANS purified samples, obtained by purifing NeuN+ and SOX10+ nuclei: 72 NeuN+ samples, 70 Sox10+ (NeuN-/Sox10+) samples, and 65 Double- (NeuN-/Sox10-) samples. NeuN positively selects neurons, and SOX10 is a robust oligodendrocyte marker. The dataset also contained 113 "Total" nuclei samples representing bulk cortex from the same individuals. DNA methylation profiling was carried out using the EPIC array. The mean age of samples was 66.0 ± 15.5 years, the ratio of males to females was 142:94.

Dataset BDR purified contains 107 samples derived from PFC tissue, originating from the Brains for Dementia Research (BDR) cohort, the purpose of which is to better understand the relationship between dementia and gene regulation in the brain. Bulk cortex tissue was FANS sorted (using the same protocol as **Dataset CortexFANS**), utilising NeuN, SOX10 and IRF8, a microglial marker, using tissue from 28 individuals resulting in 27 NeuN+ samples, 28 Sox10+ samples, 21 Double- samples, 3 IRF8+ (NeuN-/Sox10-/IRF8+) samples, 2 Triple- (NeuN-/Sox10-/IRF8-) samples, and 26 Total (unsorted nuclear) samples. DNA methylation was profiled on the EPIC array. The mean age of individuals was 80.8 ± 9.16 years, the ratio of males to females was 46:61.

Dataset Pai, is a publicly available dataset generated to investigate SZ and bipolar disorder (Pai et al., 2019), comprising 100 FANS sorted NeuN+ PFC samples. The mean age of individuals was 47.6 ± 10.5 years, the ratio of males to females was 75:25. DNA methylation was profiled on the EPIC array.

Dataset BDR bulk, also originating from the BDR brain bank, contains 1304 samples, 671 of which were PFC tissue, and 613 were occipital lobe (OCC) from matched individuals (Shireby et al., 2020). Braak staging information, a measure of neurofibrillary tangle involvement signifying AD pathology, was available across samples, with 45 samples at stage 0, 129 samples at stage I, 246 at stage II, 163 at stage III, 128 at stage IV, 206 at stage V, and 325 at stage VI, the most severe Braak stage. The mean age of individuals was 83.5 ± 9.17 years, the ratio of males to females was 667:614. DNA methylation was profiled on the EPIC array.

Dataset Adult brain is a compilation of bulk brain datasets, containing 338 samples from 188 individuals and including multiple brain regions: 38 BA11, 26 BA25, 46 BA9, 64 cerebellum, 42 hippocampus, 98 striatum, and 24 thalamus samples. Samples originate from a number of DNA methylation studies into psychiatric disorders, and are made up of post mortem tissue from 173 control samples, 44 individuals with diagnosed depression, 108 individuals diagnosed with SZ, and 13 individuals diagnosed with multiple sclerosis (MS). The mean age of individuals was 48.9 ± 18 years, the ratio of males to females was 231:73. DNA methylation was profiled on the EPIC array.

Dataset Fetal was generated as part of an investigation into DNA methylation changes that occur over early development (as of yet, unpublished). It contains 114 samples, 11 of which are cerebellum, 103 are PFC.The average age was 23.5 pcw ± 46.pcw, with a range of 6 - 456 pcw. Three samples were post-natal, with ages 2, 3, and 8 years old. The ratio of male to female was 49:64. DNA methylation was profiled on the EPIC array.

Finally, two neuronal cell line datasets were utilised: **Dataset iPSC**, which contains 93 neuronal induced iPSCs samples, and **Dataset SH-SY5Y**, which contains 156 samples from the SH-SY5Y cell line, the DNA methylation of which were profiled on the EPIC array. Neuronal cell lines are commonly used within neuroscience research as an alternative to brain tissue (the pros and cons of which are described in Section 1.3.3). **Dataset iPSC** was generated to investigate exposure to epigenetic modulators, and **Dataset SH-SY5Y** was comprised of DNA methylation data from two studies, one investigating the effect of tetrahydrocannabinol (THC) exposure on the cell line, and the second to investigate neuronal differentiation (all findings are as of yet unpublished).

4.5 Assessing the accuracy of Model 3CellPFC in simulated data

To assess the general accuracy of **Model 3CellPFC**, the simulation framework laid out in Section 3.4 was utilised (equation 3.3). The cell type specific DNA methylation profiles ($\tilde{M}_{CT.TEST_i}$) (i.e. the testing data of **Dataset CortexFANS** (n = 236, of which 83 were NeuN+, 79 were Sox10+, and 74 were Double-)) were used to simulate bulk samples with known cellular proportions. 60 samples were simulated to represent the full spectrum of combinations, with each possible cellular proportion (P_i) for each cell type between 0 and 1 in steps of 0.1 such that the total sample proportion summed to 1.

Estimates of cellular composition were highly accurate across the simulated bulk profiles; root mean squared error (RMSE) between predicted and actual proportions was 0.0363 (across all cell types) (**Figure 4.5**). Accuracy of composition estimates is an important facet of a deconvolution algorithm. These results preliminarily suggest that **Model 3CellPFC** performs well.

4.6 Comparing accuracy between Model 2CellPFC and Model 3CellPFC

To be of use to the wider research community, the novel deconvolution model **Model 3CellPFC** would need to be of equal or improved predictive accuracy when compared to the current standard, **Model 2CellPFC**. As such, estimated proportions were compared across 'Total' samples (DNA methylation profiles of un-FANS sorted nuclei) from **Dataset CortexFANS** and **Dataset BDR purified** (n = 139, **Figure 4.6**).

The variability in predictions is higher across **Model 2CellPFC** than **Model 3CellPFC** (SD of predictions from **Model 2CellPFC** are 0.253 and 0.27 for NeuN+ and NeuN-, respectively; SD of **Model 3CellPFC** are 0.0968, 0.112, and 0.0613, for NeuN+, Sox10+ and Double-, respectively; **Figure 4.6**). Cetygo is significantly higher for **Model 2CellPFC** estimates than for **Model 3CellPFC** estimates (p = 6.75e-07, mean**Model 2CellPFC** = 0.0959, mean Cetygo from **Model 3CellPFC** = 0.0788).

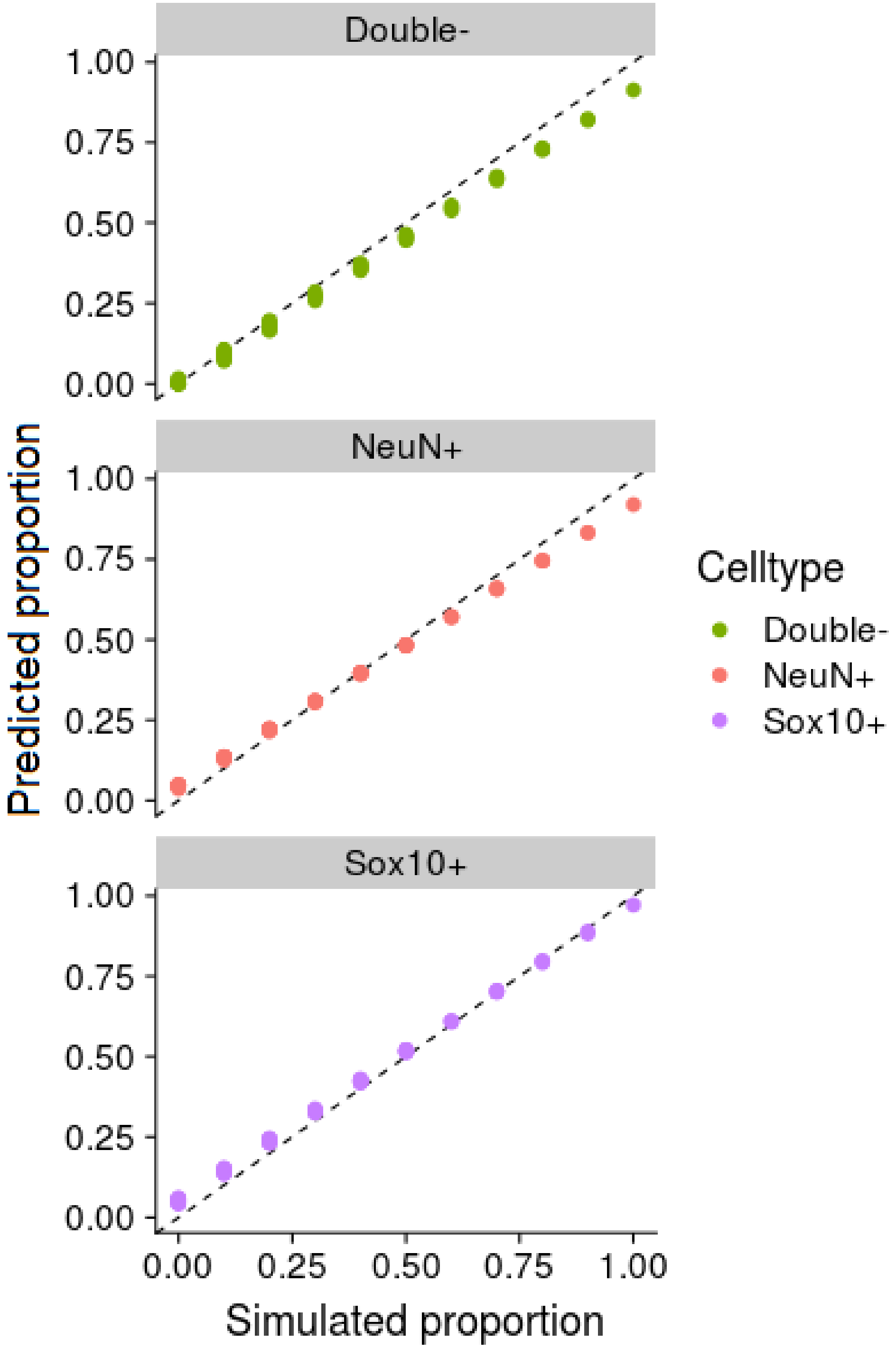


Figure 4.5: **Scatter plots of the actual and predicted proportions of simulated bulk data estimated by Model 3CellPFC.** Each plot contains the estimated proportion of one cell type across the 60 simulated samples, with Double- in green, NeuN+ in salmon, and Sox10+ in purple.

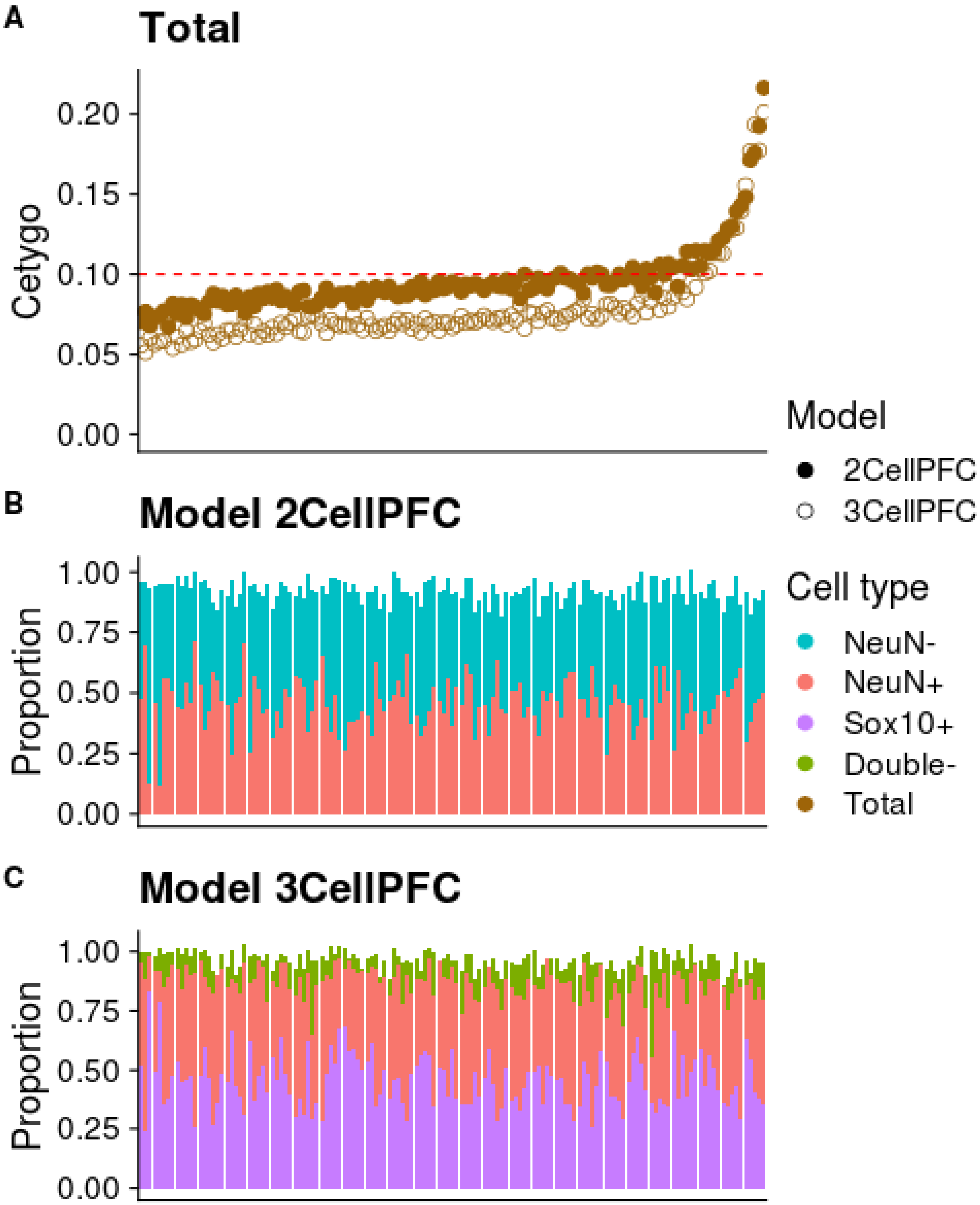


Figure 4.6: **A summary of Cetygo, and the predicted proportions estimated from Model 2CellPFC and 3CellPFC across Total samples in Dataset CortexFANS and Dataset BDR purified.** A) The Cetygo across samples. B) Stacked bar plot of the predicted proportion of cell types by **Model 2CellPFC**. C) Stacked bar plot of the predicted proportion of cell types by **Model 3CellPFC**. The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions.

Together these findings suggest that predictions made from bulk cortex DNA methylation data by **Model 3CellPFC** are more accurate and consistent than **Model 2CellPFC**.

Predictions for the Total samples were highly correlated between models, with a correlation 0.98 between proportion of NeuN+, and 0.991 between predicted proportions of NeuN- (from **Model 2CellPFC**) and the sum of Sox10+ and Double- (from **Model 3CellPFC**) (**Figure 4.7**). Of note, while predictions correlate highly, the predicted proportion of NeuN+ is systematically higher in **Model 2CellPFC** than **Model 3CellPFC** by 0.02 (mean predicted proportion of NeuN+ = 0.435 (sd = 0.253) and 0.415 (sd = 0.0968, respectively). Given the elevated Cetygo, this suggest that NeuN+ proportions are overestimated in **Model 2CellPFC**.

Figure 4.7 demonstrated that the NeuN+ predictions in **Model 2CellPFC** were systematically lower than in **Model 3CellPFC**, however, given that this was ascertained in bulk tissue, the true proportion of NeuN+ was not known. Here, estimated composition of NeuN+ was compared across NeuN+ samples, the NeuN+ samples from four datasets were utilised: testing data from **Dataset Guintivano** (n = 28) and **Dataset CortexFANS** (n = 72), and two independent datasets, **Dataset BDR purified** (n = 27), and **Dataset Pai** (n = 100).

The majority of NeuN+ samples are predicted as predominantly NeuN+ by both models, with 82.6% and 87.9% of samples predicted as >80% NeuN+ by **Model 3CellPFC** and **Model 2CellPFC** respectively (**Figure 4.8-4.11**). The predictions were highly correlated between models (Cor = 0.914, RMSE = 0.04 **Figure 4.12**), suggesting that NeuN+ predictions are equivalent between models in purified NeuN+ samples.

Of note, **Model 2CellPFC** had negative predicted estimates of NeuN- composition across many samples, which may suggest that predictions for purified samples using a model containing only two cell types may not always be biologically meaningful. In contrast, all predictions by **Model 3CellPFC** were positive.

The majority of NeuN+ samples had Cetygo less than 0.1 across both models, the soft threshold signifying that deconvolution was generally accurate (**Figure 4.13**, 213/224 (87.3%) NeuN+ samples by **Model 3CellPFC**, and 206/224 (84.4%) NeuN+ samples). This suggests that, while both models are predicting accurate compositions,

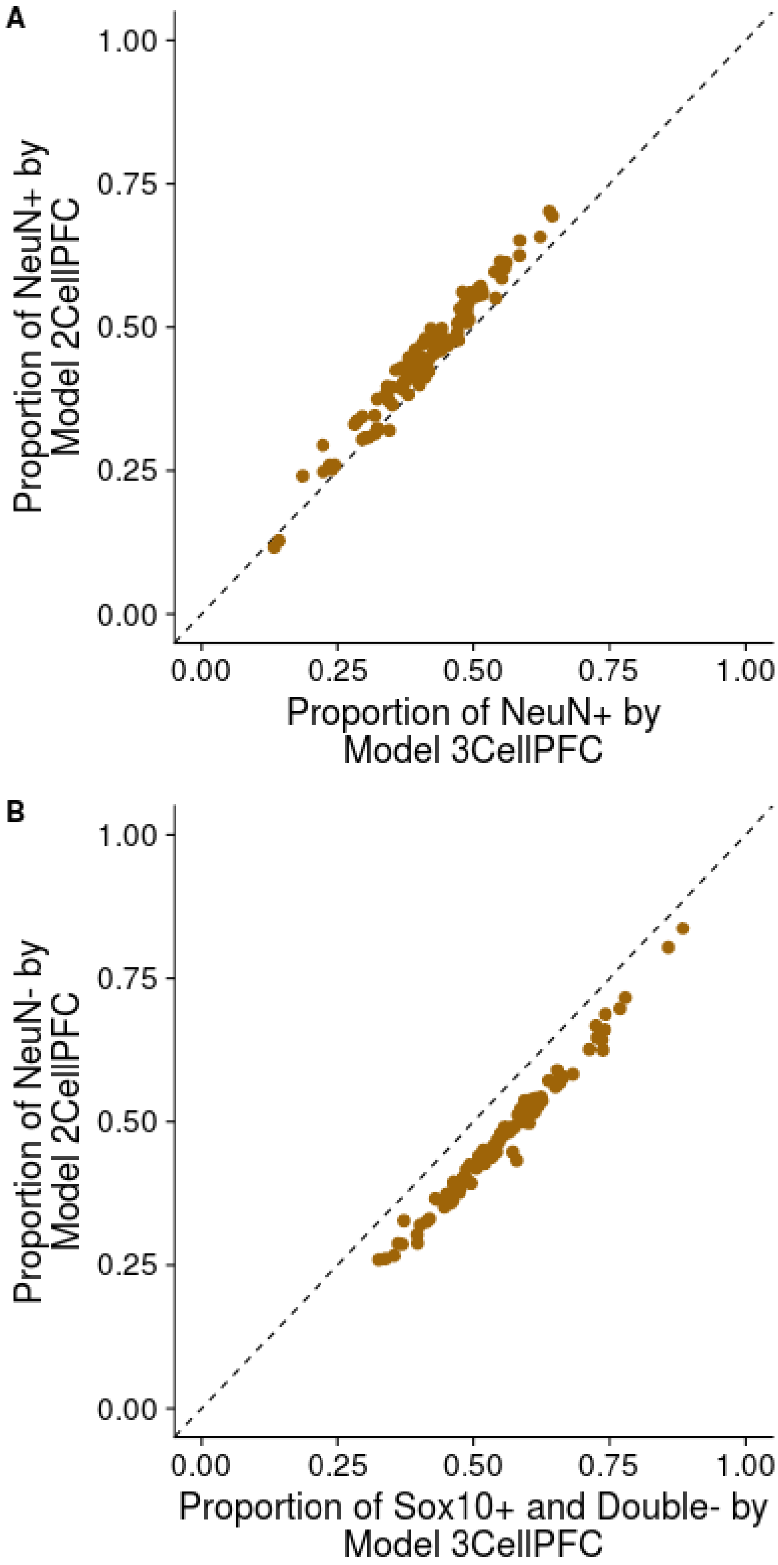


Figure 4.7: **A scatter plot of predicted proportions of Total samples in Dataset CortexFANS and Dataset BDR purified estimated by Model 2CellPFC and 3CellPFC by cell type.** Across the predicted proportions of A) NeuN+ (Cor = 0.98) and B) NeuN- and Sox10+ + Double- in **Model 2CellPFC** and **Model 3CellPFC**, respectively (Cor = 0.991). The black dashed line marks y = x.

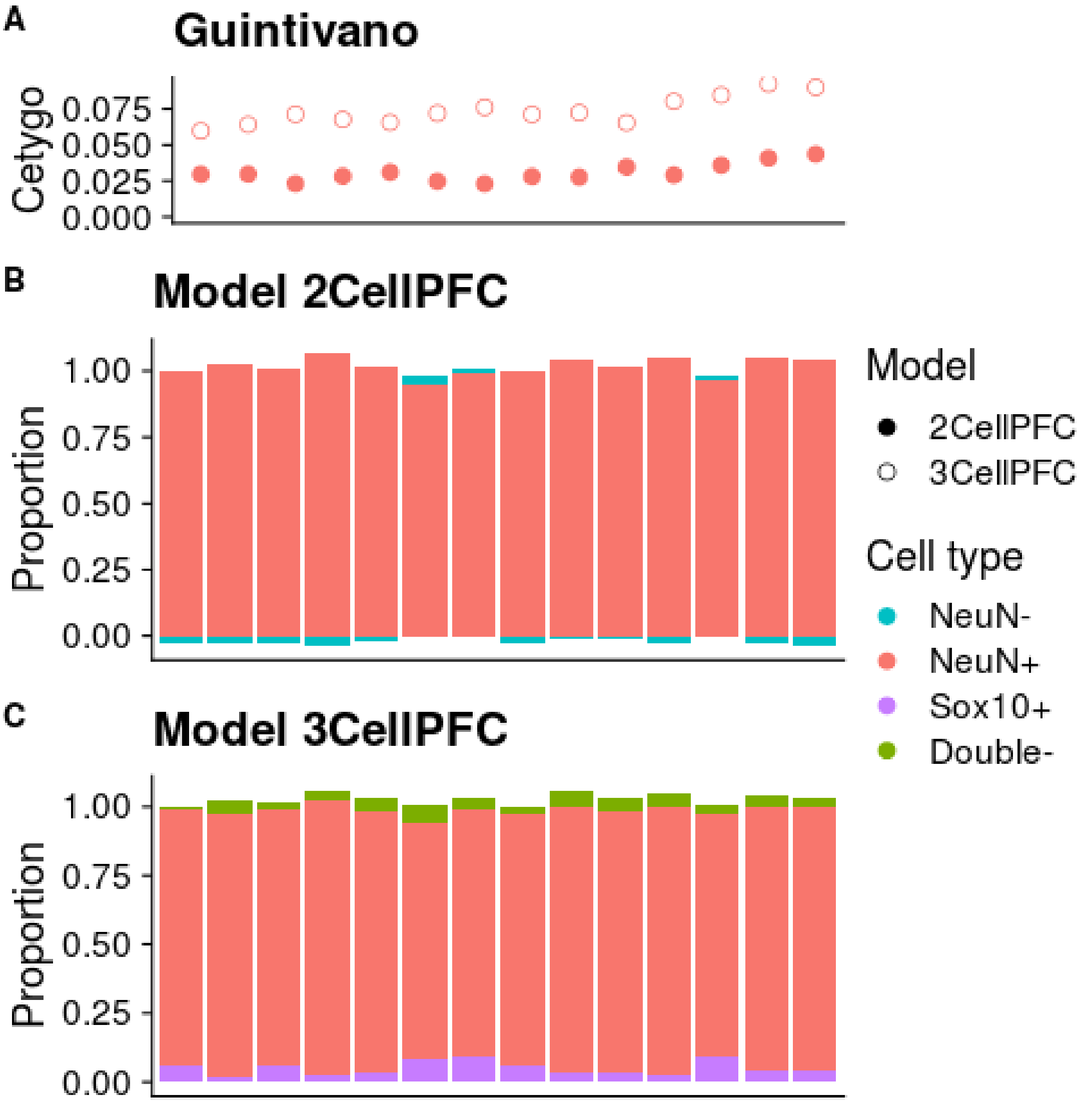


Figure 4.8: **A summary of Cetygo, estimated cell type proportions of NeuN+ samples in the testing data of Dataset Guintivano estimated using Model 2CellPFC and 3CellPFC.** A) The Cetygo across samples. B) Stacked bar plot of the predicted proportion of cell types by **Model 2CellPFC**. C) Stacked bar plot of the predicted proportion of cell types by **Model 3CellPFC**. Each bar/point represents a sample in the dataset, which are consistently ordered across plots. Bar colour represents cell type composition with NeuN- in teal, NeuN+ in pink, Sox10+ in purple, and Double- in green.

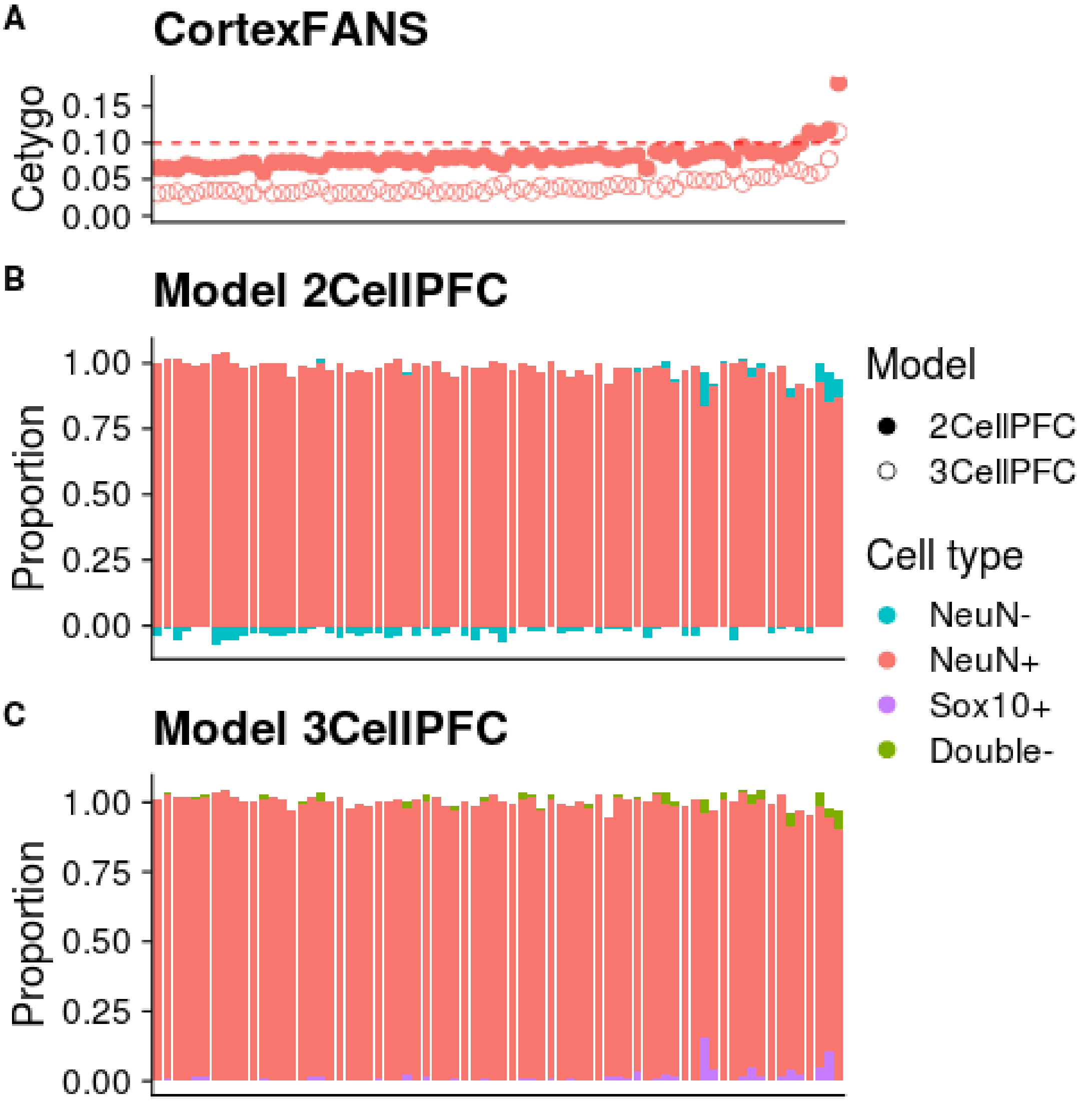


Figure 4.9: **A summary of Cetygo, estimated cell type proportions of NeuN+ samples in the testing data of Dataset CortexFANS estimated using Model 2CellPFC and 3CellPFC.** A) The Cetygo across samples. B) Stacked bar plot of the predicted proportion of cell types by **Model 2CellPFC**. C) Stacked bar plot of the predicted proportion of cell types by **Model 3CellPFC**.

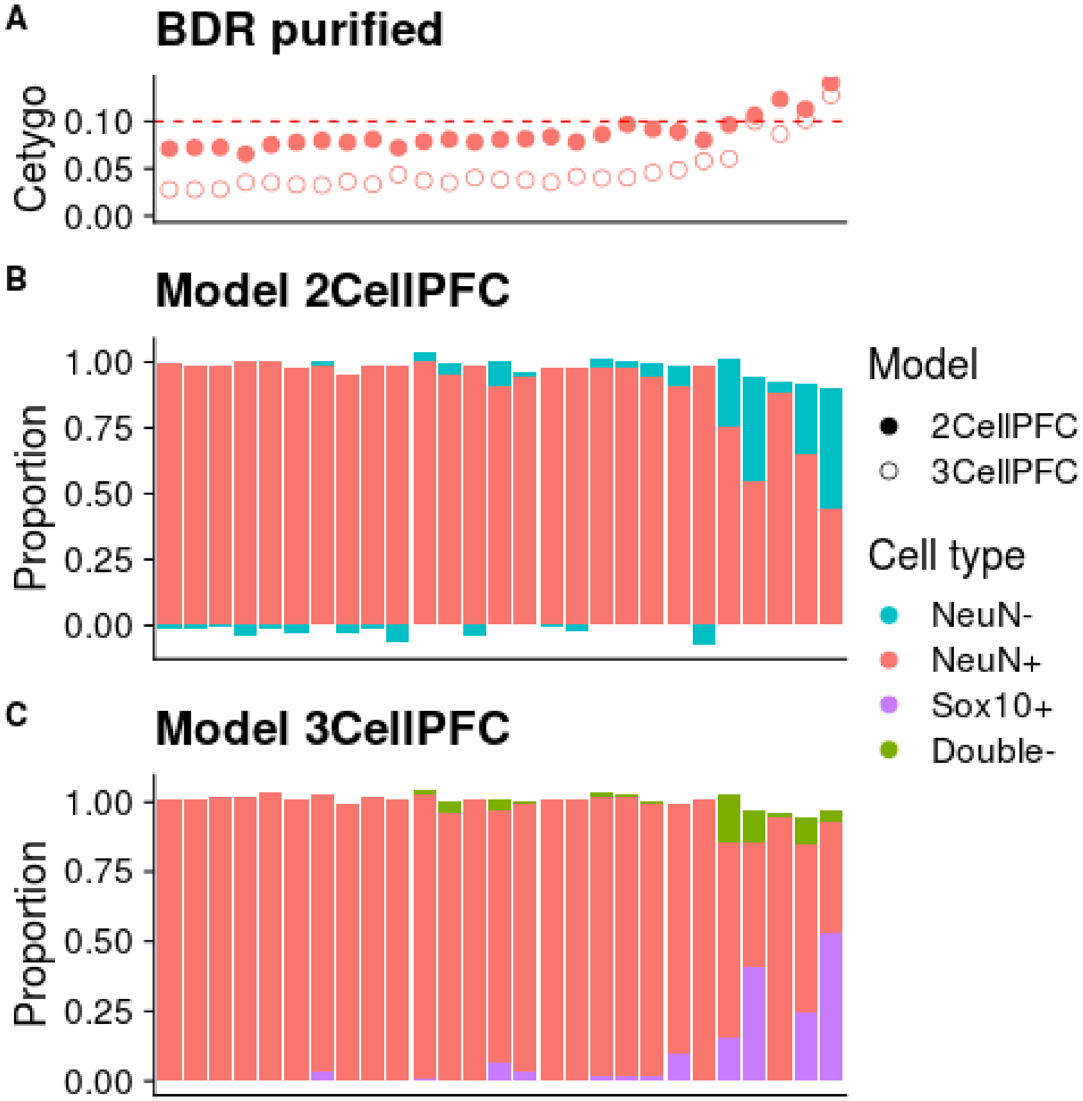


Figure 4.10: **A summary of Cetygo, estimated cell type proportions of NeuN+ samples in Dataset BDR purified estimated using Model 2CellPFC and 3CellPFC.** A) The Cetygo across samples. B) Stacked bar plot of the predicted proportion of cell types by **Model 2CellPFC**. C) Stacked bar plot of the predicted proportion of cell types by **Model 3CellPFC**.

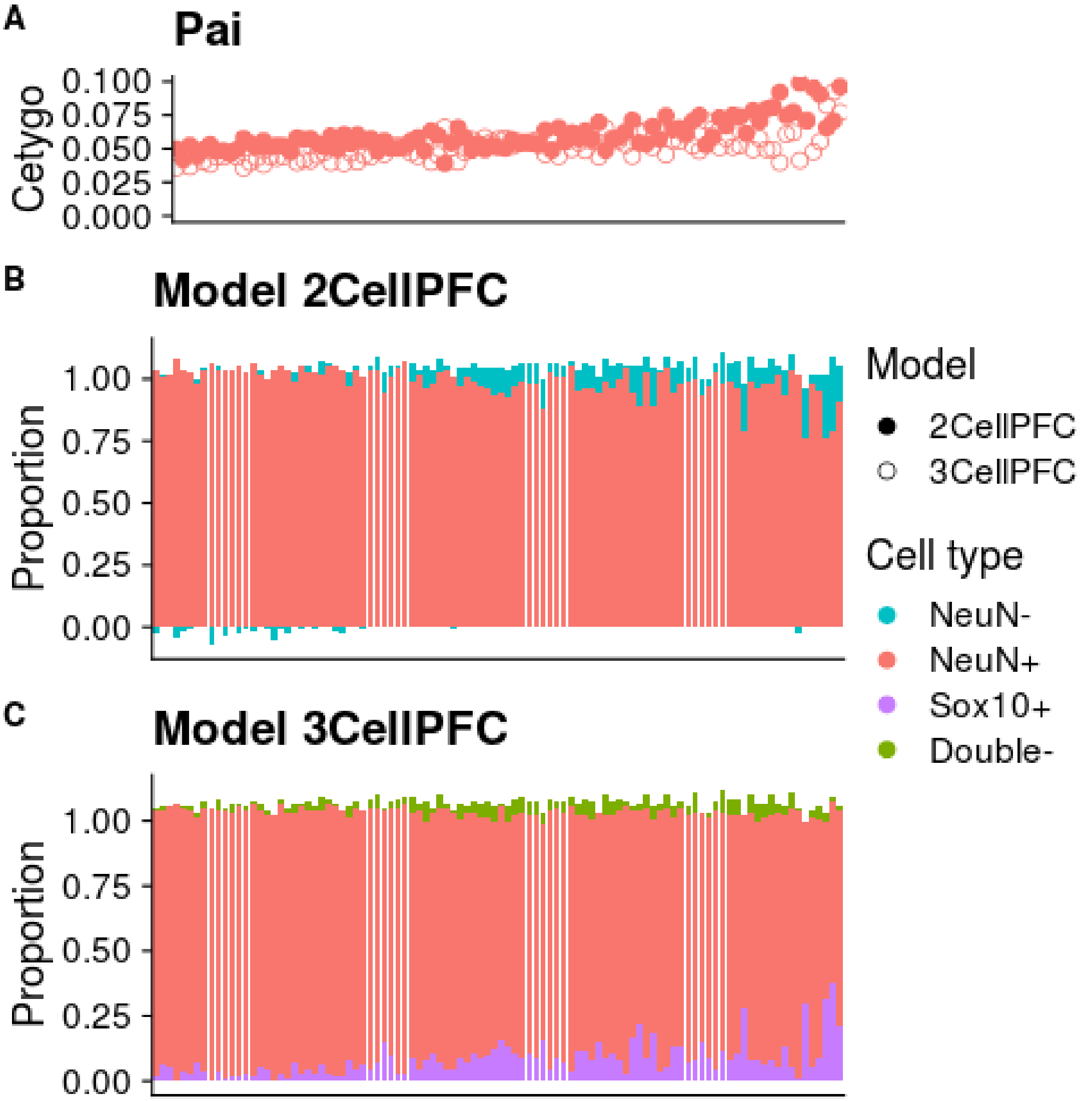


Figure 4.11: **A summary of Cetygo, estimated cell type proportions of NeuN+ samples in Dataset Pai estimated using Model 2CellPFC and 3CellPFC.** A) The Cetygo across samples. B) Stacked bar plot of the predicted proportion of cell types by **Model 2CellPFC**. C) Stacked bar plot of the predicted proportion of cell types by **Model 3CellPFC**.

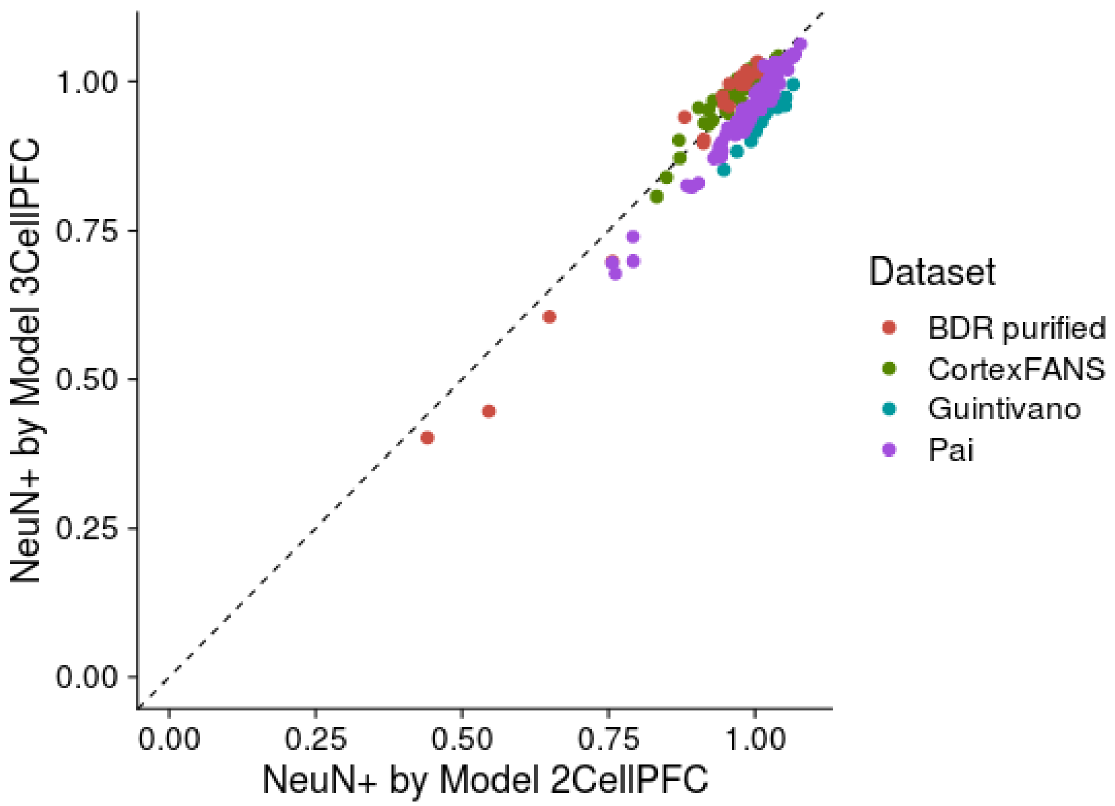


Figure 4.12: **A scatter plot between the predicted proportion of NeuN+ in NeuN+ purified samples estimated by Model 2CellPFC and Model 3CellPFC.** There is a high correlation between the predicted proportion of NeuN+ in NeuN+ purified samples between models. NeuN+ samples were taken from **Datasets BDR purified** (orange), **CortexFANS** (green), **Guintivano** (teal), and **Pai** (purple). The black dashed line marks y = x. The correlation between data is 0.914.

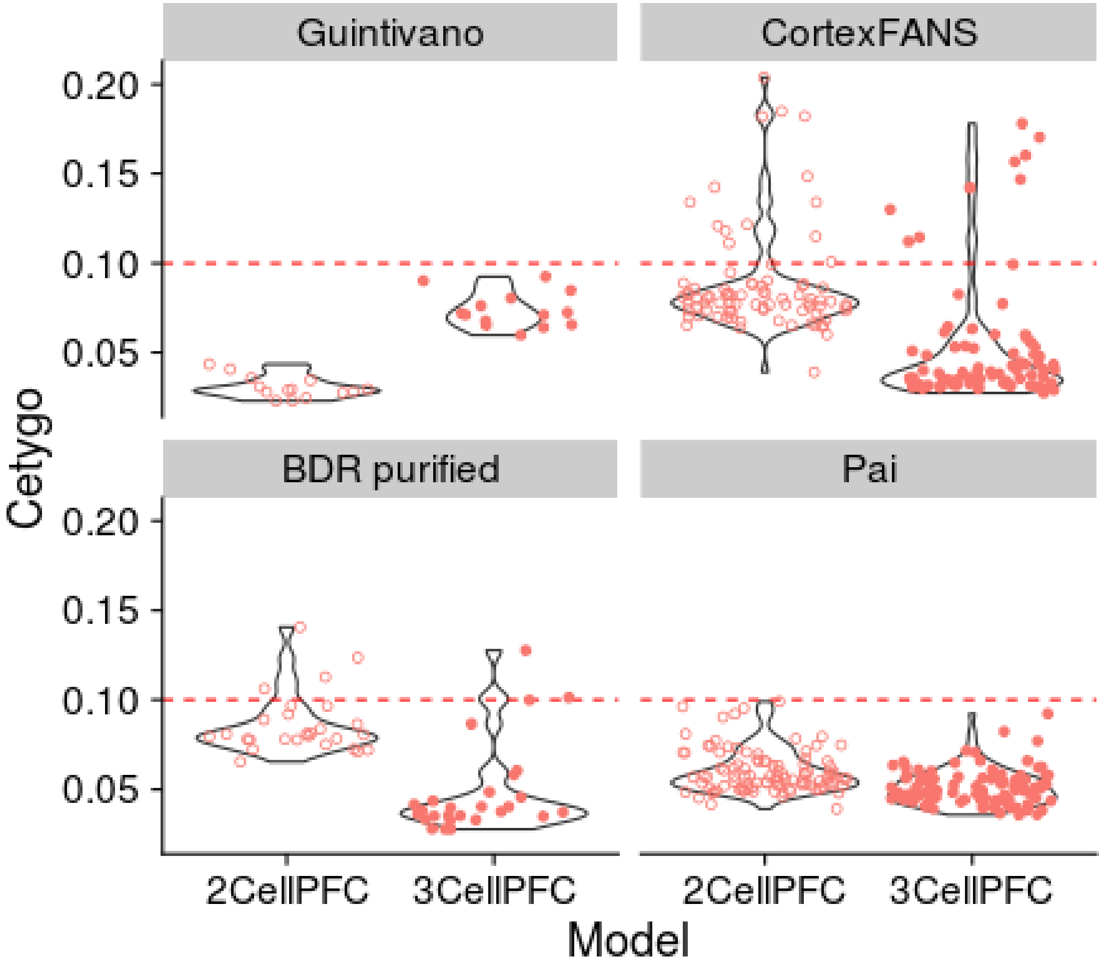


Figure 4.13: **Violin plots of Cetygo estimated from Model 2CellPFC and Model 3CellPFC across NeuN+ samples in Datasets Guintivano, CortexFANS, BDR purified, and Pai.** The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions.

Model 3CellPFC may be slightly more accurate across all samples.

Dataset Guintivano had significantly lower Cetygo in **Model 2CellPFC** than **Model 3CellPFC** (p = 2.82e-11, **Table 4.3**), while the opposite was true for **Dataset CortexFANS** (p = 5.02e-34), **BDR purified** (p = 3.79e-15), and **Dataset Pai** (p = 2.64e-08). This may be due to the aforementioned batch effects, with **Model 2CellPFC** generated from the training data of **Dataset Guintivano**, and **Model 3CellPFC** generated in the training data of **Dataset CortexFANS**, and using the same protocol and lab as **Dataset BDR purified**. The absolute mean difference of Cetygo across samples in **Dataset Pai**, the only dataset entirely independent from the model generation data, was tenfold smaller compared to other datasets, at 0.00886. Results from **Dataset Pai** suggest that **Model 3CellPFC** is slightly more accurate than **Model 2CellPFC**, although at a small magnitude, in NeuN+ predictions.

Overall, findings demonstrate that **Model 3CellPFC** is an improvement upon **Model 2CellPFC**.

4.7 Assessing model behaviour when applied to purified subsets of the cell types profiled in model reference data

The FANS separation of bulk brain tissue into purified populations using multiple cell type specific markers is a challenging process that requires experimental optimisation. As such, a reference dataset in which more than three nuclear brain populations have been profiled, one of which is a heterogeneous negatively selected population, is yet to be generated with sufficient sample size across all cell types. This means that a scenario may occur where a purified sample is deconvoluted that is not optimally profiled in the model.

To assess how deconvolution models respond to such a scenario, **Model 3CellPFC** and **Model 2CellPFC** were applied to the DNA methylation profiles of samples purified with more granularity than the samples used for model generation. Both models were applied across samples from **Dataset BDR purified** and the testing data from

Table 4.3: A paired t-test comparison of Cetygo between **Model 2CellPFC** and **Model 3CellPFC** predictions of both dataset testing data, subset by cell type. The mean difference is calculated by subtracting the mean Cetygo of **Model 3CellPFC** from **Model 2CellPFC**.

Dataset	Cell type	P value	Absolute mean difference
Guintivano Test	NeuN+	2.82e-11	0.0432
CortexFANS Test	NeuN+	5.02e-34	0.0352
BDR purified	NeuN+	3.79e-15	0.0379
Pai	NeuN+	2.64e-08	0.00886
CortexFans and **BDR purified**	Sox10+	9.27e-4	0.0046
CortexFans and **BDR purified**	Double-	6.72e-13	0.0176
BDR purified	IRF8+	0.0777	0.0164

Dataset CortexFANS, comprised of 80 Sox10+ samples (oligodendrocytes), 71 Double- samples (negative population expected to be microglia and astrocytes), 3 IRF8+ samples (microglia), and 2 Triple- samples (negative population expected to be astrocytes).

Sox10+, Double-, IRF8+ and Triple- populations are all derived from the NeuN- population (a diagram for the FANS cell purification hierarchy used can be found in **Figure 4.14**). As such it would be expected that each of these cell types are predicted as entirely NeuN- from **Model 2CellPFC**. However, each cell type makes up only a subset of the NeuN- profile, and the smaller the subset, the larger the expected difference in the DNA methylation profiles of the cell types and NeuN-, and subsequently, the higher Cetygo is expected to be. We can apply the same logic to estimates from **Model 3CellPFC**, for IRF8+ and Triple- samples which are subsets of the Double- population.

Predictive accuracy (as measured by Cetygo) was comparable across Sox10+ and Double- samples, despite the lower granularity of **Model 2CellPFC**. **Figure 4.15** shows the distribution of Cetygo for each purified cell population: Cetygo estimates for the Sox10+ and Double- populations are significantly different between models (p = 9.27e-4 and p = 6.72e-13, respectively). The Cetygo of Double- is higher in **Model 2CellPFC**, however, it is lower in Sox10+ despite only being purified in **Model 3CellPFC**. For both cell types the magnitude of difference in Cetygo between models is relatively small (0.0046 and 0.0176, respectively). Both cell types are predicted as mostly NeuN- by **Model 2CellPFC** (**Figure 4.16** and **4.17**). These findings suggest that the NeuN- samples in **Dataset Guintivano** training data may be comprised of a higher proportion of Sox10+ than Double-, resulting in a lower Cetygo in Sox10+ predictions. This was further confirmed by applying **Model 3CellPFC** to the NeuN- samples in the training data of **Dataset Guintivano**, where the mean proportion of Sox10+ was 0.761 (sd = 0.093), mean Double- was 0.291 (sd = 0.0941) (**Figure 4.18**).

IRF8+ and Triple- were not available as purified populations in either model. The Cetygo of IRF8+ is elevated (mean = 0.171 across both models). This could be explained by observations from Section 3.8.1, which showed that the IRF8+ samples had lower median array intensity, and as such would have higher noise across their DNA methylation profiles. Triple- has a lower Cetygo (mean = 0.061 across both models)

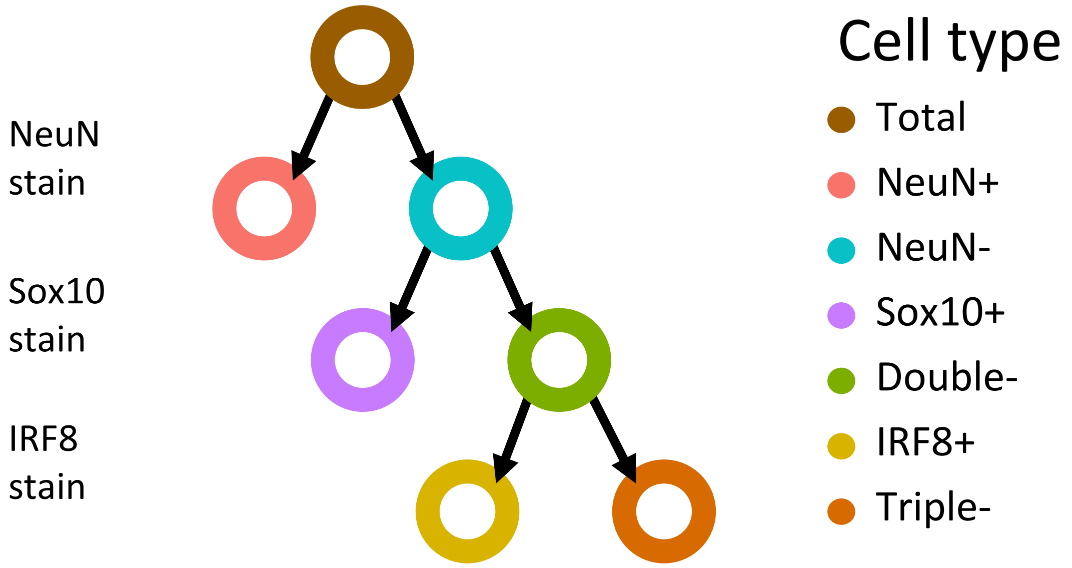


Figure 4.14: **A diagram of the cellular hierarchy in FANS sorting of PFC tissue using the protocol developed by Policicchio et al., 2020a.** NeuN+ populations are enriched for neuronal nuclei, Sox10+ are enriched for oligodendrocyte nuclei, and IRF8+ are enriched of microglial nuclei.

and is predicted as NeuN- by **Model 2CellPFC**, however, is not predicted as Double- but instead as mostly Sox10+ by **Model 3CellPFC** (**Figure 4.19**). This could suggest that, rather than Sox10+ being comprised largely of Triple- (which should be impossible), the purification of Triple- samples may not have been clean. A larger sample size would be needed for further investigation. Generally, the models may not be applicable to the more purified samples.

4.8 Investigating cell type bias in prediction accuracy in Model 3CellPFC using simulated data

A higher number of cell types in a deconvolution reference dataset might be assumed to always be beneficial, as more granularity in predictions would lead to more information gained from deconvolution. However, if the cell type specific DNA methylation profiled does not contain large enough DNA methylation differences, or a sufficient number of DNA methylation sites to distinguish cell types (or the algorithm used to generate the deconvolution model does not exploit them), similar cell types will be predicted in each others place (as seen in blood deconvolution in Section 3.4.1.3).

To investigate cell type bias in prediction accuracy, the dataset generated in Section 4.5 was again utilised, containing 60 simulated samples with every combination of cell type proportions per cell type between 0 and 1 in a step size of 0.1 so that proportions summed to 1 across cell types.

The absolute difference between predicted and actual proportion was plotted in **Figure 4.20**. **Model 3CellPFC** was found to have no obvious cell type bias, with the range and distribution of each cell type being very similar, suggesting that the cell types within the model are distinct enough so as not to easily be predicted as one and other.

This analysis could not be repeated across **Model 2CellPFC**, as the model contained only two cellular populations and given that predictions sum close to one, the predictions were not independent enough to assess cell type bias.

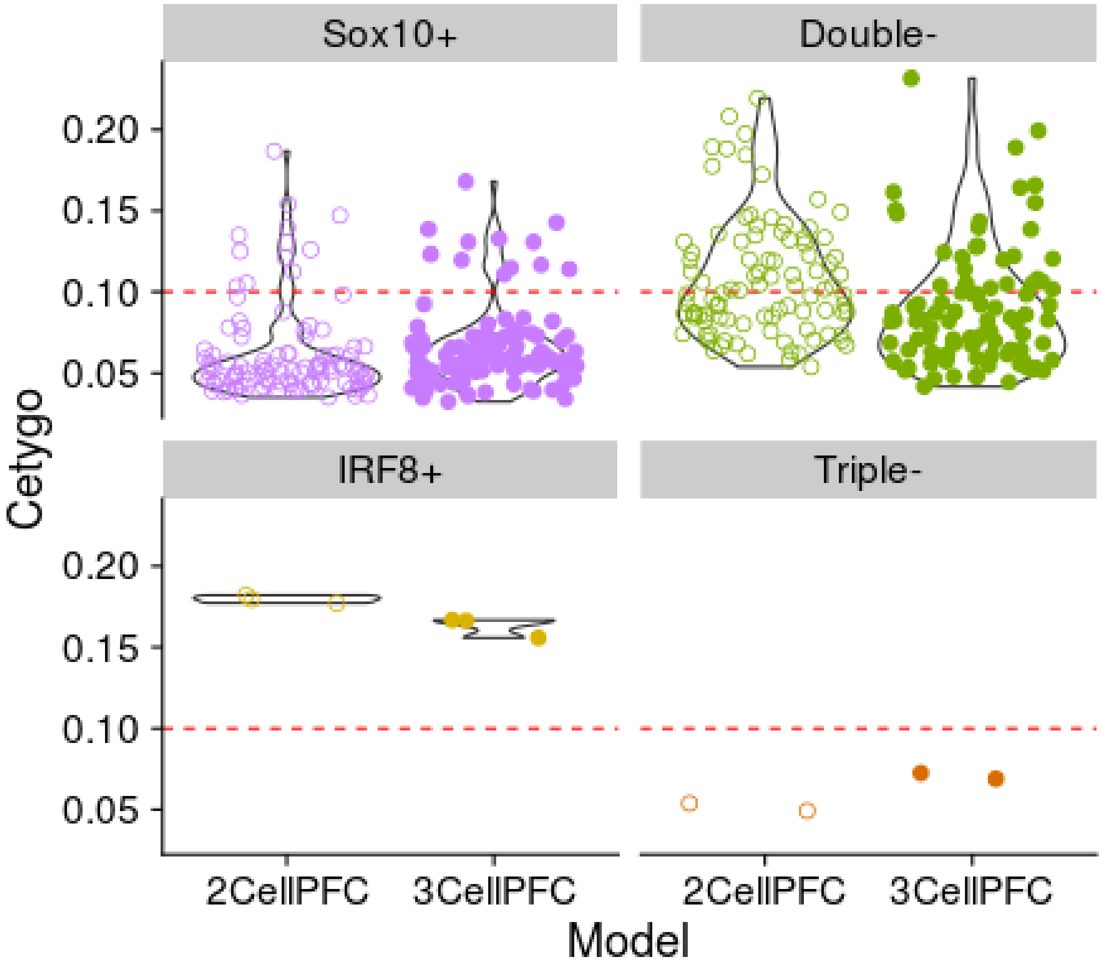


Figure 4.15: **A violin plot of Cetygo estimated using Model 2CellPFC and Model 3CellPFC across Sox10+, Double-, IRF8+, and Triple- populations.** Samples used originated from the testing data of **CortexFANS**, and the independent data, **Dataset BDR purified**. The distribution of Cetygo amongst Triple- samples could not be summarised using a violin plot due to the low sample size. The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions.

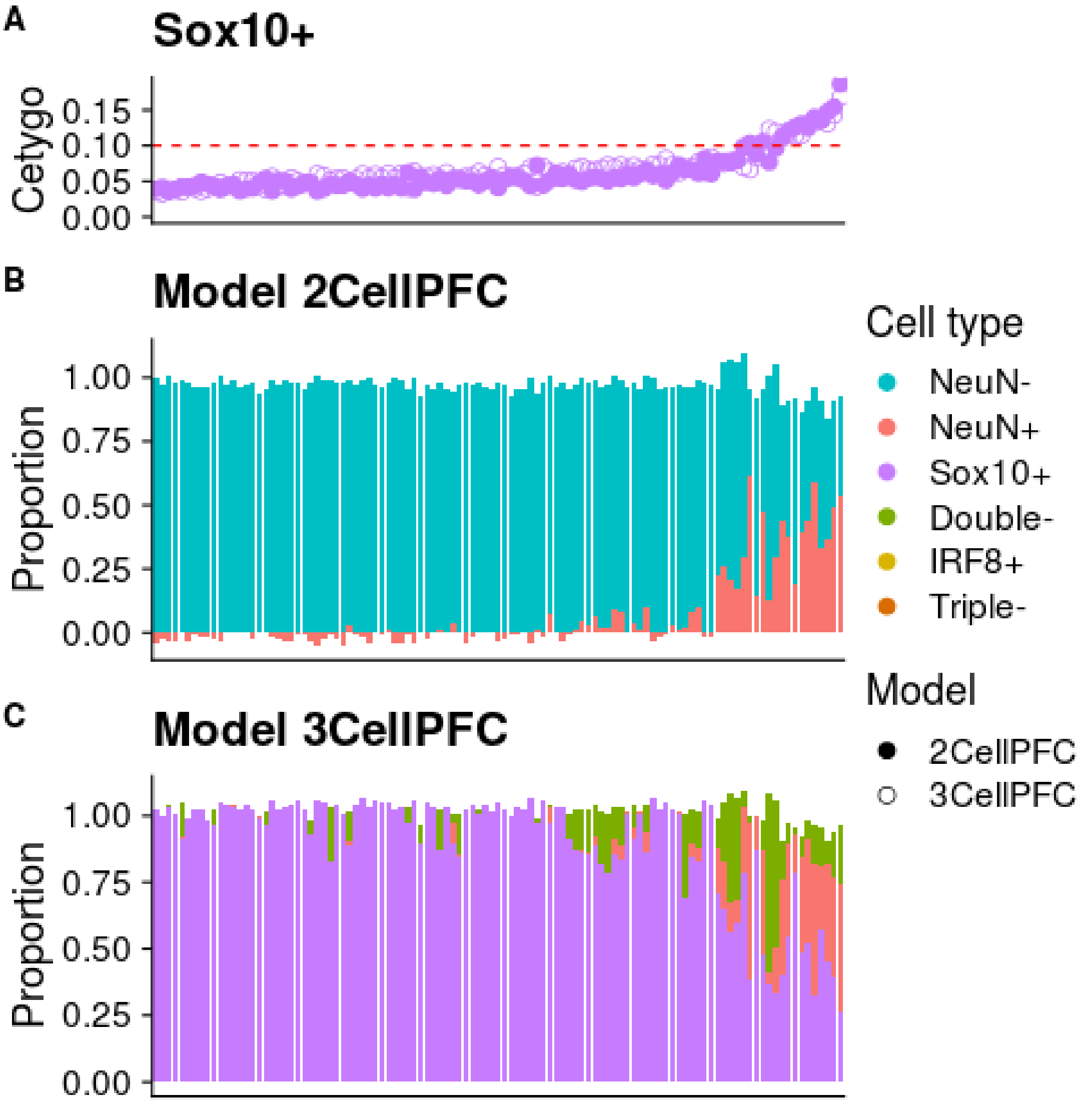


Figure 4.16: **A summary of Cetygo, and the predicted proportions estimated using Model 2CellPFC and 3CellPFC for Sox10+ samples in Dataset Cortex-FANS and Dataset BDR purified.** A) The Cetygo across samples. B) Stacked bar plot of the predicted proportion of cell types by **Model 2CellPFC**. C) Stacked bar plot of the predicted proportion of cell types by **Model 3CellPFC**. The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions.

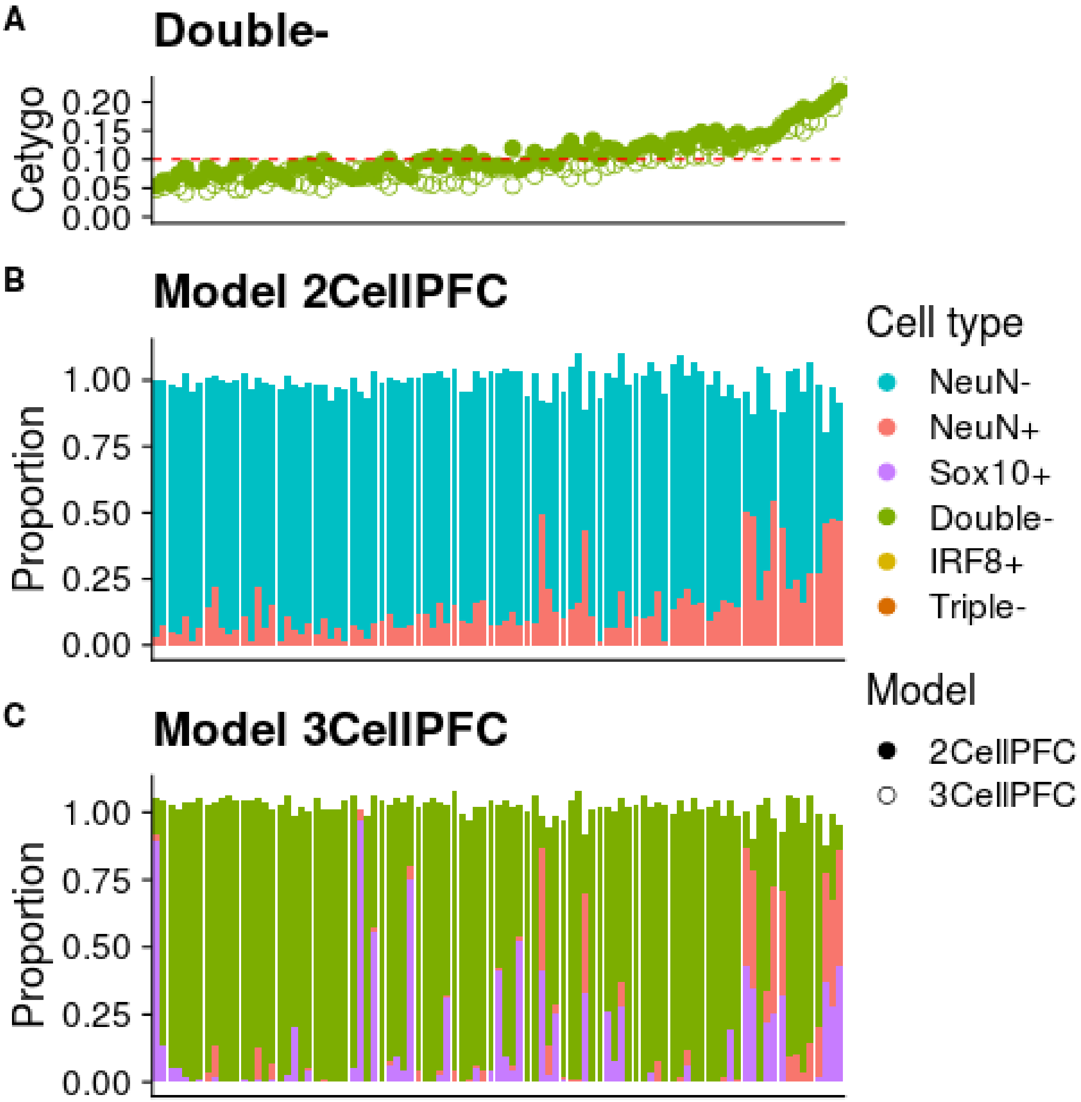


Figure 4.17: **A summary of Cetygo, and the predicted proportions estimated from Model 2CellPFC and 3CellPFC across Double- samples in Dataset CortexFANS and Dataset BDR purified.** A) The Cetygo across samples. B) Stacked bar plot of the predicted proportion of cell types by **Model 2CellPFC**. C) Stacked bar plot of the predicted proportion of cell types by **Model 3CellPFC**. The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions.

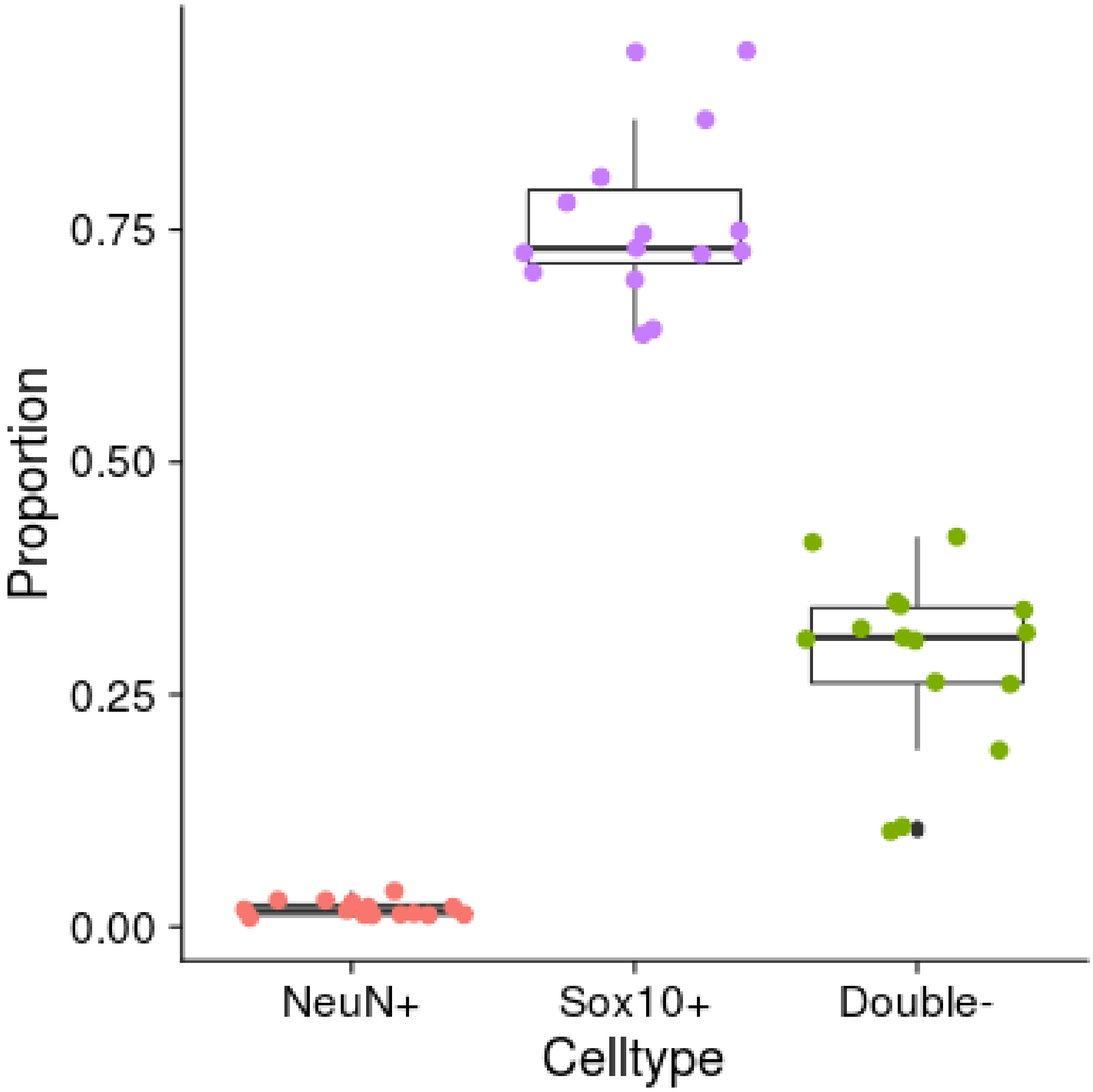


Figure 4.18: **A box plot of the predicted proportions of NeuN- samples in Dataset Guintivano training data estimated using Model 3CellPFC.** The predicted proportions of Sox10+ are higher than Double- across NeuN- samples in **Dataset Guintivano** training data.

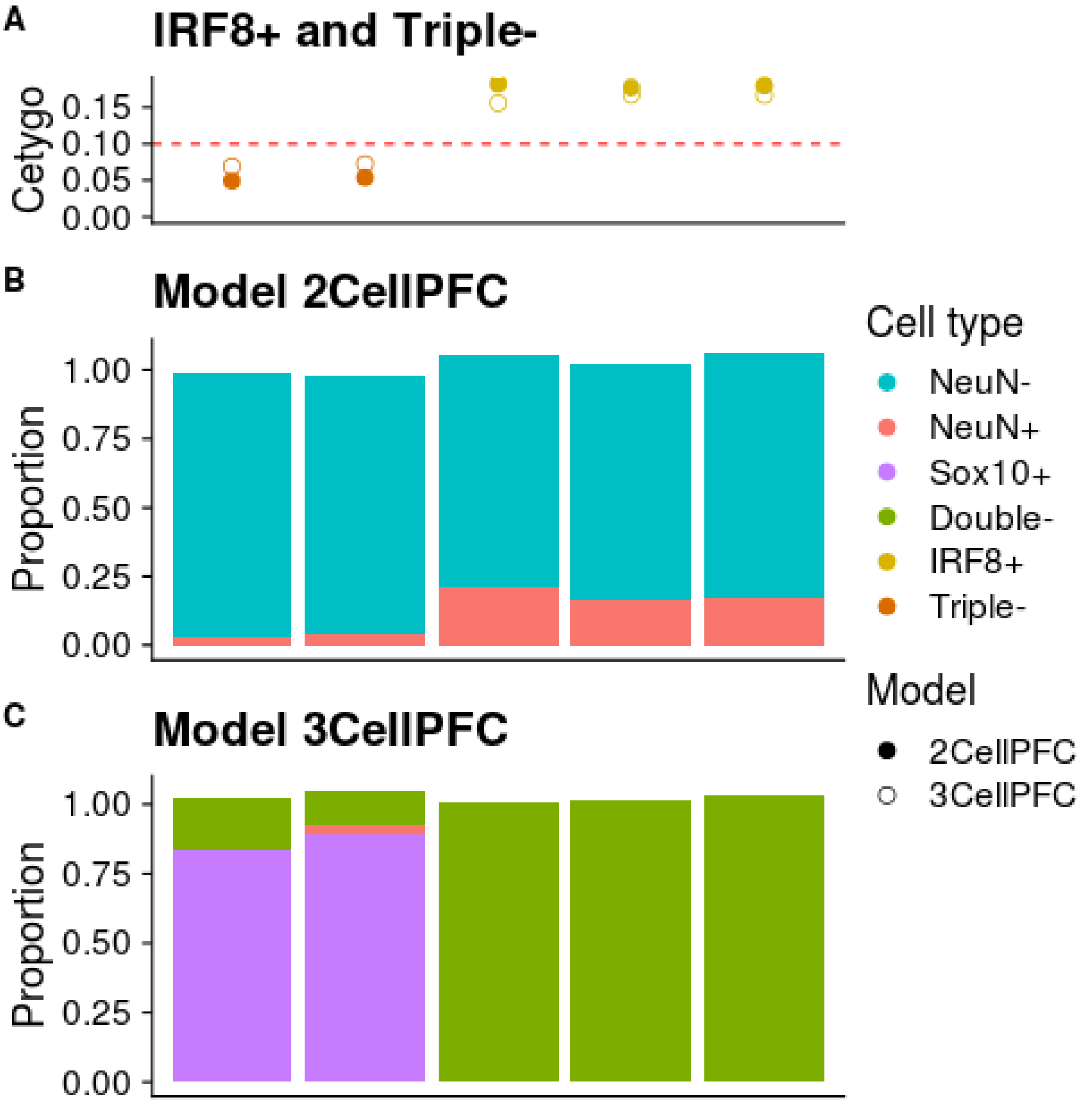


Figure 4.19: **A summary of Cetygo, and the predicted proportions estimated from Model 2CellPFC and 3CellPFC across IRF8+ and Triple- samples in Dataset BDR purified.** A) The Cetygo across samples. B) Stacked bar plot of the predicted proportion of cell types by **Model 2CellPFC**. C) Stacked bar plot of the predicted proportion of cell types by **Model 3CellPFC**. The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions.

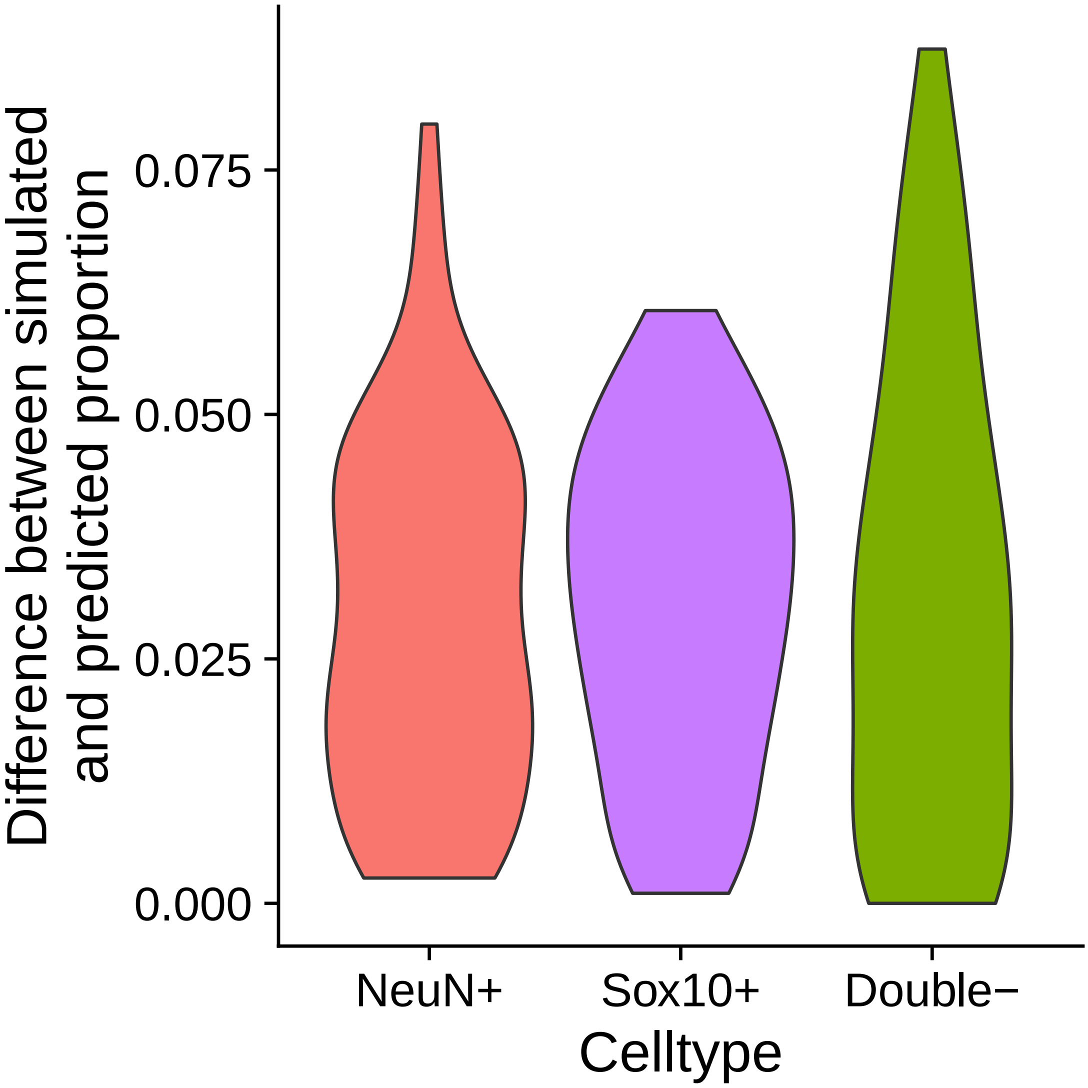


Figure 4.20: **A violin plot showing the absolute difference between predicted and actual cell type proportion in simulated data comprised of NeuN+, Sox10+ and Double- cell type populations.** Samples were simulated using **Dataset CortexFANS**, with cell type proportions set as any combination of 0-1 per cell type in steps of 0.1 that sum to 1. Proportions were predicted using **Model 3CellPFC**.

4.9 Investigating the wider applicability of the Model 3CellPFC

Disparities between brain cell type specific DNA methylation profiles can arise due to a number of drivers (see Section 4.1.1.4 for an overview), including different cellular function or context within the brain, e.g. in different brain regions, with disease progression, and during development. The DNA methylation profiles of neuronal cell lines, commonly used in neuropsychiatric disease research, will undoubtedly also differ from purified post-mortem PFC reference data. Here, the applicability of **Model 3CellPFC** was assessed across these differences, utilising Cetygo to measure model applicability, as the true cell type proportions are unknown.

4.9.1 Model 3CellPFC is applicable across all cerebral cortex and subcortical areas tested but not cerebellum

To investigate the applicability of **Model 3CellPFC** across brain regions, **Dataset Adult brain** and **Dataset BDR bulk** were utilised. **Dataset Adult brain** contained brain regions BA11 (n = 38), BA25 (n = 26), BA9 (n = 26), cerebellum (n = 64), hippocampus (n = 42), striatum (n = 98), and thalamus (n = 24). **Dataset BDR bulk** is comprised of PFC (n = 671) and OCC (n = 613) samples. Broddman area (BA) notation is used in **Dataset Adult brain** to divide brain regions based on cellular architecture, however, to be consistent across findings, BA9 and PFC from **Dataset BDR bulk** were combined under the **PFC** label, as the PFC includes BA9.

The majority of samples within each brain region had Cetygo <0.1, excluding cerebellum (CEREB) (**Figure 4.21**, **Table 4.4**). When considering the literature, it is unsurprising that cerebellum has an elevated Cetygo, as it is known to contain a neuronal population distinct from other brain regions, containing the neuronal cell subtypes: Purkinje cells and granule cells (Bartheld, Bahney and Herculano-Houzel, 2016) which do not express NeuN. This also corroborates findings in epigenetic clock research, in which DNA methylation is used to estimate 'biological' age in the brain, which show CEREB

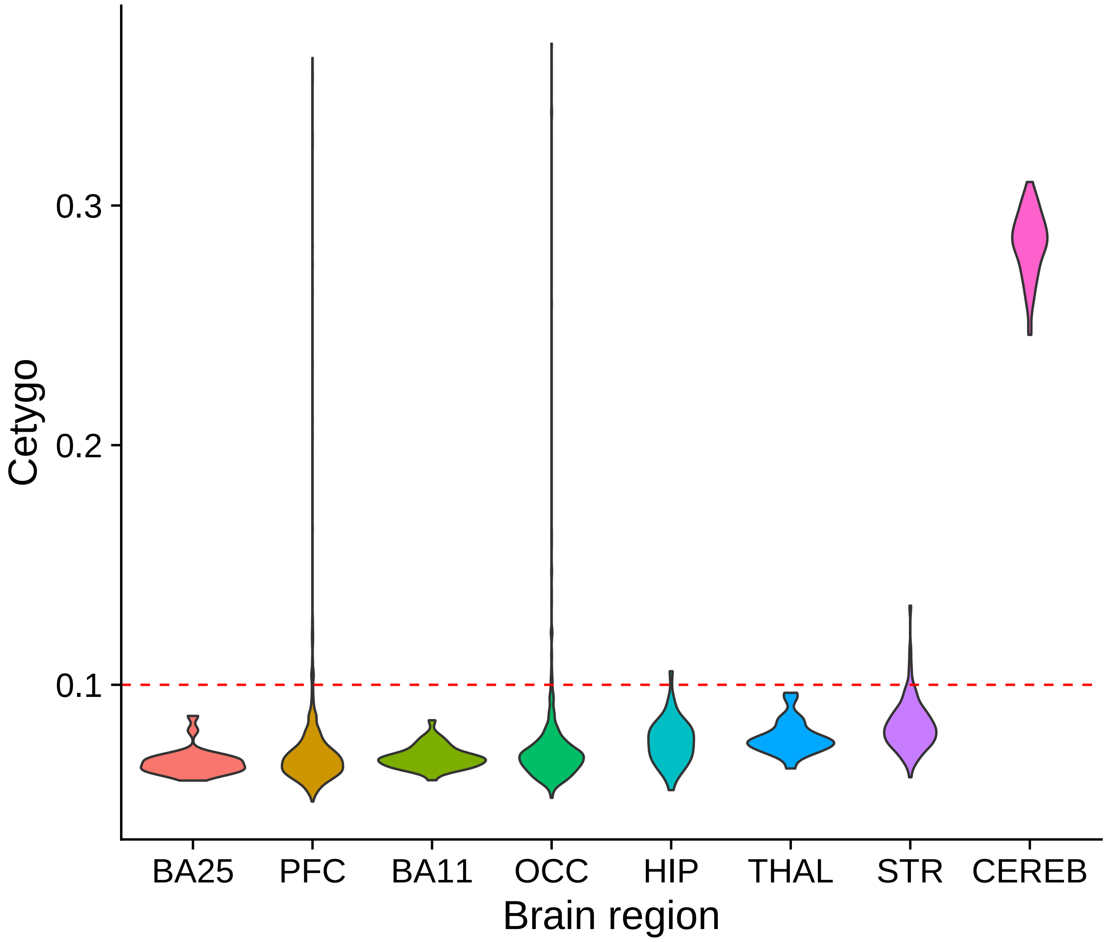


Figure 4.21: **A violin plot of Cetygo across Dataset BDR bulk and Dataset Adult brain by brain region, calculated using Model 3CellPFC.** The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions. BA = Brodmann area, PFC = Prefrontal cortex, OCC = Occipital lobe, HIP = Hippocampus, THAL = Thalamus, STR = Striatum, CEREB = Cerebellum. Region BA9 from **Dataset Adult brain** was included in the PFC region, as BA9 is a subset of the PFC. Brain regions are ordered by mean Cetygo.

tissue to have a less accurate (generally underestimated) predicted biological age than other brain regions due to its differential DNA methylation (Horvath et al., 2015).

4.9.2 Model 3CellPFC is applicable across all Braak stages of Alzheimer's disease

AD is a neurodegenerative disease, the progression of which has been found to result in differential methylation (Smith et al., 2021) and cell type specific transcriptional changes (Mathys et al., 2019). As such, **Model 3CellPFC** may not be applicable to samples where AD was more severe. To assess model applicability across AD, **Dataset BDR bulk** was utilised. Braak stage, a metric of neuropathology across brain regions associated with AD progression (Braak and Braak, 1991), had been quantified across the dataset. **Dataset BDR bulk** contained 45 samples at stage 0, 129 samples at stage I, 246 at stage II, 163 at stage III, 128 at stage IV, 206 at stage V, and 325 at stage VI, the most severe Braak stage.

If DNA methylation differences associated with AD progression were observed across the DNA methylation sites used for deconvolution in **Model 3CellPFC**, then prediction accuracy might be affected as the DNA methylation profile of cell types within the tissue deviate further from those in the reference dataset.

It was found that increased Braak stage was not related to prediction quality as measured by Cetygo ($p = 0.182$) (**Figure 4.22**, **Table 4.4**). This suggests the differential methylation of AD is not of a large enough magnitude or present at model cell type specific sites, and so does not impede prediction accuracy.

4.9.3 Model 3CellPFC is not applicable to fetal cortex samples

During early development, many changes in gene expression occur across all brain cells, as observed using singe cell RNA-seq techniques (Zhong et al., 2018; Fan et al., 2018). In these early stages, not all neurons express NeuN (Sarnat, Nochlin and Born, 1998). Therefore, it was of interest how **Model 3CellPFC**, developed using adult samples which were purified using the NeuN neuronal marker, would perform across fetal and young samples, and if there might be an age at which the brain was sufficiently developed for

Table 4.4: **A summary of the proportion of predictions with Cetygo <0.1 across Datasets BDR bulk, Adult brain, Fetal, iPSC and SH-SY5Y.** BA = Brodmann area, PFC = Prefrontal cortex, OCC = Occipital lobe, HIP = Hippocampus, THAL = Thalamus, STR = Striatum, CEREB = Cerebellum, iPSC = induced pluripotent stem cell. Low Cetygo reflects higher deconvolution accuracy.

Figure	Group	Number of samples	% Cetygo <0.1
Brain region (4.21)	BA25	26	100
	PFC	697	95.8
	BA11	38	100
	OCC	613	96.6
	HIP	42	97.6
	THAL	24	100
	STR	98	95.9
	CEREB	64	0
Braak stage (4.22)	0	45	93.3
	1	129	92.2
	2	246	96.7
	3	163	95.1
	4	128	96.1
	5	206	97.6
	6	325	96.9
Fetal (4.23)	-	136	3.67
Cell line (4.25)	iPSC	93	0
	SH-SY5Y	156	0

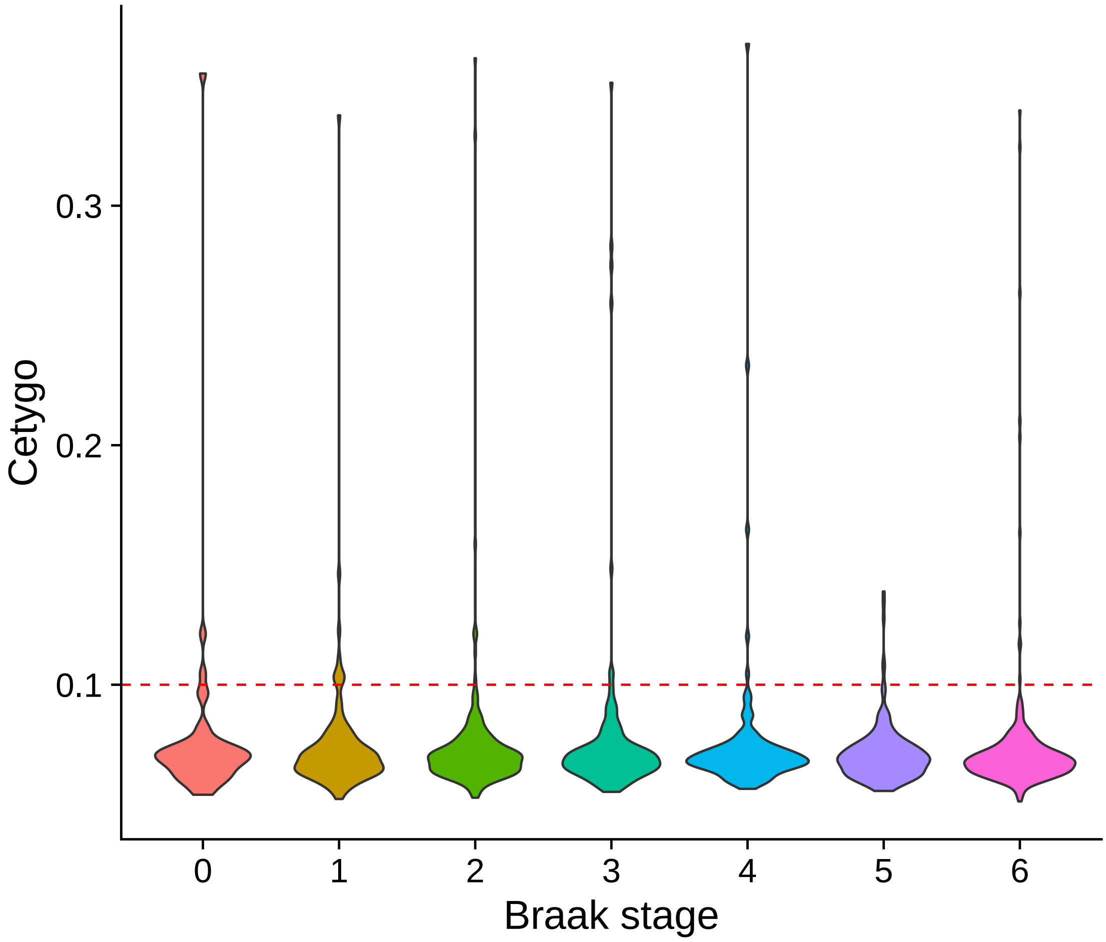


Figure 4.22: **A violin plot of Cetygo across Braak tangle stage across Dataset BDR bulk, calculated using Model 3CellPFC.** The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions.

the predictions to become accurate. **Dataset Fetal** was utilised, in which the mean age was 23.5 pcw ± 46.pcw, with a range of 6 - 456 pcw. All samples except three were prenatal, their ages being 2, 3, and 8 years old.

Cetygo was high for all but a few samples across **Model 2CellPFC** and **Model 3CellPFC** (**Figure 4.23**, **Table 4.4**), showing that low Cetygo in fetal samples was not due to the **CortexFANS** reference data alone. Cetygo was observed to decrease with age, with a steep decline as fetal age increased. All postnatal samples have an acceptable Cetygo (<0.1), showing that **Model 3CellPFC** can be applied to postnatal samples accurately. The findings suggest the need for a different, specific reference dataset containing samples below the age of ∼50 pcw to more accurately deconvolute samples in that lower age range. The exact age at which deconvolution becomes accurate is challenging to assess across this dataset due to the relative sparsity of older pre-natal and young post-natal samples.

4.9.4 Model 3CellPFC composition estimates have high Cetygo on data from neuron-like cell lines

Cell models have been utilised in neurobiology as a tool for assessing functionalities within the brain (see Section 4.1.1.4.4 for details). They are especially useful because, unlike in rodent models, cell lines can be of human origin, which makes investigating epigenetics, which is directly affected by genetics, more translatable to humans. However, there are questions as to how representative cell models are to fully formed brains purified cells, and especially adult brains, especially when they are being utilised to study disease for which the pathology only occurs in the brain later in life, such as PD (Avazzadeh et al., 2021).

We investigated the ability of **Model 3CellPFC** to accurately deconvolute cell lines, the result of which would give insight into how representative neuronal cell lines might be of (purified) cortical tissue. Two commonly used neural cell models were utilised: neuronal iPSCs (Shi, Kirwan and Livesey, 2012) (**Dataset iPSC**, n = 93), which are commonly used to study a range of neurological and neurodevelopmental disorders, including but by no means limited to PD, AD, SZ, autism (Engle, Blaha and Kleiman,

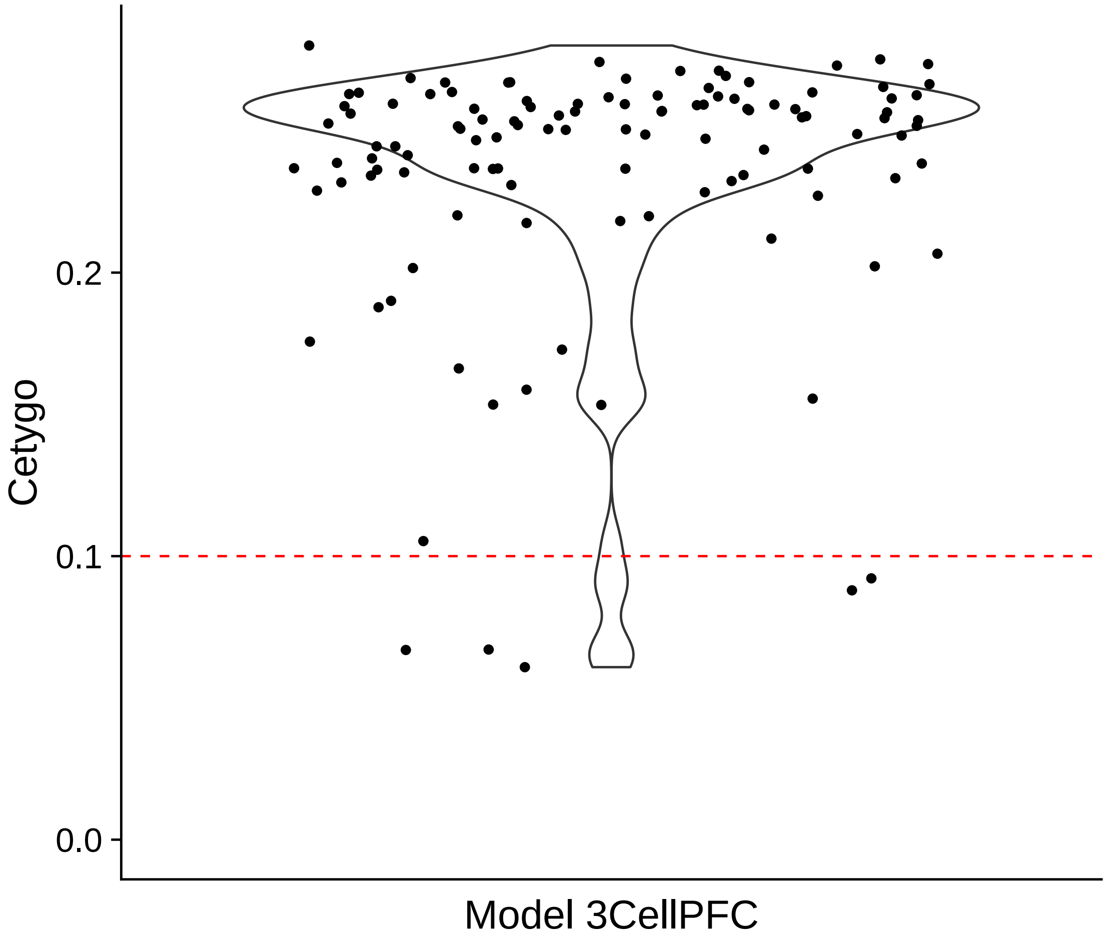


Figure 4.23: **A violin plot of Cetygo across fetal and young samples predicted by Model 3CellPFC.** The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions.

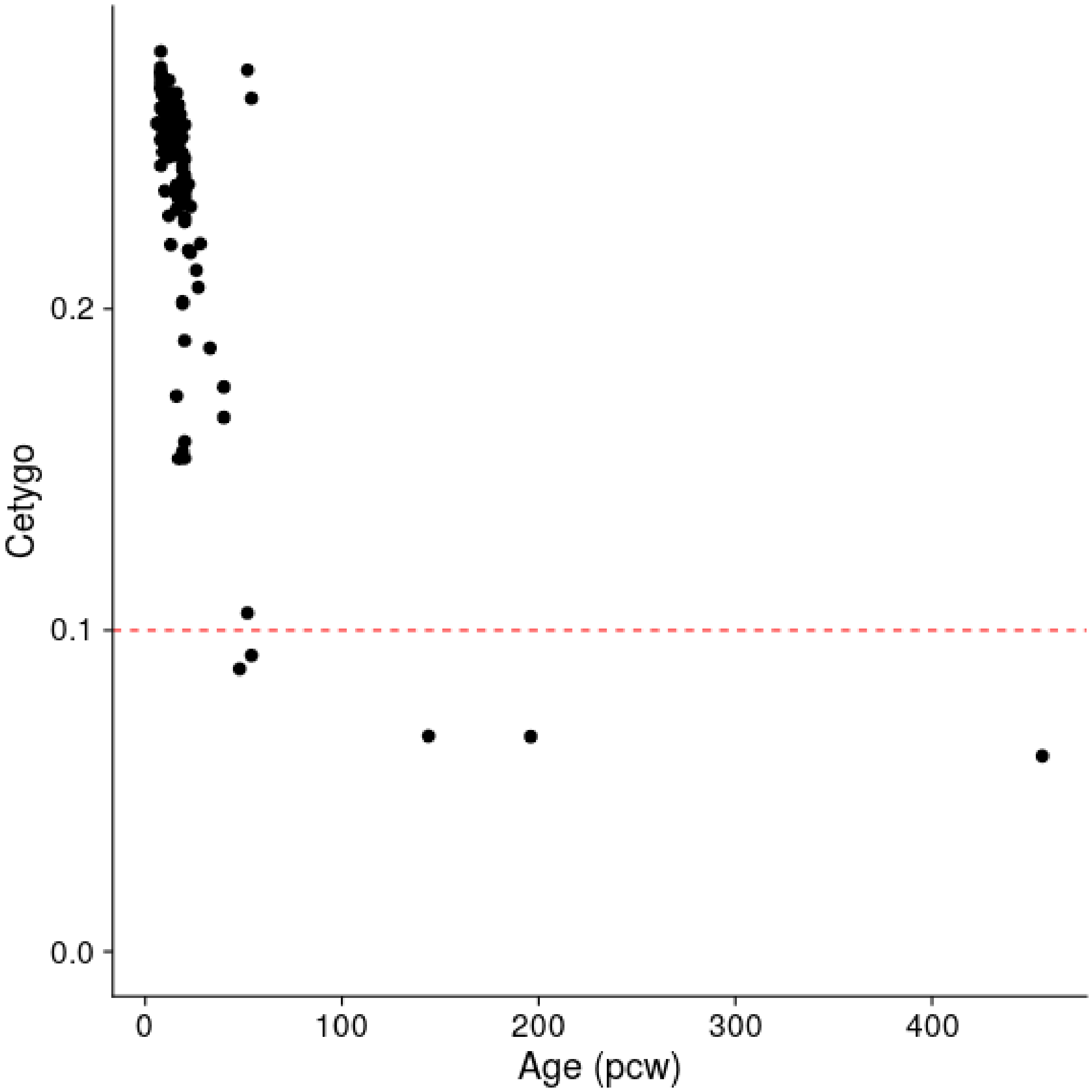


Figure 4.24: **A scatter plot showing the change in Cetygo from predictions by Model 3CellPFC as age increases in Dataset Fetal.** The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions.

2018; Dolmetsch and Geschwind, 2011), and SH-SY5Y cells (**Dataset SH-SY5Y**, n = 156), which are derived from a neuroblastoma (Biedler et al., 1978; Kovalevich and Abstract, n.d.), most commonly utilised in PD research (Constantinescu et al., 2007).

Both cell lines are derived from non-brain tissue; the iPSC cell lines are reprogrammed stem cells, which were originally keratinocytes (skin cells), and the SH-SY5Y cell line was established using a bone marrow biopsy of a metastatic neuroblastoma of a four year old female which has undergone clonal selection (Biedler et al., 1978). Even after reprogramming, cell lines have been found to maintain some 'epigenetic memory' of their cell type origin (Lister et al., 2011; Kim et al., 2011). Cell model age is also closer to fetal age (Steg et al., 2021) and as such might predict with similar accuracy as the fetal samples in Section 4.9.3.

Both cell lines had high Cetygo (mean 0.26 ± 0.00386 for iPSCs and mean 0.297 ± 0.00393 for SH-SY5Y) when using **Model 3CellPFC** (**Figure 4.25**, **Table 4.4**). Furthermore, despite both being neuronal cell lines, the average estimated proportion of NeuN+ is low (mean NeuN+ in iPSC = 0.23 ± 0.0059, SH-SY5Y = 0.365 ± 0.00732), confirming that the model, developed in adult post-mortem PFC are not applicable here. This suggests that the neuronal cell lines profiled are not comparable to purified cortical samples.

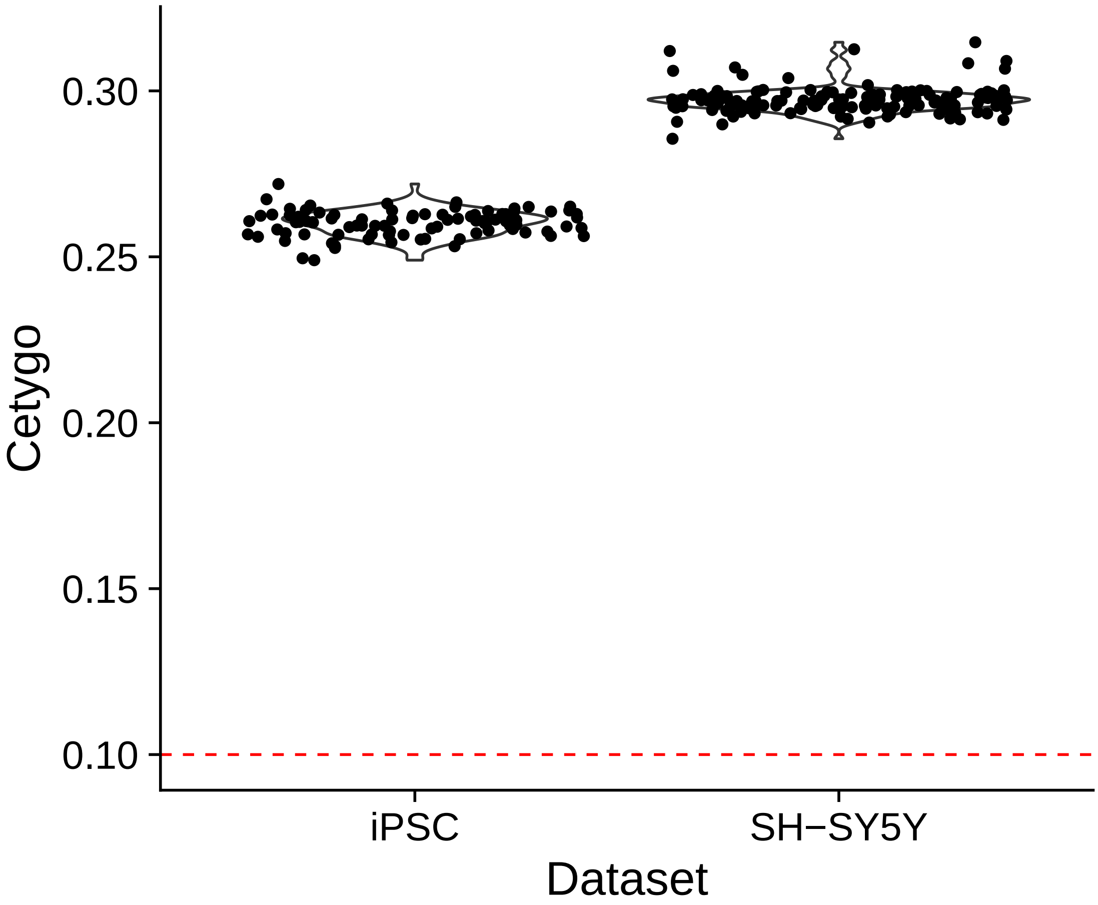


Figure 4.25: **A violin plot of Cetygo across brain-like cell line samples from predictions by Model 3CellPFC.** DNA methylation profiled from cell line samples from **Dataset iPSC** and **SH-SY5Y** were utilised, containing neuronal iPSCs and SH-SY5Y cells, respectively. The dashed line in red represents the Cetygo threshold, with values below 0.1 characterising accurate cell type proportion predictions.

4.10 Model 3CellPFC provides insights into underlying biological differences not gleaned from Model 2CellPFC

A main Chapter aim revolves around assessing the comparative utility of **Model 3CellPFC** against **Model 2CellPFC**, the latter of which was generated using the reference data from the most used PFC deconvolution method. Here, we demonstrate that **Model 3CellPFC** is advantageous in gaining biological insight that **Model 2CellPFC** does not uncover. **Dataset BDR bulk** and **Dataset Adult brain** were used to compare predicted composition between **Model 2CellPFC** and **Model 3CellPFC** across brain regions and AD.

4.10.1 Model 3CellPFC provides additional information across brain regions

Cell type composition is known to differ between brain region (Joglekar et al., 2021; Erö et al., 2018). Here, the predicted composition was compared between hippocampus (HIP), thalamus (THAL), striatum (STR), BA11, BA25, PFC and OCC. CEREB was not included due to the high Cetygo of predictions by **Model 3CellPFC** (**Figure 4.21**).

It was noted that the proportion of NeuN+ predicted by **Model 3CellPFC** was lower across all brain regions when compared to **Model 2CellPFC** predictions (**Figure 4.26**). In **Model 2CellPFC** predictions, BA25, PFC, BA11 and OCC had similar cell type profiles, with mean NeuN+ at 0.459 ± 0.103, 0.398 ± 0.134, 0.516 ± 0.0802, and 0.435 ± 0.140, respectively, and mean NeuN- 0.684 ± 0.0964, 0.624 ± 0.142, 0.591 ± 0.0771, 0.575 ± 0.147, across BA25, PFC, BA11 and OCC, respectively. In contrast, in **Model 3CellPFC** predicted proportions, while the NeuN+ samples remain similar across the four brain regions, with mean NeuN+ at 0.283 ± 0.105, 0.262 ± 0.130, 0.348 ± 0.0755, and 0.314 ± 0.146, there were differences in the proportions of Double- and Sox10+, with BA25 and BA11 exhibiting elevated Double- (mean = 0.523 ± 0.0742 and 0.442 ± 0.0577) compared to their Sox10+ proportion (mean = 0.275 ± 0.112 and 0.262

± 0.115). PFC and OCC, however, have nearly equal Sox10+ and Double- proportions (mean Sox10+ = 0.373 ± 0.216 and 0.355 ± 0.186, mean Double- = 0.340 ± 0.131 and 0.302 ± 0.116). **Model 3CellPFC**s additional granularity is also evidently beneficial in the last three brain regions, HIP, THAL and STR, where NeuN+ profiles appear the similar across the regions in **Model 2CellPFC** and **Model 3CellPFC** (mean NeuN+ by **Model 2CellPFC** = 0.205 ± 0.109, 0.164 ± 0.0839, and 0.245 ± 0.0875, across and HIP, THAL and STR, respectively, and by **Model 3CellPFC**, mean NeuN+ = 0.0786 ± 0.0851, 0.027 ± 0.0664 and 0.093 ± 0.0622), and by extension the NeuN- proportion by **Model 2CellPFC** (mean = 0.813 ± 0.0961, 0.9 ± 0.0673 and 0.854 ± 0.0537), since proportions should sum roughly to 1. This is not true for the proportions of Sox10+ and Double- predicted by **Model 3CellPFC**, where the mean Double- proportion subtracted from the mean Sox10+ proportion is 0.056, -0.066 and 0.153 per brain region, i.e. the Sox10+ proportion is higher for HIP and STR only, and by a larger magnitude in the STR tissue investigated.

This Section demonstrates the utility of **Model 3CellPFC** over **Model 2CellPFC** when accounting for cellular composition in regional analysis of brain DNA methylation profiles. Where composition estimates were said to be equivalent using **Model 2CellPFC**, but differed in non-neuronal populations, as illuminated by **Model 3CellPFC**, the additional model granularity would be expected to improve the accuracy in adjusting for cell type composition in downstream analysis.

4.10.2 Model 3CellPFC confirms known cellular composition differences in Alzheimer's disease associated with neuropathology in PFC

Cell type composition is known to alter across AD progression, in which neuronal loss and a shift in microglial cell proportions can be observed (Prinz and Priller, 2017). Here, the cell type prediction profiles of samples across Braak stage from tissues OCC and glspfc from **Dataset BDR bulk** were compared between **Model 2CellPFC** and **Model 3CellPFC**.

Across both brain regions, NeuN+ was again systematically lower in **Model**

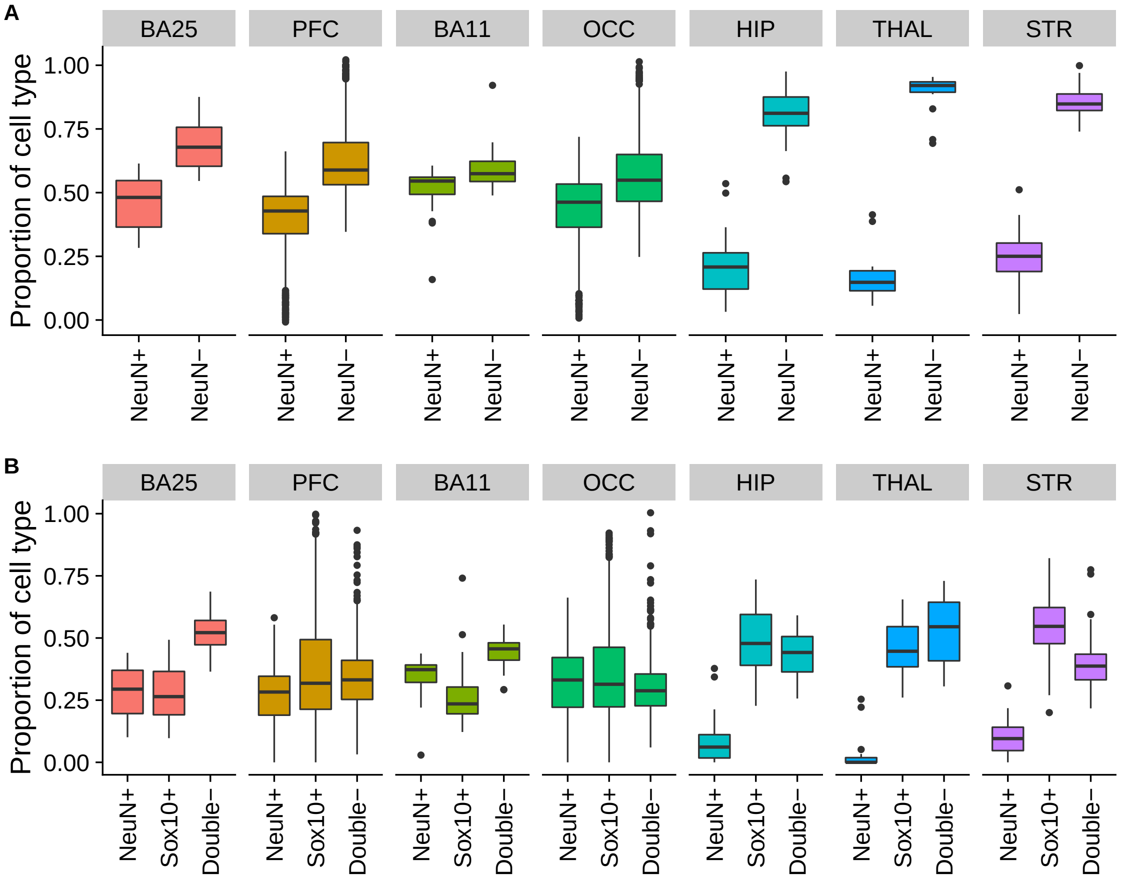


Figure 4.26: **A boxplot comparing the predicted cell type proportions by Model 2CellPFC and Model 3CellPFC across brain regions.** BA = Brodmann area, PFC = Prefrontal cortex, OCC = Occipital lobe, HIP = Hippocampus, THAL = Thalamus, STR = Striatum. Region BA9 from **Dataset Adult brain** was included in the PFC region, as BA9 is a subset of the PFC. **Model 2CellPFC** NeuN- proportions should be equivalent to the sum of **Model 3CellPFC** Sox10+ and Double- proportions.

3CellPFC compared to **Model 2CellPFC**, by 0.131 ± 0.0442. The cell type proportions in the OCC were observed not to significantly vary as Braak increased (**Figure 4.27**, **Table 4.5**). This is in line with literature, which states that AD progression, the OCC is only affected late in disease progression (Smith et al., 2001). In contrast, across the PFC, proportions of NeuN-, and the **Model 3CellPFC** counterpart Sox10+ and Double-, shift significantly as Braak stage increases (**Figure 4.28**, **Table 4.5**). In **Model 2CellPFC**, a negative and positive quadratic can be observed across Braak stage in proportions of NeuN+ and NeuN-, respectively. By increasing the granularity, **Model 3CellPFC** can detect linear shifts within the NeuN- population, with Sox10+ increasing significantly, and Double- decreasing significantly as Braak stage increases. **Model 3CellPFC** therefore improves the ability to more effectively account for cell type heterogeneity when investigating differential DNA methylation in AD, especially in brain regions for which cell type proportions shift with AD progression.

4.11 Discussion

4.11.1 Overview of results

Reference based deconvolution approaches are routinely utilised to derive cellular composition estimates from DNA methylation data generated on bulk tissue samples. These derived cell proportion estimates can be used to adjust for cellular heterogeneity between individuals in DNA methylation analyses of disease undertaken using bulk tissue. The current standard for reference based deconvolution of human cortex tissue is only able to distinguish between two cell types, neuronal and non-neuronal (i.e. glial) (Guintivano, Aryee and Kaminsky, 2013). The lack of model granularity means that differences within the proportion of glial or neuronal subtypes cannot be identified, which, given the cell type specific nature of DNA methylation, may have implications for the interpretation of DNA methylation data generated on bulk tissue. In this Chapter, a novel deconvolution model was established with the ability to distinguish between neurons, oligodendrocytes and the remaining brain cell types (primarily astrocytes and microglia), referred to as **Model 3CellPFC**.

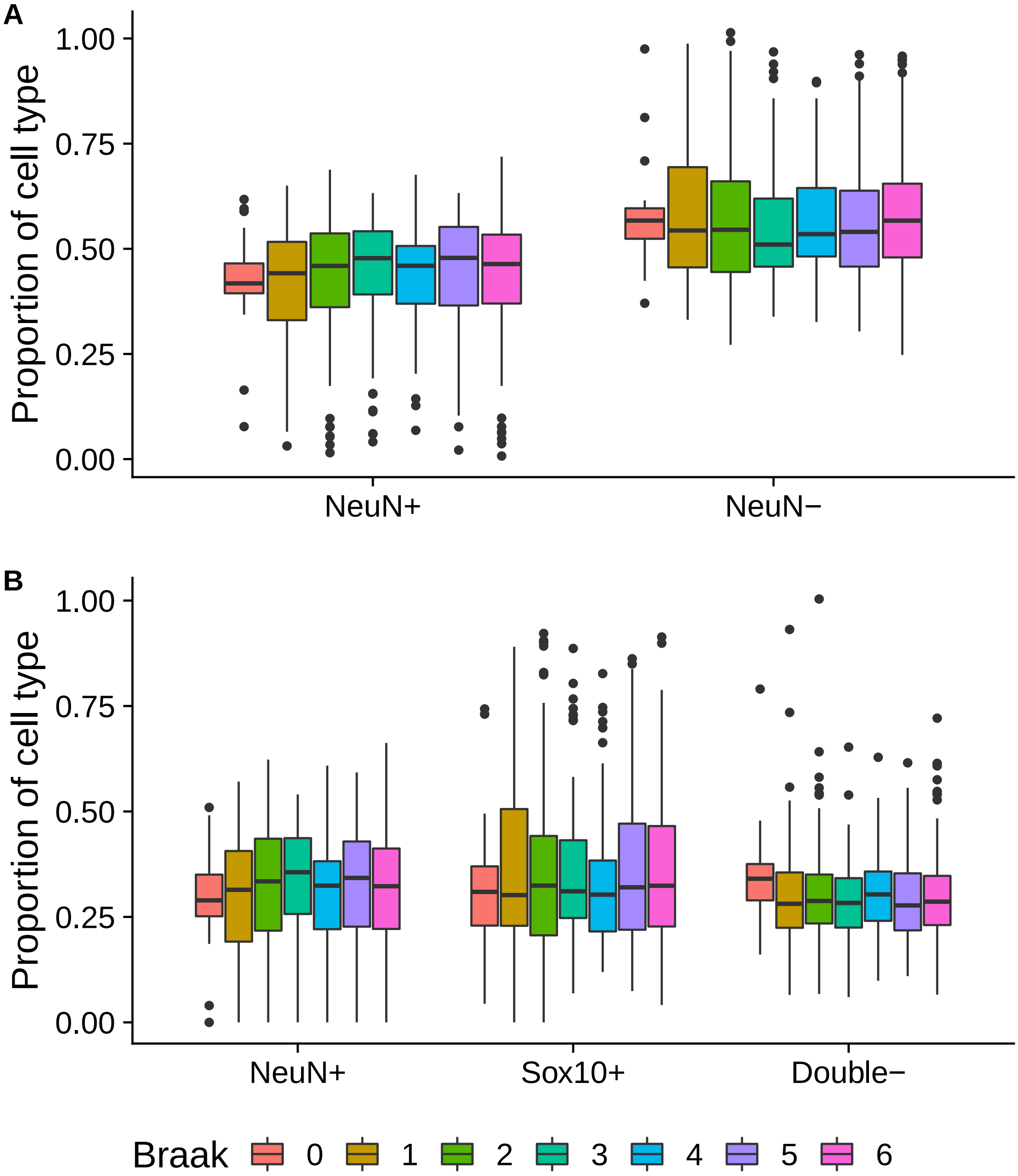


Figure 4.27: **A boxplot comparing the predicted cell type proportions from Model 2CellPFC and Model 3CellPFC across Braak stages in OCC tissue.** Cell type proportions remain consistent across Braak stage in the OCC. **Model 2CellPFC** and **Model 3CellPFC** were applied across Braak stages in **Dataset BDR bulk** PFC samples, and the proportion of each cell type plotted. A) **Model 2CellPFC** and B) **Model 3CellPFC**.

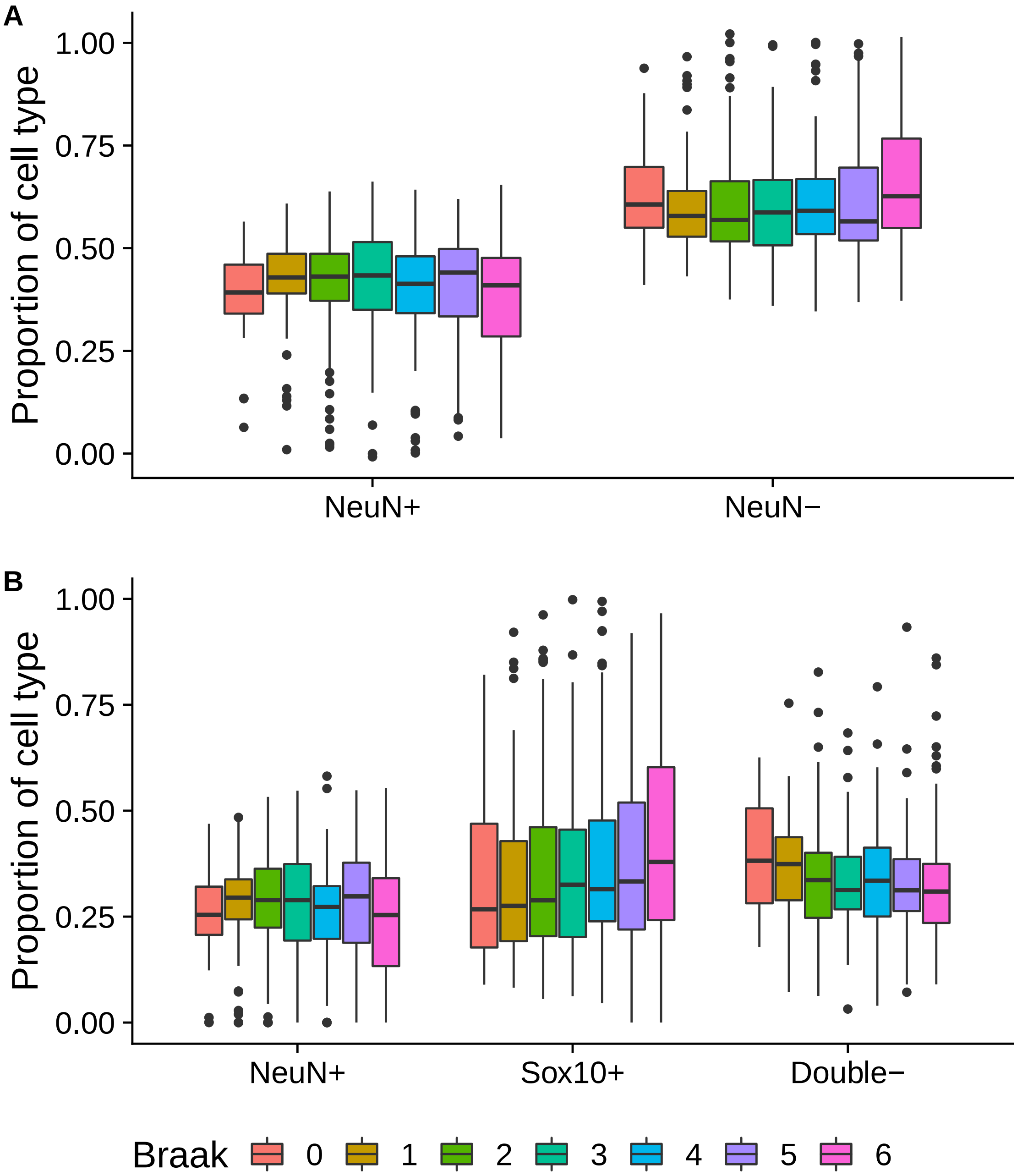


Figure 4.28: **A boxplot comparing the predicted cell type proportions from Model 2CellPFC and Model 3CellPFC across Braak stage in PFC tissue.** Cell type proportions shift across Braak stage in the PFC, which can be better identified using **Model 3CellPFC**. **Model 2CellPFC** and **Model 3CellPFC** were applied across Braak stages in **Dataset BDR bulk** PFC samples, and the proportion of each cell type plotted. A) **Model 2CellPFC** and B) **Model 3CellPFC**.

Table 4.5: **A summary of the linear modelling results testing the linear relationship between predicted cell type proportions and Braak stage in Dataset BDR bulk across PFC and OCC.** No significant difference is seen across Braak stage in the OCC, however, in PFC, NeuN-, and Sox10+ and Double- all significantly shift. The significance is stronger in **Model 3CellPFC** than **Model 2CellPFC**. Significance is represented by . = p $\leq$ 0.1, * = p $\leq$ 0.05, ** = p $\leq$ 0.01, *** = p $\leq$ 0.001.

Brain region	Model	Cell type	Model p value	Significance
OCC	Model 2CellPFC	NeuN+	0.168	
		NeuN-	0.842	
	Model 3CellPFC	NeuN+	0.455	
		Sox10+	0.831	
		Double-	0.505	
PFC	Model 2CellPFC	NeuN+	0.0766	.
		NeuN-	0.00392	**
	Model 3CellPFC	NeuN+	0.0253	*
		Sox10+	0.000207	***
		Double-	0.00465	**

Model 3CellPFC was shown to accurately predict cellular composition of cortical samples in a number of scenarios including in simulated bulk tissue DNA methylation data for which composition was known, and in independent NeuN+ purified samples. The model predictions were shown to correlate with predictions made using the two cell type deconvolution model, wit the key difference that the proportion NeuN+ were observed to be overestimated in **Model 2CellPFC**.

It is well established that differential DNA methylation patterns identified across highly heterogeneous tissue samples are liable to confounding and misinterpreted associations (Koestler et al., 2016; Adalsteinsson et al., 2012; Reinius et al., 2012; Koestler et al., 2012). **Model 3CellPFC** can be utilised to accurately estimate the cellular composition of individual cortex samples from DNA methylation data enabling the control of cellular heterogeneity in DNA methylation studies of disease.

The wider applicability of **Model 3CellPFC** was tested in this Chapter using Cetygo (Chapter 3); the model was shown to be applicable across all stages of AD progression and all cerebral cortex and subcortical areas tested. The wide applicability of the model means that fewer resources would need to be expended on the generation of subsequent reference datasets, with only very specific utility.

When compared to the two cell type PFC deconvolution model, **Model 3CellPFC** was shown to provide more information on cellular composition. For example, in the PFC as AD Braak stage increased, **Model 3CellPFC** was able to identify the underlying cellular composition changes in glial cells. When calculated with the two cell type model, glial proportions on observation appeared to have a positive quadratic association with increase in Braak stage, whereas in **Model 3CellPFC** it was shown to be the sum of a linear increase in oligodendrocytes and a linear decrease in other glial populations. Cellular composition estimates from reference based deconvolution are commonly included in EWAS analyses of diseases such as AD using bulk brain tissue DNA methylation data (Smith et al., 2020). Using the two cell type model when it does not successfully capture the shifts in glial subtypes, which have distinct DNA methylation profiles, can mean results may still be confounded by cellular heterogeneity between disease groups. As such, it would be recommended that **Model 3CellPFC** is utilised, in place of the two cell type model, to generate composition estimates for EWAS. This shift in glial subtypes may also be a feature of other disorders manifest in the brain and as such, **Model 3CellPFC** will have high utility given the models additional granularity.

4.11.2 Limitations and future work

A general caveat of the work presented in this Chapter is that only Houseman's algorithm was utilised to generate the novel deconvolution model. Houseman's algorithm is commonly used in the deconvolution of DNA methylation data, however, for the two cell type model an alternative algorithm, CETS (Guintivano, Aryee and Kaminsky, 2013), is also commonly utilised. No one algorithm performs best in all contexts (McGregor et al., 2016), however, due to its integration into commonly used QC R packages, *minfi* and *wateRmelon*, Houseman's algorithm was utilised here to benchmark the new deconvolution model. Further work will include exploring the alternative algorithms available for reference based deconvolution to optimise the use of the three cell type reference dataset.

Model 3CellPFC was found not to be applicable to DNA methylation data generated from fetal samples, reflecting the fact it was generated using reference data derived from adult donors. There are dramatic changes in DNA methylation across fetal brain development (Spiers et al., 2015) and many cells will not yet be mature and not be characterised by the expression of key cell type specific marker genes. For example, not all neurons express NeuN in early development, with only deep neurons within the cortical plate being marked at 19–22 pcw (Sarnat, Nochlin and Born, 1998). Similarly, **Model 3CellPFC** was shown not to be applicable across neural cell lines, SH-SY5Y and neurons derived from iPSCs. Cell lines have been shown to have different DNA methylation profiles to primary cells from human tissue, with DNA methylation acting as 'cell type memory' (Kim et al., 2011). Furthermore, iPSC-derived neurons are biologically young (Steg et al., 2021) and epigenetically similar to fetal neurons. As such, **Model 3CellPFC** is not applicable. Further work will include generating a cellular reference panel more appropriate for these sample types to be utilised in reference based deconvolution.

Despite being the most comprehensive brain reference deconvolution model developed to date, it still only predicts the proportions of three cell types, when in reality, both neuronal and glial cells contain many subtypes. For example, single cell technologies have been used to classify the profiles of different neural cell types in mice; single cell DNA methylation identified 41 major cell types, with 161 subtypes (Liu et al., 2021), single cell RNA-seq was utilised by Zeisel et al., 2015 uncovered 9 main brain cell types, with 47 sub-classes (Zeisel et al., 2015), and when investigating oligodendrocytes, Marques et al., 2016 identified 13 distinct subpopulations, of which 12 represented stages of the transition between oligodendrocyte precursor cells to mature oligodendrocytes (Marques et al., 2016). In the human brain, Lake et al., 2016 uncovered 16 neuronal subtypes using single nucleus RNA-seq (Lake et al., 2016), and Ziffra et al., 2021 12 distinct neural cell types in the forebrain using single cell ATAC-seq (Ziffra et al., 2021).

To further distinguish neuronal types, Kozlenkov et al., 2015 used FANS to isolate GABAergic and glutamatergic neurons using antibodies to both NeuN and SOX6, finding major DNA methylation differences between these subtypes (Kozlenkov et al., 2015). Future work should include integrating DNA methylation profiles from additional purified cell types with the aim of generating a more granular deconvolution model. With the constant development and optimisation of lab based techniques to characterise DNA methylation from purified nuclei (Policicchio et al., 2020a) and, more recently, single cells (Karemaker and Vermeulen, 2018), additional reference datasets for specific cell types are likely to become available in the near future. However, single cell DNA methylation methodologies are currently still in their infancy (especially when compared to single cell RNA-seq) and are yet to be applied extensively across human brain samples. When utilising more comprehensive cell type DNA methylation data to generate novel deconvolution models for the brain, careful consideration would be needed for the optimum number of cell types. This will most likely depend on the number of DNA methylation sites profiled (and therefore the choice of technology) that can distinguish between each cell type and the biological relevance of subtypes to the research aims. Without a clear distinction between cell types, cell type specific accuracy may be decreased, as seen in Section 3.4.1.3.

4.11.3 Conclusion

To the authors knowledge, this Chapter presents the first reference based cellular deconvolution model using DNA methylation data that distinguishes between neuronal, oligodendrocyte and remaining cell types in the brain. Results demonstrate that the model generated is highly accurate and allows users to gain additional insight into the cellular composition of bulk tissue samples when compared to the two cell type model currently used in epigenetic studies of the human brain. Subsequent composition estimates can be used to more extensively adjust for cellular heterogeneity in DNA methylation studies using brain tissue. Successfully accounting for cellular composition in DNA methylation association studies would reduce the likelihood of false positives associated with cell type proportions rather than the trait of interest, and as such would improve study replicability.

4.12 Additional methods

4.12.1 Data analysis

All analysis was carried out in R version 3.5.2. The packages used can be seen in **Table 6.1**.

5. Discussion

In the era of "open science", there is a general desire for better research, including an emphasis on data driven approaches to analysis with increased transparency on decisions made. To that end, this thesis has contributed three tools that enhance existing preprocessing pipelines for DNA methylation data and provide confidence in downstream analyses that arise from existing and novel datasets. This thesis has taken advantage of 41,029 DNA methylation datasets, obtained from publicly available repositories in addition to novel data generated by the Complex Disease Epigenomics group at the University of Exeter.

5.1 Key findings

5.1.1 Characterising the properties of bisulfite sequencing data: maximizing power and sensitivity to identify between-group differences in DNA methylation

Bisulfite sequencing (BS) is commonly used for the quantification of DNA methylation across large numbers of samples in epigenetic epidemiological studies (Gertz et al., 2011; Bundo et al., 2020; Tang et al., 2018; Fernández-Santiago et al., 2019; Rizzardi et al., 2019b). Methodologies for the optimal alignment of bisulfite sequencing data and the assessment of read quality have been given ample consideration (Andrews et al., 2010), however, data filtering processes utilised are not often consistent between studies or statistically derived. Despite the common usage of BS methods, little empirical work has been done to investigate the properties of experimental datasets and their applicability for addressing epidemiological research questions, especially in the context of sequencing read depth and sample size. Chapter 2 describes a comprehensive assessment of the impact of read depth on statistical power to identify DNA methylation differences between groups and establishes a framework for statistical analysis of BS studies.

The primary finding of Chapter 2 was that the relationship between study parameters and power is complex and interactive, with read depth and sample size both influencing the power to detect a between-group difference in DNA methylation at any specific site included in the final reduced representation bisulfite sequencing (RRBS) dataset. The relationship between both parameters and power is non-linear, with a plateau as a sufficient read depth or sample size is reached at which these factors no longer limit to power. Sample size was shown to have a larger impact on power than read depth. The simulation framework used was adapted into an interactive tool, POWEREDBiSeq, for use by the community. This tool estimates the power of BS data using simulations given the expected DNA methylation difference, minimum sample size threshold and read depth threshold across a given dataset. To the best of the authors knowledge, Chapter 2 describes the first approach for assessing statistical power across two-group BS studies.

5.1.2 Profiling the accuracy of reference based cellular deconvolution models

To address the issue of cellular heterogeneity in bulk tissue DNA methylation data, a computational solution, i.e. reference based deconvolution, has been widely adopted by researchers. However, the assumption that the reference datasets used to generate a deconvolution model will be applicable across all samples to which they are applied may be false. In Chapter 3, an error metric for reference based deconvolution algorithms, Cetygo, was established. This approach quantifies the difference between the input DNA methylation profile and the expected DNA methylation profile (calculated using the assumption cell type estimates are accurate). The inappropriate application of cellular deconvolution, either from issues of reference data applicability or poor bulk tissue DNA methylation data quality, can lead to inaccurate proportion estimates.

To the best of the authors knowledge, Chapter 3 presents the first error metric for reference based deconvolution where cellular composition is unknown. Results demonstrate that Cetygo has utility in assessing deconvolution accuracy and model applicability. Cetygo has been made publicly available and has been integrated into the commonly used DNA methylation quality control (QC) package *wateRmelon* for easy application during data preprocessing and cleaning, to allow for wider use and promote open science practices. Cetygo provides confidence about the extent that deconvolution predictions can be trusted for any given dataset, providing an indicator about how well cellular heterogeneity can be controlled for in epigenetic epidemiology using the deconvolution model applied.

5.1.3 Developing a DNA methylation reference based deconvolution model to quantify the relative abundance of three distinct neural cell types in the human cortex from bulk tissue DNA methylation data

A limitation of reference based cellular deconvolution algorithms for use on bulk tissue DNA methylation data is that they require the availability of reference data for the different cell types relevant to that specific tissue. Even when available, a reference dataset may not be comprehensive enough to estimate cellular composition for each of the major cell types within a tissue. The current standard for reference based deconvolution in the human brain can only distinguish between neuronal and and non-neuronal cell proportions (Guintivano, Aryee and Kaminsky, 2013). In Chapter 4, a novel three cell type deconvolution model was established with the ability to distinguish neurons, oligodendrocytes and the remaining brain cell types (primarily astrocytes and microglia), dubbed **Model 3CellPFC**. The lack of model granularity in current models means that shifts in the proportion of glial or neuronal subtypes cannot be identified, which, given the cell type specific nature of DNA methylation, may have implications on downstream DNA methylation analysis. **Model 3CellPFC** was validated across simulated, purified and bulk brain DNA methylation datasets, utilising Cetygo to assess model accuracy and wider applicability.

To the authors knowledge, Chapter 4 presents the first reference based cellular deconvolution model using DNA methylation data to distinguish between neuronal, oligodendrocyte and remaining cell types in the brain. The results demonstrate that this novel model is highly accurate and allows users to gain additional insight into the cellular composition of bulk brain samples when compared to the existing two cell type model. Subsequent composition estimates can be used to more comprehensively adjust for cellular heterogeneity in DNA methylation studies using brain tissue. Successfully accounting for cellular composition in methylomic studies of disease and pathology would reduce the likelihood of false positives associated with between-group differences in cell type proportions rather than the trait of interest and lead to more meaningful conclusions from studies performed using bulk brain tissue samples.

5.2 Strengths, limitations and future directions

The research in this thesis has been made possible by the open science practice of data sharing; each of the datasets utilised within this thesis were generated with another primary aim in mind. The ability to utilise preexisting datasets meant that fewer additional resources were required to undertake the analysis presented here. Furthermore, using open source data resulted in access to a much higher number of samples than could have been generated internally.

The tool presented in Chapter 2 will aid users to optimise the filtering of BS data. Doing so will negate the removal of DNA methylation sites which may have sufficient power but would have previously been removed due to more stringent filtering, increasing the potential number of true positive associations that can be uncovered.

Most DNA methylation samples generated using the Illumina DNA methylation array platform will be of high quality, with those that are not recognised through standard data QC pipelines, however, outliers may slip though the cracks. Experimental failures can occur and more so when handling manipulated samples (e.g. samples sorted by fluorescence-activated nuclear sorting (FANS)), reiterating the need for additional QC metrics and vigilant data preprocessing. While comparing between samples using methods such as principal component analysis (PCA) can be useful for identifying such outliers, it requires that a subset of samples are of high quality. Alternatively, in purified nuclei populations sorted from bulk brain tissue samples, Cetygo can be utilised in conjunction with an applicable deconvolution model (e.g. **Model 3CellPFC)** to assess the reliability of purified samples. This additional utility of Cetygo was not the error metric's primary intended purpose but a further more specific application.

A general caveat to the validation of Cetygo is that it was only applied across two tissues and using one deconvolution algorithm (i.e. Houseman's algorithm (Houseman et al., 2012)). Chapter 3 was intended as a proof of principle for Cetygo, and as such was not exhaustive, but future work will include the further characterisation of Cetygo across additional tissues and deconvolution algorithms. Given the framework for Cetygo established, there is no reason to assume that the utility of the error metric would not extend to other tissues and algorithms.

Future work will include reanalysing DNA methylation data utilising the novel tools developed in this thesis, namely, using the recommended data filtering to optimise statistical power in BS studies, using Cetygo to assess deconvolution predictions of heterogeneous tissues, and rerunning epigenome wide association study (EWAS) performed in brain tissue samples using **Model 3CellPFC** composition estimates as opposed to the current two cell type model. Cetygo can also be utilised to assess model quality of deconvolution models generated with other reference panels as they become available.

A general limitation of reference based deconvolution algorithms is that they only provide an estimated cellular proportion, rather than an actual measure of cellular abundance, and therefore cannot be used to inform users about the change in the actual quantity of specific cell types in a given bulk tissue sample. This may have implications when using composition estimates to adjust for cell type confounding in EWAS, especially since proportional changes will not be independent between cell types (as they will sum to one). Future work could include characterising how the cellular composition estimates used in EWAS influence the results of analyses, assessing the optimal way to account for cellular heterogeneity.

Reference based deconvolution is not the only type of prediction that genome wide DNA methylation profiles can be used to estimate. For example, epigenetic clocks, which estimate age (Hannum et al., 2013; Horvath, 2013; Shireby et al., 2020; Steg et al., 2021), and smoking status predictors (Bollepalli et al., 2019) have also been utilised when phenotypic data is not available. A similar framework to Cetygo could be utilised to assess the quality of other DNA methylation based predictions to allow users to explore their validity.

5.3 Conclusion

The thesis has shown that while the majority of DNAm datasets are of good quality, existing pipelines can be continually improved. To this end, the research presented in this thesis provides a suite of novel tools aimed at furthering the QC of DNA methylation data. Their adoption into existing EWAS pipelines will not only improve the general reproducibility of subsequent downstream analysis, but also encourage researchers to think more carefully about the importance of data QC and the steps needed to perform the best possible analyses of DNA methylation data.

6. Appendix

This Chapter summarises the R packages and datasets used in this thesis.

Table 6.1: A table compiling the R packages used in this thesis.

R package	Version	Citation
quadprog	1.5-8	(Berwin et al., 2019)
genefilter	1.64.0	(Gentleman et al., 2018)
tidyr	1.1.0	(Wickham and Henry, 2020)
ggplot2	3.3.2	(Wickham, 2016)
cowplot	1.0.0	(Wilke, n.d.)
scales	1.1.1	(Wickham and Seidel, 2020)
gdsfmt	1.18.1	(Zheng et al., 2017; Zheng et al., 2012)
forcats	0.5.0	(Wickham, 2020)
ggfortify	0.4.10	(Tang, Horikoshi and Li, 2016; Horikoshi and Tang, 2018)
reshape2	1.4.4	(Wickham, 2007)
viridis	0.5.1	(Garnier, 2018a)
viridisLite	0.3.0	(Garnier, 2018b)
ComplexHeatmap	1.20.0	(Gu, Eils and Schlesner, 2016)
minfi	1.28.4	(Jaffe and Irizarry, 2014)
wateRmelon	1.26.0	(Pidsley et al., 2013)
FlowSorted.Blood.EPIC	1.20.0	(Jaffe, 2018)
IlluminaHumanMethylation-450kanno.ilmn12.hg19	0.6.0	(Hansen, 2016)
dplyr	1.0.1	(Wickham et al., 2020)
bigmelon	1.8.0	(Gorrie-Stone et al., 2019)

Table 6.3: A summary of all datasets used and their dataset reference. If data is published, the GEO accession number is included. Mammalian array refers to the Illumina HorvathMammalianMethylChip40 BeadChip (Arneson et al., 2021). Where ‘Internal’ is Yes the data was generated by the Complex Disease Epigenomics Group of the University of Exeter.

Dataset reference	Use	Thesis section	Samples size	Tissue	Species	Platform	GEO accession number	Citation	Internal
mRRBS	RRBS quality control simulations and thresholding	2	125	Brain	Mouse	RRBS	GSE169234	Seiler Vellame et al., 2021	Yes
mArray	Validation of RRBS quality control thresholding	2	80	Brain	Mouse	Mam35	GSE169218	Seiler Vellame et al., 2021	Yes
Reinius	Development and validation of **Model 6CellBlood**	3	36	Purified blood	Human	HM450	NA	Reinius et al., 2012; Jaffe, 2018	No
EXTEND	Testing sex and age effects on Cetygo	3.8.2	1234	Blood	Human	EPIC	NA	NA	Yes
Understanding Society	Testing sex and age effects on Cetygo	3.8.2	1175	Blood	Human	EPIC	NA	Hannon et al., 2018	Yes

GEO	Applying Cetygo across tissues	3.5, 3.6, 3.7	32962	See **Table 7.4**	Human	HM450	See **Table 7.4**	NA	No
E-Risk	Applying Cetygo to purified blood samples	3.7	141	Blood	Human	EPIC	GSE103541	Hannon et al., 2021b	Yes
Guintivano	Reference and testing data for **Model 2CellPFC**	3, 4	58	FANS sorted brain	Human	HM450	GSE41826	Guintivano, Aryee and Kaminsky, 2013	No
Pai	Applying Cetygo to purified brain samples	3.7	100	FANS sorted brain	Human	EPIC	GSE112179	Pai et al., 2019	No
CortexFANS	Reference and testing data for **Model 3CellPFC**	4	430	FANS sorted brain and bulk nuclei	Human	EPIC	NA	NA	Yes

Fetal	Testing **Model 3CellPFC** applicability	4.9.3	114	Brain	Human	EPIC	NA	NA	Yes
BDR bulk	Testing **Model 3CellPFC** applicability	4.9, 4.10	1304	Brain	Human	EPIC	NA	Shireby et al., 2020	Yes
SH-SY5Y	Testing **Model 3CellPFC** applicability	4.9.4	156	SH-SY5Y cell line	Human	EPIC	NA	NA	Yes
iPSC	Testing **Model 3CellPFC** applicability	4.9.4	93	Neuronal iPSC	Human	EPIC	NA	NA	Yes
Adult brain	Testing **Model 3CellPFC** applicability	4.9, 4.10	656	Brain	Human	HM450	NA	NA	Yes
BDR purified	Testing Cetygo's relationship with median array intensity, assessing **Model 3CellPFC** accuracy	3.8.1, 4.6	107	FANS sorted brain	Human	EPIC	NA	NA	Yes
PPMI	Testing Cetygo's relationship with median array intensity	3.8.1	524	Blood	Human	EPIC	NA	Marek et al., 2011	No

IoP	Testing sex, age and ethnicity effects on Cetygo	3.8.2	800	Blood	Human	HM450	NA	Hannon et al., 2021a	Yes
EUGEI	Testing sex, age and ethnicity effects on Cetygo	3.8.2	934	Blood	Human	EPIC	NA	Hannon et al., 2021a	Yes

7. Supplementary

7.1 Additional file 2 for Chapter 2

Table 7.1: A summary of RRBS information on total number of reads aligned, unaligned ambiguously aligned, and total number of reads, as well as the number of methylated and unmethylated CpGs, CpH, and CHH's, and total number of cytosines.

File	Total Reads	Aligned Reads	Unaligned Reads	Ambiguously Aligned Reads	No Genomic Sequence
A17	41880099	30573692	3795509	7510898	0
A18	40840556	27905983	4128664	8805909	0
A19	38149633	28526024	3230934	6392675	0
A20	39923463	27666915	4008499	8248049	0
A21	41774444	28116006	4216628	9441810	0
A22	39522963	26578881	3977077	8967005	0
A23	38435933	26161761	3764542	8509630	0
A24	29293342	19740284	2984647	6568411	0
B17	49445903	34425016	4802916	10217971	0
B18	56432912	38651211	5454714	12326987	0
B20	39251636	27131520	3679661	8440455	0
B21	42804395	28838147	4125199	9841049	0
B22	40142801	26872788	4074140	9195873	0
B23	39281326	27043336	3613280	8624710	0
B24	38111425	25802755	3687679	8620991	0
C17	40845662	28858392	4106808	7880462	0
C18	43610153	30253616	4679667	8676870	0
C19	46982891	31601186	5170200	10211505	0
C20	43701908	29881533	4704394	9115981	0

C21	40149792	27509462	4211115	8429215	0
C22	34733075	23441779	3779411	7511885	0
C23	42000250	28927452	4471836	8600962	0
D17	60399990	40221525	5926540	14251925	0
D18	40258231	28007709	3699046	8551476	0
D19	41532105	30519981	3535381	7476743	0
D20	39529355	27670777	3666246	8192332	0
D23	43306624	29304944	4141819	9859861	0
E17	50569776	35375485	4775531	10418760	0
E18	49438556	35048186	4627471	9762899	0
E19	41611870	30630275	3686645	7294950	0
E20	48555567	33935710	4497139	10122718	0
E21	39216076	27428068	3763298	8024710	0
E22	39332856	27412307	3702352	8218197	0
F17	36058814	25791027	3715691	6552096	0
F18	46007974	31032876	5131600	9843498	0
F19	43423424	29786406	4621873	9015145	0
F20	51385356	35251741	5375082	10758533	0
F21	58098364	39911326	6467430	11719608	0
F22	36040200	25806904	3688198	6545098	0
F23	36057904	24581716	3880848	7595340	0
F24	40760452	28454831	4236981	8068640	0
G17	56991435	38268316	5738485	12984634	0
G18	40009124	27574638	3974730	8459756	0
G19	35664370	25489141	3659247	6515982	0
G20	40426822	28021106	4005234	8400482	0
G21	36554575	26544241	3348950	6661384	0
G22	33173424	23741348	3188779	6243297	0
G23	39357183	28742647	3433682	7180854	0
G24	36689068	24600208	3621071	8467789	0

H17	38944878	26584455	3858916	8501507	0
H18	40949091	27889527	4117805	8941759	0
H21	35747744	25386000	3633619	6728125	0
H22	32593940	24083208	2809405	5701327	0
H23	44268506	30073459	4331362	9863685	0
H24	54266000	36236129	5485124	12544747	0
I20	48427654	31601649	5360400	11465605	0
J17	40494488	28273878	3790196	8430414	0
J19	40155376	26604086	4374308	9176982	0
J20	48165426	31733559	5209242	11222625	0
J21	64753886	42811874	7299447	14642565	0
J22	30831563	20253154	3260535	7317874	0
J23	27370506	17920098	3044675	6405733	0
J24	43592380	28803722	4703515	10085143	0
K17	52293318	33024875	8461426	10807017	0
K18	44322853	28335662	6827679	9159512	0
K19	36737137	23448819	5590633	7697685	0
K20	40104398	25652609	6501976	7949813	0
K21	32369889	20775830	4814795	6779264	0
K23	34762909	22145741	5417682	7199486	0
K24	35301655	22612158	5320065	7369432	0
L17	52730639	33453378	9481956	9795305	0
L18	38983284	24662737	6700251	7620296	0
L19	37683930	23937536	6580803	7165591	0
L21	44918910	28262171	7791552	8865187	0
L22	36364278	23070712	6096658	7196908	0
L23	32357159	20478143	5464711	6414305	0
L24	40112977	25428061	6938864	7746052	0
M17	36560772	23533306	4508753	8518713	0
M18	38587833	24665293	4804586	9117954	0

M19	31448423	20000578	3916625	7531220	0
M20	40896857	26392694	4686817	9817346	0
M21	40684607	26215897	4695388	9773322	0
M22	31708133	20512022	3691651	7504460	0
M23	27232403	17666663	3211724	6354016	0
M24	33642028	21742961	3756040	8143027	0
N24	34616573	22128095	5510762	6977716	0
O18	38666144	25000247	5757222	7908675	0
O19	36475897	23452070	5416477	7607350	0
O20	45687986	29251620	6965362	9471004	0
O21	36930614	23708955	5610643	7611016	0
O22	33480166	21535284	4917711	7027171	0
O23	39364157	25358514	5645904	8359739	0
O24	40365543	25988227	5781749	8595567	0
P17	54391187	35046831	7430253	11914103	0
P18	39604500	25933028	5369193	8302279	0
P19	38396979	24796624	5381245	8219110	0
P20	42984319	27654840	5664124	9665355	0
P21	34968374	22607951	4691606	7668817	0
P22	36191048	23441433	4624068	8125547	0
P23	43204658	30955559	5071930	7177169	0
P24	45844641	29724020	5853897	10266724	0
Q17	50082775	31987334	9777682	8317759	0
Q18	41111464	25639180	7785282	7687002	0
Q19	37153947	23169305	7136320	6848322	0
Q20	48448135	30617324	8547080	9283731	0
Q21	41967258	26366404	7621142	7979712	0
Q22	39019644	24394370	7479370	7145904	0
Q23	36722207	23277320	6964438	6480449	0
Q24	46948240	29333258	8682789	8932193	0

S17	51905516	34072598	7550954	10281964	0
S18	48383714	31392247	6724426	10267041	0
S19	38277931	24780565	5239546	8257820	0
S20	41081378	26737922	5901654	8441802	0
S21	35177820	22806826	5162894	7208100	0
S22	32736906	21324700	4415513	6996693	0
S23	40209977	26154937	5446167	8608873	0
S24	38495536	25051742	5508137	7935657	0
T17	53703606	34927963	7946309	10829334	0
T18	50436064	32717544	7212502	10506018	0
T19	37989252	24654077	5373241	7961934	0
T20	48571382	31351343	6418446	10801593	0
T21	39896409	25897882	5598217	8400310	0
T22	36249811	23565243	4830538	7854030	0
T23	36313314	25075295	4740966	6497053	0
T24	39799603	25832402	5606111	8361090	0

File	Total Cs	Methylated CpGs	Unmethylated CpGs
A17	389838307	14882627	77568966
A18	342036958	18821083	56977682
A19	368448377	12263056	76607141
A20	347321060	16003011	63973258
A21	341242424	20379081	53634899
A22	322124399	19629372	49759134
A23	320674525	17363433	53419299
A24	239415233	14008573	38132075
B17	422263711	20749813	72708273
B18	469285887	26472652	75488667
B20	327318411	17915309	53558467
B21	343789697	22058184	49815171

B22	320308984	20657183	46467866
B23	328652925	17495810	53877375
B24	309617540	18340016	46958888
C17	364034916	16768310	66402202
C18	382105757	18898515	66379294
C19	396735887	22374655	64700938
C20	374061753	19067353	64145264
C21	346823363	17695025	59728369
C22	290458759	15995345	47102036
C23	364030094	17722925	63769917
D17	472161488	34032282	62388967
D18	336210766	18781835	54001325
D19	378896244	15778374	71161530
D20	342216041	16911587	60183285
D23	346568262	22983810	49146826
E17	428706419	22119147	71484585
E18	428742587	20734888	74687837
E19	385149284	14873090	74110772
E20	412990359	21215151	69459611
E21	333996088	16735983	57412984
E22	328890486	18366276	52139557
F17	298642917	15461956	50803677
F18	344732246	25281166	44944992
F19	337003751	21741494	49220727
F20	403968339	25948246	59550897
F21	460462144	27820734	71375093
F22	303105418	15008006	52227920
F23	275750522	18065987	39967906
F24	329674062	18533468	52851041
G17	457119460	30412696	63594038

G18	334162358	18650143	53143089
G19	319036638	13624022	59497952
G20	348059806	17641230	59560399
G21	334274866	13452697	63121261
G22	296222627	13000560	54010198
G23	363409105	14222885	69130389
G24	292891592	20191685	39455097
H17	317565574	18492234	49129418
H18	333829454	19427761	51465951
H21	300650895	14299120	52345924
H22	302511675	11601796	58658697
H23	363920961	22188071	55369513
H24	433334450	28511046	62419457
I20	386162300	25558529	55018367
J17	343001200	18105109	57482068
J19	299202997	21586177	38973905
J20	362217823	26996573	46522103
J21	489686813	33332047	69127219
J22	226059700	17502490	27541288
J23	199376092	15780118	24021500
J24	324887993	22879715	43549625
K17	326625039	30823075	31287385
K18	279699662	26398960	26628444
K19	235711468	22344313	22311152
K20	252730442	22272722	27050482
K21	204226489	20110882	17889595
K23	221581328	20214741	22492553
K24	225726327	21713601	20665220
L17	328458762	27136128	39106752
L18	242292092	23033334	23251973

L19	232301659	21426925	23678112
L21	274213836	27304855	24616481
L22	226495084	21811744	21073536
L23	198541871	19470677	18017055
L24	247373704	22949274	24621254
M17	256955680	20690990	30337618
M18	268788815	22795125	28812653
M19	219492675	17919295	25109654
M20	298737933	23566478	35329140
M21	294430007	22784600	35387773
M22	227186422	18381871	25994689
M23	195946650	15197699	23851743
M24	241622992	20542992	25642217
N24	219175721	18794446	24217698
O18	258940376	21689148	29429369
O19	247323422	20421267	28997118
O20	303355253	26027340	33875505
O21	244418046	20568025	28138385
O22	223676222	19882156	23455305
O23	262701433	23864651	26745612
O24	271857611	24217852	28495619
P17	383936364	27965387	52133313
P18	286296227	18936159	42162047
P19	269542503	19947530	36235417
P20	306047281	21857273	41985588
P21	245675891	18261120	32664946
P22	255645496	21129423	29967056
P23	367875147	15083869	69234423
P24	323526671	24959376	40825163
Q17	322247875	23455008	44120253

Q18	250013750	22850362	25384837
Q19	225579064	19921812	24651661
Q20	298198834	28008875	29286577
Q21	256468809	23609509	26058436
Q22	239431441	20399203	27342060
Q23	230538726	18436174	28683474
Q24	285636107	26426304	28997330
S17	361369387	25043455	50318660
S18	324553595	28081365	35064749
S19	259110083	21968563	28856419
S20	277550643	23210112	31719756
S21	234364281	18996446	27889622
S22	220960197	19195910	23621819
S23	269417591	23404993	29067157
S24	261959916	21072345	31427820
T17	373437419	25449000	53142643
T18	336507301	28247462	38423880
T19	251677674	20672183	29430042
T20	329243666	27901436	36482777
T21	267037446	21658141	31728676
T22	244310133	22232518	24379342
T23	279749478	14676328	46564779
T24	269134926	21067999	33295837

File	Methylated CpHs	Unmethylated CpHs	Methylated CHHs	Unmethylated CHHs
A17	562670	87481786	1212961	208129297
A18	612287	75962615	1310482	188352809
A19	500544	82890395	1080435	195106806
A20	577035	77713646	1185012	187869098
A21	650601	75501991	1405429	189670423

A22	605281	71392440	1286783	179451389
A23	561595	71350908	1163971	176815319
A24	441677	53031041	923893	132877974
B17	704177	93803260	1583159	232715029
B18	858774	104079909	1923803	260462082
B20	678100	72506974	1500004	181159557
B21	657144	75909828	1504274	193845096
B22	628086	70521890	1445905	180588054
B23	581787	72848664	1301427	182547862
B24	567278	68434046	1282890	174034422
C17	701164	81003395	1532098	197627747
C18	778014	84764285	1660717	209624932
C19	862960	87785213	1891638	219120483
C20	789810	82816246	1664211	205578869
C21	729723	77049951	1543280	190077015
C22	643721	64014946	1382123	161320588
C23	745009	80720430	1553834	199517979
D17	1093727	103852803	2554774	268238935
D18	670955	74591234	1495970	186669447
D19	676042	84882665	1495925	204901708
D20	680230	76199669	1486337	186754933
D23	773011	76532684	1775737	195356194
E17	840014	94936316	1900762	237425595
E18	824258	95092547	1844873	235558184
E19	673648	85989249	1519918	207982607
E20	798974	91613408	1765951	228137264
E21	644465	73939515	1438157	183824984
E22	676819	72606920	1510574	183590340
F17	537166	65737470	1284992	164817656
F18	758138	74799232	1895188	197053530

F19	679815	73764279	1627591	189969845
F20	828330	88613567	1967395	227059904
F21	903310	101259618	2098595	257004794
F22	548162	66910261	1264282	167146787
F23	568971	60151130	1357106	155639422
F24	635936	72517245	1503704	183632668
G17	892359	100914282	2079302	259226783
G18	604949	74005850	1344363	186413964
G19	493609	71236860	1088320	173095875
G20	601024	77479096	1342786	191435271
G21	507351	74677921	1142328	181373308
G22	469632	66039414	1042607	161660216
G23	539273	81513661	1218965	196783932
G24	585364	64708391	1387616	166563439
H17	657251	69945015	1453862	177887794
H18	676890	73647483	1519206	187092163
H21	462561	66546255	1054647	165942388
H22	528575	67797381	1177203	162748023
H23	675429	80723760	1552315	203411873
H24	868351	95689953	2013014	243832629
I20	932293	85034820	2031337	217586954
J17	691700	76120016	1559711	189042596
J19	585245	65297286	1427286	171333098
J20	720735	79355072	1676369	206946971
J21	937846	107312058	2162630	276815013
J22	448463	49228684	1054403	130284372
J23	414397	43415185	995855	114749037
J24	620818	70973149	1420380	185444306
K17	793917	68632178	1899888	193188596
K18	685333	58827247	1643809	165515869

K19	596669	49869106	1442103	139148125
K20	538138	53150262	1214113	148504725
K21	520573	42900638	1279124	121525677
K23	520150	46785269	1211270	130357345
K24	557010	47608343	1331260	133850893
L17	669516	68720742	1518677	191306947
L18	573942	50873133	1355891	143203819
L19	520348	48561147	1209958	136905169
L21	689125	57245985	1667327	162690063
L22	523805	47620864	1200746	134264389
L23	492757	41532309	1185226	117843847
L24	563977	51783388	1314981	146140830
M17	622991	55890350	1389078	148024653
M18	683609	58390342	1502197	156604889
M19	562224	47780779	1235897	126884826
M20	712035	65755394	1475214	171899672
M21	741084	64695027	1621428	169200095
M22	582027	49688178	1302483	131237174
M23	491399	42865731	1092056	112448022
M24	636159	52986460	1430862	140384302
N24	458918	46090019	1034387	128580253
O18	590111	55138559	1336021	150757168
O19	574202	52904719	1313476	143112640
O20	698605	64595787	1574045	176583971
O21	555798	51860445	1248396	142046997
O22	523757	47668982	1182830	130963192
O23	682883	55877350	1635205	153895732
O24	617776	58171593	1359709	158995062
P17	850729	83280750	1839556	217866629
P18	586745	62183346	1267772	161160158

P19	594754	58294061	1280861	153189880
P20	663577	66780707	1394949	173365187
P21	545425	53166599	1177660	139860141
P22	587446	55608880	1262565	147090126
P23	581985	81970249	1247507	199757114
P24	743009	70292761	1632949	185073413
Q17	594478	68036680	1352551	184688905
Q18	542261	52227693	1252453	147756144
Q19	474926	47095604	1089536	132345525
Q20	641598	62517329	1447231	176297224
Q21	567671	53757129	1321605	151154459
Q22	488792	50115273	1098403	139987710
Q23	451995	48462057	1021830	133483196
Q24	646798	59778437	1516669	168270569
S17	767780	77532400	1712611	205994481
S18	806667	69336609	1846738	189417467
S19	663212	55466162	1530966	150624761
S20	665765	59374722	1495277	161085011
S21	547368	49802442	1224331	135904072
S22	537781	47215676	1216477	129172534
S23	682111	57414986	1556642	157291702
S24	596048	56032095	1296559	151535049
T17	792761	80390206	1776087	211886722
T18	778586	71745840	1748894	195562639
T19	576275	53502056	1294985	146202133
T20	815023	70752149	1825847	191466434
T21	624123	56892673	1409108	154724725
T22	617613	52339542	1409348	143331770
T23	516381	61317443	1150588	155523959
T24	612025	57751579	1354742	155052744

7.2 Supplementary for Chapter 3

Table 7.4: A summary of **Dataset GEO**, compiled by collaborators (Tyler Gorrie-Stone, Leonard Schalkwyk)

GEO Accession	Sample Size	Tissue
GSE100386	46	Lymph Node
GSE100503	13	Breast
GSE100561	12	Blood
GSE100653	18	Breast
GSE101443	8	Breast
GSE101641	48	Epithelial
GSE101673	24	Epithelial
GSE101764	149	mucosa
GSE101961	121	Breast
GSE102119	146	Ovary
GSE102177	36	Blood
GSE102504	25	Blood
GSE102970	48	Sperm
GSE103010	5	Bone Marrow
GSE103186	191	Intestines, Digestive System
GSE103413	28	Buccal, Liver, Breast, Chorion
GSE103768	57	Adipose
GSE103911	58	T Cells
GSE104087	40	Epithelial
GSE104287	48	Blood
GSE104376	45	Blood
GSE104471	72	Blood, Epithelial
GSE104778	72	Blood
GSE104812	48	Blood
GSE105018	1658	Blood

GSE105109	384	Brain
GSE105123	108	Blood
GSE105260	9	Kidney
GSE106089	84	Placenta
GSE106556	20	Rectum, Colon
GSE107038	40	Liver
GSE107226	12	Lung
GSE107353	113	Blood
GSE107459	127	Blood
GSE107737	24	Blood
GSE108058	30	Sperm
GSE108143	2	Embryonic Stem Cells
GSE108423	21	Blood
GSE108462	181	plasma
GSE108576	48	Breast, Lung
GSE108785	6	Blood
GSE109042	53	Buccal
GSE109430	36	Blood
GSE109446	58	Epithelial
GSE109914	113	Blood
GSE110607	104	Blood, B Cells, T Cells, Granulocyes
GSE111223	259	Saliva
GSE111396	61	Fibroblast
GSE112047	16	Prostate
GSE112314	22	Saliva
GSE112696	12	T Cells
GSE112877	96	Fibroblast
GSE114753	156	Sperm
GSE114935	47	Blood
GSE115797	48	Skin

GSE115920	6	Sperm
GSE116300	44	Blood
GSE116754	2	Embryonic Stem Cells
GSE116924	1	Blood
GSE117050	38	T Cells
GSE118260	20	Intestines, Saliva
GSE118570	43	T Cells
GSE119684	45	Blood
GSE120062	38	Placenta
GSE120250	88	Placenta
GSE120307	34	Blood
GSE121633	480	Blood
GSE124366	215	Buccal, Blood
GSE124565	22	Neutrophils
GSE125105	699	Blood
GSE125895	68	Brain
GSE126017	54	Sperm
GSE127824	24	Blood
GSE128068	112	Blood
GSE38240	11	Bone, Lymph Node, Prostate
GSE42861	689	Blood
GSE43976	95	Blood
GSE47915	4	Prostate
GSE48472	42	Blood, Liver, Muscle, Pancreas, Adipose, Buccal, Saliva
GSE49149	196	Pancreas
GSE49618	9	T Cells, B Cells, Blood
GSE51032	845	Blood
GSE52025	62	Skin
GSE54375	2	Lung

GSE55491	24	Blood
GSE56596	5	Nervous System/Spinal Cord
GSE57992	1	Fibroblast
GSE59065	97	Blood
GSE59524	24	Adipose
GSE60655	36	Muscle
GSE61107	48	Brain
GSE61278	110	Liver
GSE61441	92	Kidney
GSE61454	269	Liver, Adipose, Muscle
GSE61461	17	Fibroblast, induced Pluripotent Stem Cells, Embryonic Stem Cells
GSE61496	312	Blood
GSE62219	60	Blood
GSE62929	12	Blood
GSE63106	62	Cartilage
GSE63409	24	T Cells
GSE63669	23	Brain
GSE63695	97	Cartilage
GSE65057	24	Liver
GSE65078	4	Fibroblast
GSE65163	72	Epithelial
GSE65638	16	Blood
GSE66077	6	Embryonic Stem Cells
GSE66210	60	Blood, Chorion
GSE66313	40	Breast
GSE66351	190	Brain, Buccal
GSE66552	43	Blood
GSE66562	23	NK, T Cells

GSE66836	19	Lung
GSE67097	6	Skin
GSE67170	89	T Cells, Blood
GSE67393	117	Blood
GSE67419	24	Buccal
GSE67444	70	Blood
GSE67485	11	Liver, Intestines
GSE67733	12	Fibroblast
GSE68379	1000	Blood, urogenital_system, Lung, Digestive System, Nervous System/Spinal Cord, Skin, Kidney, Thyroid, Pancreas, Breast, Bone
GSE68777	40	Blood
GSE68825	135	Lung
GSE68838	267	Colon
GSE69502	127	Kidney, Nervous System/Spinal Cord, Brain, Muscle
GSE69852	6	Liver
GSE70460	18	Brain
GSE70478	38	Blood, Neutrophils
GSE70737	11	Embryonic Stem Cells, Neuron, Epithelial, Fibroblast
GSE71678	343	Placenta
GSE71719	46	Placenta
GSE71955	135	T Cells
GSE72021	171	serous
GSE72120	72	Saliva

GSE72354	34	T Cells
GSE72364	12	T Cells
GSE72556	96	Saliva
GSE72867	69	Blood
GSE73103	355	Blood
GSE73115	180	Blood
GSE73412	74	Blood
GSE73549	92	Prostate, Lymph Node
GSE73626	18	Cartilage
GSE73745	12	Saliva
GSE74193	675	Brain
GSE74214	18	Breast
GSE74432	115	Blood
GSE74548	174	Blood
GSE74738	21	Blood, Placenta, Chorion
GSE74877	7	Fibroblast, Bone Marrow, Breast, Blood
GSE75196	24	Placenta
GSE75248	335	Placenta
GSE75405	24	Blood
GSE75434	9	Buccal
GSE75546	12	Rectum
GSE75704	166	Brain
GSE76372	9	Lymph Node, induced Pluripotent Stem Cells
GSE76503	48	Blood
GSE77135	21	Fibroblast
GSE77353	23	Brain, Buccal
GSE77797	18	Blood
GSE77954	48	Colon, Intestines, Rectum, Liver

GSE77965	22	Colon, Blood
GSE79064	18	Brain, Liver
GSE79100	93	Kidney
GSE79185	45	Breast, Nervous System/Spinal Cord, Colon, Lung, Ovary, Prostate, Kidney
GSE79257	111	Blood
GSE79329	34	Blood
GSE79366	2	Epithelial
GSE79695	44	Bone Marrow
GSE80261	216	Buccal
GSE80377	4	Blood
GSE80468	60	Blood
GSE80794	12	Breast
GSE80969	4	Blood, Fibroblast
GSE81224	10	Epithelial, Ovary
GSE81438	20	Ovary
GSE82084	36	Blood
GSE83691	4	Liver
GSE83842	12	Lung
GSE83944	48	Neutrophils
GSE84743	48	Intestines
GSE85042	71	Blood
GSE85210	253	Blood
GSE85506	47	Blood
GSE85566	115	Epithelial
GSE85647	23	Blood
GSE85828	75	Blastocyst, Fibroblast
GSE86258	14	Prostate
GSE86402	6	Colon

GSE86829	5	Blood
GSE87056	31	Liver
GSE87095	122	B Cells
GSE87571	732	Blood
GSE87582	21	Blood, T Cells
GSE87640	125	Blood
GSE87648	384	Blood
GSE87655	6	Muscle
GSE88824	83	Neutrophils, T Cells, NK, B Cells, Blood
GSE88883	100	Breast
GSE89251	136	T Cells
GSE89474	10	Blood
GSE89852	37	Liver
GSE89925	3	Fibroblast
GSE90871	24	Brain
GSE92909	6	Breast
GSE93208	19	Chorion
GSE93933	126	Blood
GSE93963	6	Intestines
GSE94326	4	Brain
GSE95486	24	Blood
GSE95761	6	Embryonic Stem Cells
GSE95816	1	Fibroblast
GSE97362	235	Blood
GSE97529	36	Bone
GSE97784	24	Buccal
GSE98056	69	Blood
GSE98203	88	Neuron
GSE98224	48	Placenta

GSE98938	7	Brain, Chorion
GSE99511	68	Cervix/Vagina
GSE99553	84	Digestive System
GSE99624	48	Blood
GSE99716	7	Embryonic Stem Cells, Lung, induced Pluripotent Stem Cells, Placenta
GSE99863	257	Blood

Bibliography

References that are used in Chapter 2 are not included within this bibliography because it was published and contains its own bibliography.

B. T. Adalsteinsson, H. Gudnason, T. Aspelund, T. B. Harris, L. J. Launer, G. Eiriksdottir, A. V. Smith and V. Gudnason (Oct. 2012). 'Heterogeneity in White Blood Cells Has Potential to Confound DNA Methylation Measurements'. In: *PLOS ONE* 7.10, e46705.

A. Adan, G. Nel Alizada, Y. Mur Kiraz, Y. Baran and A. Nalbant (2017). 'Flow cytometry: basic principles and applications'. In: *Crit Rev Biotechnol* 37.2, pp. 163–176.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter (2002). *Molecular Biology of the Cell*. 6th. Garland Science, pp. 373–380.

V. G. Allfrey, R. Faulkner and A. E. Mirsky (May 1964). 'ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS'. In: *Proceedings of the National Academy of Sciences* 51.5, pp. 786–794.

S. Andrews, F. Krueger, A. Segonds-Pichon, L. Biggins, C. Krueger and S. Wingett (2010). *FastQC*.

A. Anguita-Ruiz, M. Bustos-Aibar, J. Plaza-Díaz, A. Mendez-Gutierrez, J. Alcalá-Fdez, C. M. Aguilera and F. J. Ruiz-Ojeda (Mar. 2021). 'Omics Approaches in Adipose Tissue and Skeletal Muscle Addressing the Role of Extracellular Matrix in Obesity and Metabolic Dysfunction'. In: *International Journal of Molecular Sciences 2021, Vol. 22, Page 2756* 22.5, p. 2756.

A. Arneson, A. Haghani, M. J. Thompson, M. Pellegrini, S. Bin Kwon, H. Vu, C. Z. Li, A. T. Lu, B. Barnes, K. D. Hansen et al. (Jan. 2021). 'A mammalian methylation array for profiling methylation levels at conserved sequences'. In: *bioRxiv*, p. 2021.01.07.425637.

C. Aschwanden (Aug. 2015). *Science Isn't Broken*.

S. Avazzadeh, J. M. Baena, C. Keighron, Y. Feller-Sanchez and L. R. Quinlan (Mar. 2021). 'Modelling Parkinson's Disease: iPSCs towards Better Understanding of Human Pathology'. In: *Brain Sciences 2021, Vol. 11, Page 373* 11.3, p. 373.

F. A. Azevedo, L. R. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. Ferretti, R. E. Leite, W. J. Filho, R. Lent and S. Herculano-Houzel (Apr. 2009). 'Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain'. In: *Journal of Comparative Neurology* 513.5, pp. 532–541.

M. Bachman, S. Uribe-Lewis, X. Yang, M. Williams, A. Murrell and S. Balasubramanian (Sept. 2014). '5-Hydroxymethylcytosine is a predominantly stable DNA modification'. In: *Nature Chemistry 2014 6:12* 6.12, pp. 1049–1055.

K. M. Bakulski, J. I. Feinberg, S. V. Andrews, J. Yang, S. Brown, S. L. McKenney, F. Witter, J. Walston, A. P. Feinberg and M. D. Fallin (May 2016a). 'DNA methylation of cord blood cell types: Applications for mixed cell birth studies'. In: *Epigenetics* 11.5, pp. 354–362.

K. M. Bakulski, A. Halladay, V. W. Hu, J. Mill and M. D. Fallin (Sept. 2016b). 'Epigenetic Research in Neuropsychiatric Disorders: the "Tissue Issue"'. In: *Current Behavioral Neuroscience Reports* 3.3, pp. 264–274.

C. S. v. Bartheld, J. Bahney and S. Herculano-Houzel (Dec. 2016). 'The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting'. In: *Journal of Comparative Neurology* 524.18, pp. 3865–3895.

D. Beck, M. Ben Maamar and M. K. Skinner (May 2021). 'Genome-wide CpG density and DNA methylation analysis method (MeDIP, RRBS, and WGBS) comparisons'. In: *Epigenetics*, pp. 1–13.

J. T. Bell and T. D. Spector (Oct. 2012). 'DNA methylation studies using twins: what are they telling us?' In: *Genome Biology 2012 13:10* 13.10, pp. 1–6.

R. Berwin, A. Turlach, A. Weingessel and C. Moler (2019). *quadprog: Functions to Solve Quadratic Programming Problems*.

S. K. Bhattacharya, S. Ramchandani, N. Cervoni and M. Szyf (Feb. 1999). 'A mammalian protein with specific demethylase activity for mCpG DNA'. In: *Nature* 397.6720, pp. 579–583.

M. Bibikova, J. Le, B. Barnes, S. Saedinia-Melnyk, L. Zhou, R. Shen and K. L. Gunderson (Oct. 2009). 'Genome-wide DNA methylation profiling using Infinium® assay'. In: *Epigenomics* 1.1, pp. 177–200.

J. L. Biedler, S. Roffler-Tarlov, M. Schachner and L. S. Freedman (1978). 'Multiple Neurotransmitter Synthesis by Human Neuroblastoma Cell Lines and Clones'. In: *Cancer Research* 38.11 Part 1.

A. Bird, M. Taggart, M. Frommer, O. J. Miller and D. Macleod (Jan. 1985). 'A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA'. In: *Cell* 40.1, pp. 91–99.

S. Bollepalli, T. Korhonen, J. Kaprio, S. Anders and M. Ollikainen (Oct. 2019). 'EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data'. In: *Epigenomics* 11.13, pp. 1469–1486.

M. J. Bonder, R. Luijk, D. V. Zhernakova, M. Moed, P. Deelen, M. Vermaat, M. van Iterson, F. van Dijk, M. van Galen, J. Bot et al. (Dec. 2016). 'Disease variants alter transcription factor levels and methylation of their binding sites'. In: *Nature Genetics 2016 49:1* 49.1, pp. 131–138.

H. Braak and E. Braak (1991). 'Neuropathological stageing of Alzheimer-related changes'. In: *Acta Neuropathol* 82, pp. 239–259.

N. J. Bray, S. Kapur and J. Price (2012). 'Investigating schizophrenia in a dish: possibilities, potential and limitations'. In: *World Psychiatry* 11.3, p. 153.

O. D. Buhule, R. L. Minster, N. L. Hawley, M. Medvedovic, G. Sun, S. Viali, R. Deka, S. T. McGarvey and D. E. Weeks (2014). 'Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale'. In: *Frontiers in Genetics* 0.SEP, p. 354.

M. Bundo, J. Ueda, Y. Nakachi, K. Kasai, T. Kato, K. Iwamoto and T. Kato (Dec. 2020). 'Decreased DNA methylation at promoters and gene-specific neuronal hypermethylation in the prefrontal cortex of patients with bipolar disorder'. In: *medRxiv*, p. 2020.12.10.20246405.

H.-M. Byun, K. D. Siegmund, F. Pan, D. J. Weisenberger, G. Kanel, P. W. Laird and A. S. Yang (Dec. 2009). 'Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns'. In: *Human Molecular Genetics* 18.24, pp. 4808–4817.

W. Cao, H. Lee, W. Wu, A. Zaman, S. McCorkle, M. Yan, J. Chen, Q. Xing, N. Sinnott-Armstrong, H. Xu et al. (July 2020). 'Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma'. In: *Nature Communications 2020 11:1* 11.1, pp. 1–19.

J. E. Castillo-Fernandez, T. D. Spector and J. T. Bell (Aug. 2014). 'Epigenetics of discordant monozygotic twins: Implications for disease'. In: *Genome Medicine* 6.7, pp. 1–16.

H. Cedar and Y. Bergman (May 2009). 'Linking DNA methylation and histone modification: patterns and paradigms'. In: *Nature Reviews Genetics 2009 10:5* 10.5, pp. 295–304.

X.-S. Chen, N. Huang, N. Michael and L. Xiao (Dec. 2015). 'Advancements in the Underlying Pathogenesis of Schizophrenia: Implications of DNA Methylation in Glial Cells'. In: *Frontiers in Cellular Neuroscience* 9.DEC, pp. 1–8.

Y. A. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson and R. Weksberg (2013). 'Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray'. In: *Epigenetics* 8.2, pp. 203–209.

F. Ciccarone, S. Tagliatesta, P. Caiafa and M. Zampieri (Sept. 2018). 'DNA methylation dynamics in aging: how far are we from understanding the mechanisms?' In: *Mechanisms of Ageing and Development* 174, pp. 3–17.

S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini and S. E. Jacobsen (Mar. 2008). 'Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning'. In: *Nature* 452.7184, pp. 215–219.

A. L. Comes, D. Czamara, K. Adorjan, H. Anderson-Schmidt, T. F. Andlauer, M. Budde, K. Gade, M. Hake, J. L. Kalman, S. Papiol et al. (Dec. 2020). 'The role of environmental stress and DNA methylation in the longitudinal course of bipolar disorder'. In: *International Journal of Bipolar Disorders* 8.1, p. 9.

R. Constantinescu, A. T. Constantinescu, H. Reichmann and B. Janetzky (2007). 'Neuronal differentiation and long-term culture of the human neuroblastoma line SH-SY5Y'. In: *Journal of Neural Transmission, Supplementa* 72, pp. 17–28.

D. N. Cooper, M. H. Taggart and A. P. Bird (Feb. 1983). 'Unmethlated domains in vertebrate DNA'. In: *Nucleic Acids Research* 11.3, pp. 647–658.

D. Cotter, L. Hudson and S. Landau (Aug. 2005). 'Evidence for orbitofrontal pathology in bipolar disorder and major depression, but not in schizophrenia'. In: *Bipolar disorders* 7.4, pp. 358–369.

M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp et al. (Dec. 2010). 'Histone H3K27ac separates active from poised enhancers and predicts developmental state'. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.50, pp. 21931–21936.

A. Crujeiras, A. Diaz-Lagares, J. Moreno-Navarrete, J. Sandoval, D. Hervas, A. Gomez, W. Ricart, F. Casanueva, M. Esteller and J. Fernandez-Real (Dec. 2016). 'Genome-wide DNA methylation pattern in visceral adipose tissue differentiates insulin-resistant from insulin-sensitive obese subjects'. In: *Translational Research* 178, pp. 13–24.

CUSABIO (n.d.). *Four Common Histone Modifications*.

C. Davegårdh, S. García-Calzón, K. Bacos and C. Ling (Aug. 2018). *DNA methylation in the pathogenesis of type 2 diabetes in humans*.

C. Davegårdh, E. Hall Wedin, C. Broholm, T. I. Henriksen, M. Pedersen, B. K. Pedersen, C. Scheele and C. Ling (Jan. 2019). 'Sex influences DNA methylation and gene expression in human skeletal muscle myoblasts and myotubes'. In: *Stem Cell Research and Therapy* 10.1, pp. 1–17.

P. L. De Jager, G. Srivastava, K. Lunnon, J. Burgess, L. C. Schalkwyk, L. Yu, M. L. Eaton, B. T. Keenan, J. Ernst, C. McCabe et al. (Sept. 2014). 'Alzheimer's disease: Early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci'. In: *Nature Neuroscience* 17.9, pp. 1156–1163.

A. M. Deaton and A. Bird (May 2011). 'CpG islands and the regulation of transcription'. In: *Genes & Development* 25.10, pp. 1010–1022.

S. Dedeurwaerder, M. Defrance, M. Bizet, E. Calonne, G. Bontempi and F. Fuks (Nov. 2014). 'A comprehensive overview of Infinium HumanMethylation450 data processing'. In: *Briefings in Bioinformatics* 15.6, pp. 929–941.

R. Dolmetsch and D. H. Geschwind (June 2011). 'The Human Brain in a Dish: The Promise of iPSC-Derived Neurons'. In: *Cell* 145.6, pp. 831–834.

J. van Dongen, M. J. Bonder, K. F. Dekkers, M. G. Nivard, M. van Iterson, G. Willemsen, M. Beekman, A. van der Spek, J. B. J. van Meurs, L. Franke et al. (Mar. 2018). 'DNA methylation signatures of educational attainment'. In: *npj Science of Learning 2018 3:1* 3.1, pp. 1–14.

J. van Dongen, S. D. Gordon, A. F. McRae, V. V. Odintsova, H. Mbarek, C. E. Breeze, K. Sugden, S. Lundgren, J. E. Castillo-Fernandez, E. Hannon et al. (Sept. 2021). 'Identical twins carry a persistent epigenetic signature of early genome programming'. In: *Nature Communications 2021 12:1* 12.1, pp. 1–14.

M. K. Donovan, A. D'Antonio-Chronowska, M. D'Antonio and K. A. Frazer (Feb. 2020). 'Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants'. In: *Nature Communications 2020 11:1* 11.1, pp. 1–14.

A. W. Drong, G. Nicholson, Å. K. Hedman, E. Meduri, E. Grundberg, K. S. Small, S.-Y. Shin, J. T. Bell, F. Karpe, N. Soranzo et al. (Feb. 2013). 'The Presence of Methylation Quantitative Trait Loci Indicates a Direct Genetic Influence on the Level of DNA Methylation in Adipose Tissue'. In: *PLOS ONE* 8.2, e55923.

I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul et al. (Sept. 2012). 'An integrated encyclopedia of DNA elements in the human genome'. In: *Nature 2012 489:7414* 489.7414, pp. 57–74.

S. J. Engle, L. Blaha and R. J. Kleiman (Nov. 2018). 'Best Practices for Translational Disease Modeling Using Human iPSC-Derived Neurons'. In: *Neuron* 100.4, pp. 783–797.

EpiGentek (n.d.). *Advanced Epigenetic Overview of Histone Modifications*.

J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne et al. (Mar. 2011). 'Mapping and analysis of chromatin state dynamics in nine human cell types'. In: *Nature 2011 473:7345* 473.7345, pp. 43–49.

C. Erö, M.-O. Gewaltig, D. Keller and H. Markram (Nov. 2018). 'A Cell Atlas for the Mouse Brain'. In: *Frontiers in Neuroinformatics* 0, p. 84.

X. Fan, J. Dong, S. Zhong, Y. Wei, Q. Wu, L. Yan, J. Yong, L. Sun, X. Wang, Y. Zhao et al. (June 2018). 'Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis'. In: *Cell Research 2018 28:7* 28.7, pp. 730–745.

R. Fernández-Santiago, A. Merkel, G. Castellano, S. Heath, Á. Raya, E. Tolosa, M.-J. Martí, A. Consiglio and M. Ezquerra (July 2019). 'Whole-genome DNA hyper-methylation in iPSC-derived dopaminergic neurons from Parkinson's disease patients'. In: *Clinical Epigenetics 2019 11:1* 11.1, pp. 1–7.

J. P. Fortin, A. Labbe, M. Lemire, B. W. Zanke, T. J. Hudson, E. J. Fertig, C. M. Greenwood and K. D. Hansen (Dec. 2014). 'Functional normalization of 450k methylation array data improves replication in large cancer studies'. In: *Genome Biology* 15.11, p. 503.

M. F. Fraga, E. Ballestar, M. F. Paz, S. Ropero, F. Setien, M. L. Ballestar, D. Heine-Suñer, J. C. Cigudosa, M. Urioste, J. Benitez et al. (July 2005). 'Epigenetic differences arise during the lifetime of monozygotic twins'. In: *Proceedings of the National Academy of Sciences* 102.30, pp. 10604–10609.

H. B. Fraser, L. L. Lam, S. M. Neumann and M. S. Kobor (Feb. 2012). 'Population-specificity of human DNA methylation'. In: *Genome Biology 2012 13:2* 13.2, pp. 1–12.

L. P. Freedman, I. M. Cockburn and T. S. Simcoe (June 2015). 'The economics of reproducibility in preclinical research'. In: *PLoS Biology* 13.6, pp. 1–9.

M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy and C. L. Paul (Mar. 1992). 'A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.' In: *Proceedings of the National Academy of Sciences* 89.5, pp. 1827–1831.

K. Fu, G. Bonora and M. Pellegrini (Mar. 2020). 'Interactions between core histone marks and DNA methyltransferases predict DNA methylation patterns observed in human cells and tissues'. In: *Epigenetics* 15.3, pp. 272–282.

J. A. Gagnon-Bartsch and T. P. Speed (July 2012). 'Using control genes to correct for unwanted variation in microarray data'. In: *Biostatistics* 13.3, pp. 539–552.

E. R. Gamazon, J. A. Badner, L. Cheng, C. Zhang, D. Zhang, N. J. Cox, E. S. Gershon, J. R. Kelsoe, T. A. Greenwood, C. M. Nievergelt et al. (Jan. 2012). 'Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants'. In: *Molecular Psychiatry 2013 18:3* 18.3, pp. 340–346.

Q. Gan, T. Yoshida, O. G. McDonald and G. K. Owens (Jan. 2007). 'Concise Review: Epigenetic Mechanisms Contribute to Pluripotency and Cell Lineage Determination of Embryonic Stem Cells'. In: *STEM CELLS* 25.1, pp. 2–9.

M. Gardiner-Garden and M. Frommer (July 1987). 'CpG Islands in vertebrate genomes'. In: *Journal of Molecular Biology* 196.2, pp. 261–282.

S. Garnier (2018a). *viridis: Default Color Maps from 'matplotlib'.*

S. Garnier (2018b). *viridisLite: Default Color Maps from 'matplotlib' (Lite Version).*

G. Gasparoni, S. Bultmann, P. Lutsik, T. F. Kraus, S. Sordon, J. Vlcek, V. Dietinger, M. Steinmaurer, M. Haider, C. B. Mulholland et al. (July 2018). 'DNA methylation analysis on purified neurons and glia dissects age and Alzheimer's disease-specific changes in the human cortex'. In: *Epigenetics and Chromatin* 11.1, pp. 1–19.

T. R. Gaunt, H. A. Shihab, G. Hemani, J. L. Min, G. Woodward, O. Lyttleton, J. Zheng, A. Duggirala, W. L. McArdle, K. Ho et al. (Mar. 2016). 'Systematic identification of genetic influences on methylation across the human life course'. In: *Genome Biology 2016 17:1* 17.1, pp. 1–14.

R. Gentleman, V. Carey, W. Huber and F. Hahne (2018). *genefilter: genefilter: methods for filtering genes from high-throughput experiments*.

J. Gertz, K. E. Varley, T. E. Reddy, K. M. Bowling, F. Pauli, S. L. Parker, K. S. Kucera, H. F. Willard and R. M. Myers (Aug. 2011). 'Analysis of DNA Methylation in a Three-Generation Family Reveals Widespread Genetic Influence on Epigenetic Regulation'. In: *PLoS Genetics* 7.8. Ed. by W. A. Bickmore, e1002228.

J. R. Gibbs, M. P. v. d. Brug, D. G. Hernandez, B. J. Traynor, M. A. Nalls, S.-L. Lai, S. Arepalli, A. Dillman, I. P. Rafferty, J. Troncoso et al. (May 2010). 'Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain'. In: *PLOS Genetics* 6.5, e1000952.

D. Globisch, M. Münzel, M. Müller, S. Michalakis, M. Wagner, S. Koch, T. Brückl, M. Biel and T. Carell (2010). 'Tissue Distribution of 5-Hydroxymethylcytosine and Search for Active Demethylation Intermediates'. In: *PLOS ONE* 5.12, e15367.

T. Gómez-Isla, R. Hollister, H. West, S. Mui, J. H. Growdon, R. C. Petersen, J. E. Parisi and B. T. Hyman (Jan. 1997). 'Neuronal loss correlates with but exceeds neurofibrillary tangles in Alzheimer's disease'. In: *Annals of Neurology* 41.1, pp. 17–24.

T. J. Gorrie-Stone, M. C. Smart, A. Saffari, K. Malki, E. Hannon, J. Burrage, J. Mill, M. Kumari and L. C. Schalkwyk (Mar. 2019). 'Bigmelon: tools for analysing large DNA methylation datasets'. In: *Bioinformatics* 35.6. Ed. by J. Kelso, pp. 981–986.

E. Grundberg, E. Meduri, J. K. Sandling, Å. K. Hedman, S. Keildson, A. Buil, S. Busche, W. Yuan, J. Nisbet, M. Sekowska et al. (Nov. 2013). 'Global Analysis of DNA Methylation Variation in Adipose Tissue from Twins Reveals Links to Disease-Associated Variants in Distal Regulatory Elements'. In: *The American Journal of Human Genetics* 93.5, pp. 876–890.

H. Gu, C. Bock, T. S. Mikkelsen, N. Jäger, Z. D. Smith, E. Tomazou, A. Gnirke, E. S. Lander and A. Meissner (Feb. 2010). 'Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution'. In: *Nature Methods* 7.2, pp. 133–136.

H. Gu, Z. D. Smith, C. Bock, P. Boyle, A. Gnirke and A. Meissner (Apr. 2011). 'Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling.' In: *Nature protocols* 6.4, pp. 468–81.

Z. Gu, R. Eils and M. Schlesner (2016). 'Complex heatmaps reveal patterns and correlations in multidimensional genomic data'. In: *Bioinformatics*.

F. Guénard, A. Tchernof, Y. Deshaies, S. Biron, O. Lescelleur, L. Biertho, S. Marceau, L. Pérusse and M.-C. Vohl (June 2017). 'Genetic regulation of differentially methylated genes in visceral adipose tissue of severely obese men discordant for the metabolic syndrome'. In: *Translational Research* 184, pp. 1–11.

J. Guintivano, M. J. Aryee and Z. A. Kaminsky (Mar. 2013). 'A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression.' In: *Epigenetics* 8.3, pp. 290–302.

J. U. Guo, Y. Su, J. H. Shin, J. Shin, H. Li, B. Xie, C. Zhong, S. Hu, T. Le, G. Fan et al. (Dec. 2013). 'Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain'. In: *Nature Neuroscience 2013 17:2* 17.2, pp. 215–222.

E. Hannon, E. L. Dempster, G. Mansell, J. Burrage, N. Bass, M. M. Bohlken, A. Corvin, C. J. Curtis, D. Dempster, M. Di Forti et al. (Feb. 2021a). 'Dna methylation meta-analysis reveals cellular alterations in psychosis and markers of treatment-resistant schizophrenia'. In: *eLife* 10, pp. 1–53.

E. Hannon, E. Dempster, J. Viana, J. Burrage, A. R. Smith, R. Macdonald, D. St Clair, C. Mustard, G. Breen, S. Therman et al. (Aug. 2016). 'An integrated genetic-epigenetic analysis of schizophrenia: Evidence for co-localization of genetic associations and differential DNA methylation'. In: *Genome Biology* 17.1.

E. Hannon, T. J. Gorrie-Stone, M. C. Smart, J. Burrage, A. Hughes, Y. Bao, M. Kumari, L. C. Schalkwyk and J. Mill (Nov. 2018). 'Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits'. In: *The American Journal of Human Genetics* 103.5, pp. 654–665.

E. Hannon, K. Lunnon, L. Schalkwyk and J. Mill (Nov. 2015a). 'Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes'. In: *Epigenetics* 10.11, pp. 1024–1032.

E. Hannon, G. Mansell, E. Walker, M. F. Nabais, J. Burrage, A. Kepa, J. Best-Lane, A. Rose, S. Heck, T. E. Moffitt et al. (Mar. 2021b). 'Assessing the co-variability of DNA methylation across peripheral cells and tissues: Implications for the interpretation of findings in epigenetic epidemiology'. In: *PLOS Genetics* 17.3, e1009443.

E. Hannon, H. Spiers, J. Viana, R. Pidsley, J. Burrage, T. M. Murphy, C. Troakes, G. Turecki, M. C. O'Donovan, L. C. Schalkwyk et al. (Nov. 2015b). 'Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci'. In: *Nature Neuroscience 2016 19:1* 19.1, pp. 48–54.

G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao et al. (Jan. 2013). 'Genome-wide methylation profiles reveal quantitative views of human aging rates.' In: *Molecular cell* 49.2, pp. 359–367.

K. D. Hansen, B. Langmead and R. A. Irizarry (Oct. 2012). 'BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions'. In: *Genome Biology* 13.10, R83.

K. D. Hansen (2016). *IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays*.

H. Hayatsu (2008). *Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis - A personal account*.

B. He, C. Zhang, X. Zhang, Y. Fan, H. Zeng, J. Liu, H. Meng, D. Bai, J. Peng, Q. Zhang et al. (July 2021). 'Tissue-specific 5-hydroxymethylcytosine landscape of the human genome'. In: *Nature Communications 2021 12:1* 12.1, pp. 1–12.

M. Horikoshi and Y. Tang (2018). *ggfortify: Data Visualization Tools for Statistical Analysis Results*.

S. Horvath (Dec. 2013). 'DNA methylation age of human tissues and cell types'. In: *Genome Biology 2013 14:10* 14.10, pp. 1–20.

S. Horvath, V. Mah, A. T. Lu, J. S. Woo, O.-W. Choi, A. J. Jasinska, J. A. Riancho, S. Tung, N. S. Coles, J. Braun et al. (2015). 'The cerebellum ages slowly according to the epigenetic clock.' In: *Aging* 7.5, pp. 294–306.

S. Horvath, J. A. Zoller, A. Haghani, A. J. Jasinska, K. Raj, C. E. Breeze, J. Ernst, K. L. Vaughan and J. A. Mattison (Sept. 2021). 'Epigenetic clock and methylation studies in the rhesus macaque'. In: *GeroScience 2021*, pp. 1–13.

E. A. Houseman, M. L. Kile, D. C. Christiani, T. A. Ince, K. T. Kelsey and C. J. Marsit (Dec. 2016). 'Reference-free deconvolution of DNA methylation data and mediation by cell composition effects'. In: *BMC Bioinformatics* 17.1, p. 259.

E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke and K. T. Kelsey (Dec. 2012). 'DNA methylation arrays as surrogate measures of cell mixture distribution'. In: *BMC Bioinformatics* 13.1, p. 86.

E. A. Houseman, J. Molitor and C. J. Marsit (May 2014). 'Reference-free cell mixture adjustments in analysis of DNA methylation data'. In: *Bioinformatics* 30.10, pp. 1431–1439.

K. Y. Y. Huang, Y. J. Huang and P. Y. Chen (Apr. 2018). 'BS-Seeker3: Ultrafast pipeline for bisulfite sequencing'. In: *BMC Bioinformatics* 19.1, p. 111.

A. Hughes, M. Smart, T. Gorrie-Stone, E. Hannon, J. Mill, Y. Bao, J. Burrage, L. Schalkwyk and M. Kumari (Nov. 2018). 'Socioeconomic Position and DNA Methylation Age Acceleration Across the Life Course'. In: *American Journal of Epidemiology* 187.11, pp. 2346–2354.

S. F. Ibrahim and G. Van Den Engh (June 2007). 'Flow cytometry and cell sorting'. In: *Advances in biochemical engineering/biotechnology* 106, pp. 19–39.

R. S. Illingworth, U. Gruenewald-Schneider, S. Webb, A. R. W. Kerr, K. D. James, D. J. Turner, C. Smith, D. J. Harrison, R. Andrews and A. P. Bird (Sept. 2010). 'Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome'. In: *PLOS Genetics* 6.9, e1001134.

Illumina (n.d.[a]). *Illumina Support*.

Illumina (n.d.[b]). *Infinium Methylation Coverage*.

J.-P. Issa (Jan. 2014). 'Aging and epigenetic drift: a vicious cycle'. In: *The Journal of Clinical Investigation* 124.1, pp. 24–29.

A. E. Jaffe (2018). *FlowSorted.Blood.450k: Illumina HumanMethylation data on sorted blood cell populations*.

A. E. Jaffe and R. A. Irizarry (Feb. 2014). 'Accounting for cellular heterogeneity is critical in epigenome-wide association studies'. In: *Genome Biology* 15.2, R31.

S. Jäkel and L. Dimou (Feb. 2017). 'Glial Cells and Their Function in the Adult Brain: A Journey through the History of Their Ablation'. In: *Frontiers in Cellular Neuroscience* 0, p. 24.

H. S. Jang, W. J. Shin, J. E. Lee and J. T. Do (May 2017). 'CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function'. In: *Genes 2017, Vol. 8, Page 148* 8.6, p. 148.

B. R. Jasny, N. Wigginton, M. McNutt, T. Bubela, S. Buck, R. Cook-Deegan, T. Gardner, B. Hanson, C. Hustad, V. Kiermer et al. (Aug. 2017). 'Fostering reproducibility in industry-academia research Sharing can pose challenges for collaborations'. In: *Science* 357.6353, pp. 759–761.

A. Joglekar, A. Prjibelski, A. Mahfouz, P. Collier, S. Lin, A. K. Schlusche, J. Marrocco, S. R. Williams, B. Haase, A. Hayes et al. (Jan. 2021). 'A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain'. In: *Nature Communications 2021 12:1* 12.1, pp. 1–16.

P. A. Jones (July 2012). *Functions of DNA methylation: Islands, start sites, gene bodies and beyond*.

B. R. Joubert, J. F. Felix, P. Yousefi, K. M. Bakulski, A. C. Just, C. Breton, S. E. Reese, C. A. Markunas, R. C. Richmond, C. J. Xu et al. (Apr. 2016). 'DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis'. In: *The American Journal of Human Genetics* 98.4, pp. 680–696.

I. D. Karemaker and M. Vermeulen (Sept. 2018). *Single-Cell DNA Methylation Profiling: Technologies and Biological Applications*.

N. Kazmi, H. R. Elliott, K. Burrows, T. Tillin, A. D. Hughes, N. Chaturvedi, T. R. Gaunt and C. L. Relton (Jan. 2020). 'Associations between high blood pressure and DNA methylation'. In: *PLOS ONE* 15.1, e0227728.

N. J. Kessler, R. A. Waterland, A. M. Prentice and M. J. Silver (2018). *Establishment of environmentally sensitive DNA methylation states in the very early human embryo*. Tech. rep.

K. Kim, R. Zhao, A. Doi, K. Ng, J. Unternaehrer, P. Cahan, H. Hongguang, Y. H. Loh, M. J. Aryee, M. W. Lensch et al. (Dec. 2011). 'Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells'. In: *Nature Biotechnology* 29.12, pp. 1117–1119.

H. K. Kimelberg (2010). 'Functions of Mature Mammalian Astrocytes: A Current View'. In:

A. Klungland and A. B. Robertson (June 2017). 'Oxidized C5-methyl cytosine bases in DNA: 5-Hydroxymethylcytosine; 5-formylcytosine; and 5-carboxycytosine'. In: *Free Radical Biology and Medicine* 107, pp. 62–68.

A. Koch, S. C. Joosten, Z. Feng, T. C. De Ruijter, M. X. Draht, V. Melotte, K. M. Smits, J. Veeck, J. G. Herman, L. V. Neste et al. (July 2018). *Analysis of DNA methylation in cancer: Location revisited.*

D. C. Koestler, M. J. Jones, J. Usset, B. C. Christensen, R. A. Butler, M. S. Kobor, J. K. Wiencke and K. T. Kelsey (Mar. 2016). 'Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL)'. In: *BMC Bioinformatics* 17.1, pp. 1–21.

D. C. Koestler, C. J. Marsit, B. C. Christensen, W. Accomando, S. M. Langevin, E. A. Houseman, H. H. Nelson, M. R. Karagas, J. K. Wiencke and K. T. Kelsey (Aug. 2012). 'Peripheral Blood Immune Cell Methylation Profiles Are Associated with Nonhematopoietic Cancers'. In: *Cancer Epidemiology and Prevention Biomarkers* 21.8, pp. 1293–1302.

J. Kovalevich and D. L. Abstract (n.d.). 'Considerations for the Use of SH-SY5Y Neuroblastoma Cells in Neurobiology'. In: *Methods in Molecular Biology* 1078 ().

A. Kozlenkov, M. Wang, P. Roussos, S. Rudchenko, M. Barbu, M. Bibikova, B. Klotzle, A. J. Dwork, B. Zhang, Y. L. Hurd et al. (Nov. 2015). 'Substantial DNA methylation differences between two major neuronal subtypes in human brain'. In: *Nucleic Acids Research* 44.6, pp. 2593–2612.

S. E. Kreps and D. L. Kriner (Oct. 2020). 'Model uncertainty, political contestation, and public trust in science: Evidence from the COVID-19 pandemic'. In: *Science Advances* 6.43, eabd4563.

F. Krueger and S. R. Andrews (June 2011). 'Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications'. In: *Bioinformatics* 27.11, pp. 1571–1572.

F. Krueger, B. Kreck, A. Franke and S. R. Andrews (Feb. 2012). 'DNA methylome analysis using short bisulfite sequencing data'. In: *Nature Methods* 9.2, pp. 145–151.

S. Kumar, V. Chinnusamy and T. Mohapatra (Dec. 2018). 'Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond'. In: *Frontiers in Genetics* 0, p. 640.

C. Ladd-Acosta, J. Pevsner, S. Sabunciyan, R. H. Yolken, M. J. Webster, T. Dinkins, P. A. Callinan, J. B. Fan, J. B. Potash and A. P. Feinberg (Dec. 2007). 'DNA methylation signatures within the human brain'. In: *American Journal of Human Genetics* 81.6, pp. 1304–1315.

P. W. Laird (Feb. 2010). 'Principles and challenges of genome-wide DNA methylation analysis'. In: *Nature Reviews Genetics 2010 11:3* 11.3, pp. 191–203.

B. B. Lake, R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, D. Gao, H.-L. Fung, S. Chen et al. (June 2016). 'Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain'. In: *Science* 352.6293, pp. 1586–1590.

S. M. Langevin, E. A. Houseman, B. C. Christensen, J. K. Wiencke, H. H. Nelson, M. R. Karagas, C. J. Marsit and K. T. Kelsey (July 2011). 'The influence of aging, environmental exposures and local sequence features on the variation of DNA methylation in blood'. In: *Epigenetics* 6.7, pp. 908–919.

F. Larsen, G. Gundersen, R. Lopez and H. Prydz (Aug. 1992). 'CpG islands as gene markers in the human genome'. In: *Genomics* 13.4, pp. 1095–1107.

L. Laurent, E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. K. Sung, I. Rigoutsos, J. Loring et al. (Mar. 2010). 'Dynamic changes in the human methylome during differentiation'. In: *Genome Research* 20.3, pp. 320–331.

M. K. Lee, Y. Hong, S. Y. Kim, S. J. London and W. J. Kim (Sept. 2016). 'DNA methylation and smoking in Korean adults: epigenome-wide association study'. In: *Clinical Epigenetics* 8.1, pp. 1–17.

J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe and J. D. Storey (Mar. 2012). 'The SVA package for removing batch effects and other unwanted variation in high-throughput experiments'. In: *Bioinformatics* 28.6, pp. 882–883.

J. T. Leek and J. D. Storey (Sept. 2007). 'Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis'. In: *PLOS Genetics* 3.9, e161.

S. Lent, H. Xu, L. Wang, Z. Wang, C. Sarnowski, M.-F. Hivert and J. Dupuis (Sept. 2018). 'Comparison of novel and existing methods for detecting differentially methylated regions'. In: *BMC Genetics 2018 19:1* 19.1, pp. 27–31.

K. M. Lenz and L. H. Nelson (Apr. 2018). 'Microglia and Beyond: Innate Immune Cells As Regulators of Brain Development and Behavioral Function'. In: *Frontiers in Immunology* 0.APR, p. 698.

M. E. Levine, A. T. Lu, A. Quach, B. H. Chen, T. L. Assimes, S. Bandinelli, L. Hou, A. A. Baccarelli, J. D. Stewart, Y. Li et al. (2018). 'An epigenetic biomarker of aging for lifespan and healthspan'. In: *Aging* 10.4, pp. 573–591.

J. J. Levy, A. J. Titus, C. L. Petersen, Y. Chen, L. A. Salas and B. C. Christensen (Mar. 2020). 'MethylNet: an automated and modular deep learning approach for DNA methylation analysis'. In: *BMC Bioinformatics 2020 21:1* 21.1, pp. 1–15.

E. Li, T. H. Bestor and R. Jaenisch (June 1992). 'Targeted mutation of the DNA methyltransferase gene results in embryonic lethality'. In: *Cell* 69.6, pp. 915–926.

H. Li, A. Sharma, K. Luo, Z. S. Qin, X. Sun and H. Liu (June 2020). 'DeconPeaker, a Deconvolution Model to Identify Cell Types Based on Chromatin Accessibility in ATAC-Seq Data of Mixture Samples'. In: *Frontiers in Genetics* 11, p. 392.

Y. Li, H. Zheng, Q. Wang, C. Zhou, L. Wei, X. Liu, W. Zhang, Y. Zhang, Z. Du, X. Wang et al. (Feb. 2018). 'Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys'. In: *Genome Biology* 19.1.

E. Libertini, S. C. Heath, R. A. Hamoudi, M. Gut, M. J. Ziller, A. Czyz, V. Ruotti, H. G. Stunnenberg, M. Frontini, W. H. Ouwehand et al. (June 2016). 'Information recovery from low coverage whole-genome bisulfite sequencing'. In: *Nature Communications* 7.1, pp. 1–7.

R. Lister and J. R. Ecker (June 2009). 'Finding the fifth base: Genome-wide sequencing of cytosine methylation'. In: *Genome Research* 19.6, pp. 959–966.

R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo et al. (Nov. 2009). 'Human DNA methylomes at base resolution show widespread epigenomic differences'. In: *Nature* 462.7271, pp. 315–322.

R. Lister, M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman et al. (Feb. 2011). 'Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells'. In: *Nature 2010 471:7336* 471.7336, pp. 68–73.

C. Liu, R. E. Marioni, A. K. Hedman, L. Pfeiffer, P. C. Tsai, L. M. Reynolds, A. C. Just, Q. Duan, C. G. Boer, T. Tanaka et al. (Feb. 2018). 'A DNA methylation biomarker of alcohol consumption'. In: *Molecular Psychiatry* 23.2, pp. 422–433.

H. Liu, J. Zhou, W. Tian, C. Luo, A. Bartlett, A. Aldridge, J. Lucero, J. K. Osteen, J. R. Nery, H. Chen et al. (Oct. 2021). 'DNA methylation atlas of the mouse brain at single-cell resolution'. In: *Nature 2021 598:7879* 598.7879, pp. 120–128.

J. Liu, M. Morgan, K. Hutchison and V. D. Calhoun (2010). 'A Study of the Influence of Sex on Genome Wide Methylation'. In: *PLOS ONE* 5.4, e10028.

Y. Liu, M. J. Aryee, L. Padyukov, M. D. Fallin, E. Hesselberg, A. Runarsson, L. Reinius, N. Acevedo, M. Taub, M. Ronninger et al. (Feb. 2013). 'Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis'. In: *Nature Biotechnology* 31.2, pp. 142–147.

M. W. Logue, A. K. Smith, E. J. Wolf, H. Maniates, A. Stone, S. A. Schichman, R. E. McGlinchey, W. Milberg and M. W. Miller (Nov. 2017). 'The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples'. In: *Epigenomics* 9.11, pp. 1363–1371.

C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano and G. Kroemer (June 2013). 'The Hallmarks of Aging'. In: *Cell* 153.6, pp. 1194–1217.

K. Lunnon, E. Hannon, R. G. Smith, E. Dempster, C. Wong, J. Burrage, C. Troakes, S. Al-Sarraj, A. Kepa, L. Schalkwyk et al. (Feb. 2016). 'Variation in 5-hydroxymethylcytosine across human cortex and cerebellum'. In: *Genome Biology 2016 17:1* 17.1, pp. 1–15.

K. Lunnon, R. Smith, E. Hannon, P. L. De Jager, G. Srivastava, M. Volta, C. Troakes, S. Al-Sarraj, J. Burrage, R. Macdonald et al. (Sept. 2014). 'Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease'. In: *Nature Neuroscience* 17.9, pp. 1164–1170.

G. H. Lushington and R. Chaguturu (Jan. 2016). *Biomedical research: A house of cards?*

P. Lutsik, M. Slawski, G. Gasparoni, N. Vedeneev, M. Hein and J. Walter (Mar. 2017). 'MeDeCom: discovery and quantification of latent components of heterogeneous methylomes'. In: *Genome Biology 2017 18:1* 18.1, pp. 1–20.

P. E. Lutz, A. Tanti, A. Gasecka, S. Barnett-Burns, J. J. Kim, Y. Zhou, G. G. Chen, M. Wakid, M. Shaw, D. Almeida et al. (Dec. 2017). 'Association of a history of child abuse with impaired myelination in the anterior cingulate cortex: Convergent epigenetic, transcriptional, and morphological evidence'. In: *American Journal of Psychiatry* 174.12, pp. 1185–1194.

O. El-Maarri (2003). 'Methods: DNA methylation'. In: *Advances in Experimental Medicine and Biology*. Vol. 544. Kluwer Academic/Plenum Publishers, pp. 197–204.

G. Mansell, T. J. Gorrie-Stone, Y. Bao, M. Kumari, L. S. Schalkwyk, J. Mill and E. Hannon (May 2019). 'Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array'. In: *BMC Genomics 2019 20:1* 20.1, pp. 1–15.

K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury et al. (Dec. 2011). 'The Parkinson Progression Marker Initiative (PPMI)'. In: *Progress in Neurobiology* 95.4, pp. 629–635.

S. Marques, A. Zeisel, S. Codeluppi, D. v. Bruggen, A. M. Falcão, L. Xiao, H. Li, M. Häring, H. Hochgerner, R. A. Romanov et al. (June 2016). 'Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system'. In: *Science* 352.6291, pp. 1326–1329.

D. Martin, J. Xu, C. Porretta and C. D. Nichols (Feb. 2017). 'Neurocytometry: Flow cytometric sorting of specific neuronal populations from human and rodent brain'. In: *ACS chemical neuroscience* 8.2, p. 356.

A. Matevossian and S. Akbarian (2008). 'Neuronal nuclei isolation from human post-mortem brain tissue'. In: *Journal of Visualized Experiments* 20, p. 914.

H. Mathys, J. Davila-Velderrain, Z. Peng, F. Gao, S. Mohammadi, J. Z. Young, M. Menon, L. He, F. Abdurrob, X. Jiang et al. (May 2019). ‘Single-cell transcriptomic analysis of Alzheimer’s disease’. In: *Nature*, p. 1.

N. Matigian, L. Windus, H. Smith, C. Filippich, C. Pantelis, J. McGrath, B. Mowry and N. Hayward (Sept. 2007). ‘Expression profiling in monozygotic twins discordant for bipolar disorder reveals dysregulation of the WNT signalling pathway’. In: *Molecular psychiatry* 12.9, pp. 815–825.

N. S. McCarthy, P. E. Melton, G. Cadby, S. Yazar, M. Franchina, E. K. Moses, D. A. Mackey and A. W. Hewitt (Nov. 2014). ‘Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns’. In: *BMC Genomics 2014 15:1* 15.1, pp. 1–11.

K. McGregor, S. Bernatsky, I. Colmegna, M. Hudson, T. Pastinen, A. Labbe and C. M. Greenwood (Dec. 2016). ‘An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies’. In: *Genome Biology* 17.1, p. 84.

A. F. McRae, J. E. Powell, A. K. Henders, L. Bowdler, G. Hemani, S. Shah, J. N. Painter, N. G. Martin, P. M. Visscher and G. W. Montgomery (May 2014). ‘Contribution of genetic variation to transgenerational inheritance of DNA methylation’. In: *Genome Biology 2014 15:5* 15.5, pp. 1–10.

A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe et al. (2008). ‘Genome-scale DNA methylation maps of pluripotent and differentiated cells’. In: *Nature* 454.7205, pp. 766–770.

M. Mellén, P. Ayata, S. Dewell, D. Kriaucionis and N. Heintz (Dec. 2012). ‘MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system’. In: *Cell* 151.7, pp. 1417–1430.

I. Mendizabal, S. Berto, N. Usui, K. Toriumi, P. Chatterjee, C. Douglas, I. Huh, H. Jeong, T. Layman, C. A. Tamminga et al. (July 2019). ‘Cell type-specific epigenetic links to schizophrenia risk in the brain’. In: *Genome Biology* 20.1.

L. Y. M. Middleton, J. Dou, J. Fisher, J. A. Heiss, V. K. Nguyen, A. C. Just, J. Faul, E. B. Ware, C. Mitchell, J. A. Colacino et al. (2020). ‘Saliva cell type DNA methylation reference panel for epidemiology studies in children’. In:

J. Mill and B. T. Heijmans (July 2013). 'From promises to practical strategies in epigenetic epidemiology'. In: *Nature Reviews Genetics 2013 14:8* 14.8, pp. 585–594.

J. Mill, T. Tang, Z. Kaminsky, T. Khare, S. Yazdanpanah, L. Bouchard, P. Jia, A. Assadzadeh, J. Flanagan, A. Schumacher et al. (2008). 'Epigenomic Profiling Reveals DNA-Methylation Changes Associated with Major Psychosis'. In:

J. L. Min, G. Hemani, G. D. Smith, C. Relton and M. Suderman (Dec. 2018). 'Meffil: efficient normalization and analysis of very large DNA methylation datasets'. In: *Bioinformatics* 34.23, p. 3983.

X. Ming, Z. Zhang, Z. Zou, C. Lv, Q. Dong, Q. He, Y. Yi, Y. Li, H. Wang and B. Zhu (June 2020). 'Kinetics and mechanisms of mitotic inheritance of DNA methylation and their roles in aging-associated methylome deterioration'. In: *Cell Research 2020 30:11* 30.11, pp. 980–996.

C. Montano, M. A. Taub, A. Jaffe, E. Briem, J. I. Feinberg, R. Trygvadottir, A. Idrizi, A. Runarsson, B. Berndsen, R. C. Gur et al. (May 2016). 'Association of DNA methylation differences with schizophrenia in an epigenome-wide association study'. In: *JAMA Psychiatry* 73.5, pp. 506–514.

L. D. Moore, T. Le and G. Fan (Jan. 2013). *DNA methylation and its basic function.*

R. H. Mulder, A. Neumann, C. A. M. Cecil, E. Walton, L. C. Houtepen, A. J. Simpkin, J. Rijlaarsdam, B. T. Heijmans, T. R. Gaunt, J. F. Felix et al. (Mar. 2021). 'Epigenome-wide change and variation in DNA methylation in childhood: trajectories from birth to late adolescence'. In: *Human Molecular Genetics* 30.1, pp. 119–134.

M. Münzel, D. Globisch, T. Brückl, M. Wagner, V. Welzmiller, S. Michalakis, M. Müller, M. Biel and T. Carell (July 2010). 'Quantification of the Sixth DNA Base Hydroxymethylcytosine in the Brain'. In: *Angewandte Chemie International Edition* 49.31, pp. 5375–5377.

Z. Mussa, J. Tome-Garcia, Y. Jiang, S. Akbarian and N. M. Tsankova (Apr. 2021). 'Isolation of Adult Human Astrocyte Populations from Fresh-frozen Cortex using Fluorescence-Activated Nuclei Sorting'. In: *Journal of visualized experiments : JoVE* 2021.170.

X. Nan, H. H. Ng, C. A. Johnson, C. D. Laherty, B. M. Turner, R. N. Eisenman and A. Bird (May 1998). ‘Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex’. In: *Nature* 393.6683, pp. 386–389.

K.-A. Nave (Nov. 2010). ‘Myelination and support of axonal integrity by glia’. In: *Nature 2010 468:7321* 468.7321, pp. 244–252.

E. P. Neff (Aug. 2019). ‘Animal models of Alzheimer’s disease embrace diversity’. In: *Lab Animal 2019 48:9* 48.9, pp. 255–259.

A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn and A. A. Alizadeh (2015). ‘Robust enumeration of cell subsets from tissue expression profiles HHS Public Access’. In: *Nat Methods* 12.5, pp. 453–457.

F. Nifosì, T. Toffanin, H. Follador, F. Zonta, G. Padovan, G. Pigato, C. Carollo, M. Ermani, P. Amistà and G. I. Perini (Oct. 2010). ‘Reduced right posterior hippocampal volume in women with recurrent familial pure depressive disorder’. In: *Psychiatry research* 184.1, pp. 23–28.

S. Numata, T. Ye, M. Herman and B. K. Lipska (2014). ‘DNA methylation changes in the postmortem dorsolateral prefrontal cortex of patients with schizophrenia’. In: *Frontiers in Genetics* 5.JUL.

S. Numata, T. Ye, T. M. Hyde, X. Guitart-Navarro, R. Tao, M. Wininger, C. Colantuoni, D. R. Weinberger, J. E. Kleinman and B. K. Lipska (Feb. 2012). ‘DNA Methylation Signatures in Development and Aging of the Human Prefrontal Cortex’. In: *The American Journal of Human Genetics* 90.2, pp. 260–272.

N. F. P. d. Oliveira, B. F. d. Souza and M. d. C. Coêlho (Sept. 2020). ‘UV Radiation and Its Relation to DNA Methylation in Epidermal Cells: A Review’. In: *Epigenomes 2020, Vol. 4, Page 23* 4.4, p. 23.

A. H. Olsson, P. Volkov, K. Bacos, T. Dayeh, E. Hall, E. A. Nilsson, C. Ladenvall, T. Rönn and C. Ling (Nov. 2014). ‘Genome-Wide Associations between Genetic and Epigenetic Variation Influence mRNA Expression and Insulin Secretion in Human Pancreatic Islets’. In: *PLOS Genetics* 10.11, e1004735.

V. Onuchic, R. J. Hartmaier, D. N. Boone, M. L. Samuels, R. Y. Patel, W. M. White, V. D. Garovic, S. Oesterreich, M. E. Roth, A. V. Lee et al. (Nov. 2016). ‘Epigenomic

Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types'. In: *Cell Reports* 17.8, pp. 2075–2086.

L. D. Orozco, L. Rubbi, L. J. Martin, F. Fang, F. Hormozdiari, N. Che, A. D. Smith, A. J. Lusis and M. Pellegrini (Apr. 2014). 'Intergenerational genomic DNA methylation patterns in mouse hybrid strains'. In: *Genome Biology 2014 15:5* 15.5, pp. 1–13.

S. Påhlman, A. I. Ruusala, L. Abrahamsson, M. E. Mattsson and T. Esscher (1984). 'Retinoic acid-induced differentiation of cultured human neuroblastoma cells: a comparison with phorbolester-induced differentiation'. In: *Cell differentiation* 14.2, pp. 135–144.

S. Pai, P. Li, B. Killinger, L. Marshall, P. Jia, J. Liao, A. Petronis, P. E. Szabó and V. Labrie (Dec. 2019). 'Differential methylation of enhancer at IGF2 is associated with abnormal dopamine synthesis in major psychosis'. In: *Nature Communications* 10.1.

V. Patil, R. L. Ward and L. B. Hesson (June 2014). 'The evidence for functional non-CpG methylation in mammalian cells'. In: *Epigenetics* 9.6, pp. 823–828.

B. S. Pedersen, D. A. Schwartz, I. V. Yang and K. J. Kechris (Nov. 2012). 'Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values'. In: *Bioinformatics* 28.22, pp. 2986–2988.

E. Pennisi (Oct. 2010). '1000 Genomes project gives new map of genetic diversity'. In: *Science* 330.6004, pp. 574–575.

T. J. Peters, M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, R. V Lord, S. J. Clark and P. L. Molloy (Jan. 2015). 'De novo identification of differentially methylated regions in the human genome'. In: *Epigenetics & Chromatin 2015 8:1* 8.1, pp. 1–16.

D. Phillips, Y. Wang, N. L. Pedersen, S. Hägg, K. Institutet and C. Reynolds (2019). 'Loneliness and DNA methylation of variants in the conserved transcriptional response to adversity (CTRA) pathway'. In: *Behavior Genetics* 49.532.

R. Pidsley, J. Viana, E. Hannon, H. Spiers, C. Troakes, S. Al-saraj, N. Mechawar, G. Turecki, L. C. Schalkwyk, N. J. Bray et al. (2014). 'Methylomic profiling of human brain tissue supports a neurodevelopmental origin for schizophrenia'. In: *Genome biology* 15.10.

R. Pidsley, C. C. Y Wong, M. Volta, K. Lunnon, J. Mill and L. C. Schalkwyk (May 2013). 'A data-driven approach to preprocessing Illumina 450K methylation array data'. In: *BMC Genomics* 14.1, p. 293.

R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. Van Djik, B. Muhlhausler, C. Stirzaker and S. J. Clark (Oct. 2016). 'Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling'. In: *Genome Biology* 17.1.

E. Pishva, B. Creese, A. R. Smith, W. Viechtbauer, P. Proitsi, D. L. van den Hove, C. Ballard, J. Mill and K. Lunnon (May 2020). 'Psychosis-associated DNA methylomic variation in Alzheimer's disease cortex'. In: *Neurobiology of Aging* 89, pp. 83–88.

N. Plongthongkum, D. H. Diep and K. Zhang (Oct. 2014). *Advances in the profiling of DNA modifications: Cytosine methylation and beyond*.

S. S. Policicchio, J. P. Davies, B. Chioza, J. Burrage, J. Mill, E. L. Dempster and S. Policicchio (Oct. 2020a). 'Fluorescence-activated nuclei sorting (FANS) on human post-mortem cortex tissue enabling the isolation of distinct neural cell populations for multiple omic profiling Neurodegeneration Method Development Community Complex Disease Epigenetics Group'. In:

S. Policicchio, S. Washer, J. Viana, A. Iatrou, J. Burrage, E. Hannon, G. Turecki, Z. Kaminsky, J. Mill, E. L. Dempster et al. (Dec. 2020b). 'Genome-wide DNA methylation meta-analysis in the brains of suicide completers'. In: *Translational Psychiatry* 10.1, p. 69.

E. M. Price and W. P. Robinson (Mar. 2018). 'Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned'. In: *Frontiers in Genetics* 0.MAR, p. 83.

F. Prinz, T. Schlange and K. Asadullah (Sept. 2011). *Believe it or not: How much can we rely on published data on potential drug targets?*

M. Prinz and J. Priller (Jan. 2017). 'The role of peripheral immune cells in the CNS in steady state and disease'. In: *Nature Neuroscience 2017 20:2* 20.2, pp. 136–144.

I. Rafael, M. Aryee and K. D. Hansen (2017). 'Package 'bumphunter''. In:

E. Rahmani, R. Schweiger, L. Shenhav, T. Wingert, I. Hofer, E. Gabel, E. Eskin and E. Halperin (Sept. 2018). 'BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference'. In: *Genome Biology 2018 19:1* 19.1, pp. 1–18.

E. Rahmani, N. Zaitlen, Y. Baran, C. Eng, D. Hu, J. Galanter, S. Oh, E. G. Burchard, E. Eskin, J. Zou et al. (May 2016). 'Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies'. In: *Nature Methods* 13.5, pp. 443–445.

E.-A. Raiber, G. Portella, S. Martínez Cuesta, R. Hardisty, P. Murat, Z. Li, M. Iurlaro, W. Dean, J. Spindel, D. Beraldi et al. (Oct. 2018). '5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells'. In: *Nature Chemistry 2018 10:12* 10.12, pp. 1258–1266.

K. Raj, B. Szladovits, A. Haghani, J. A. Zoller, C. Z. Li, P. Black, D. Maddox, T. R. Robeck and S. Horvath (Aug. 2021). 'Epigenetic clock and methylation studies in cats'. In: *GeroScience 2021*, pp. 1–16.

L. E. Reinius, N. Acevedo, M. Joerink, G. Pershagen, S.-E. Dahlén, D. Greco, C. Söderhäll, A. Scheynius and J. Kere (July 2012). 'Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility'. In: *PLoS ONE* 7.7. Ed. by A. H. Ting, e41361.

M. Rhein, L. Hagemeier, M. Klintschar, M. Muschler, S. Bleich and H. Frieling (2015). 'DNA methylation results depend on DNA integrity – role of post mortem interval'. In: *Frontiers in Genetics* 0.MAY, p. 182.

R. C. Richmond, A. J. Simpkin, G. Woodward, T. R. Gaunt, O. Lyttleton, W. L. McArdle, S. M. Ring, A. D. Smith, N. J. Timpson, K. Tilling et al. (Apr. 2015). 'Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC)'. In: *Human Molecular Genetics* 24.8, pp. 2201–2217.

A. D. Riggs (1975). *X inactivation, differentiation, and DNA méthylation*. Tech. rep., pp. 9–25.

L. F. Rizzardi, P. F. Hickey, V. Rodriguez DiBlasi, R. Tryggvadóttir, C. M. Callahan, A. Idrizi, K. D. Hansen and A. P. Feinberg (Jan. 2019a). 'Neuronal brain-region-specific

DNA methylation and chromatin accessibility are associated with neuropsychiatric trait heritability'. In: *Nature Neuroscience 2019 22:2* 22.2, pp. 307–316.

L. F. Rizzardi, P. F. Hickey, V. Rodriguez DiBlasi, R. Tryggvadóttir, C. M. Callahan, A. Idrizi, K. D. Hansen and A. P. Feinberg (Feb. 2019b). 'Neuronal brain-region-specific DNA methylation and chromatin accessibility are associated with neuropsychiatric trait heritability'. In: *Nature Neuroscience* 22.2, pp. 307–316.

T. Rönn, P. Volkov, C. Davegårdh, T. Dayeh, E. Hall, A. H. Olsson, E. Nilsson, Å. Tornberg, M. Dekker Nitert, K.-F. Eriksson et al. (June 2013). 'A Six Months Exercise Intervention Influences the Genome-wide DNA Methylation Pattern in Human Adipose Tissue'. In: *PLoS Genetics* 9.6. Ed. by J. M. Greally, e1003572.

J. A. Y. Roubroeks, R. G. Smith, D. L. A. v. d. Hove and K. Lunnon (Oct. 2017). 'Epigenetics and DNA methylomic profiling in Alzheimer's disease and other neurodegenerative diseases'. In: *Journal of Neurochemistry* 143.2, pp. 158–170.

P. Rovira, C. Sánchez-Mora, M. Pagerols, V. Richarte, M. Corrales, C. Fadeuilhe, L. Vilar-Ribó, L. Arribas, G. Shireby, E. Hannon et al. (June 2020). 'Epigenome-wide association study of attention-deficit/hyperactivity disorder in adults'. In: *Translational Psychiatry 2020 10:1* 10.1, pp. 1–12.

L. A. Salas, D. C. Koestler, R. A. Butler, H. M. Hansen, J. K. Wiencke, K. T. Kelsey and B. C. Christensen (Dec. 2018). 'An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray'. In: *Genome Biology* 19.1, p. 64.

H. B. Sarnat, D. Nochlin and D. E. Born (Mar. 1998). 'Neuronal nuclear antigen (NeuN): a marker of neuronal maturation in the early human fetal nervous system'. In: *Brain and Development* 20.2, pp. 88–94.

A. Schäfer and R. S. Baric (Feb. 2017). 'Epigenetic Landscape during Coronavirus Infection'. In: *Pathogens 2017, Vol. 6, Page 8* 6.1, p. 8.

B. Schmitz, A. Radbruch, T. Kümmel, C. Wickenhauser, H. Korb, M. Hansmann, J. Thiele and R. Fischer (May 1994). 'Magnetic activated cell sorting (MACS) — a new immunomagnetic method for megakaryocytic cell isolation: Comparison of different separation techniques'. In: *European Journal of Haematology* 52.5, pp. 267–275.

D. Seiler Vellame, I. Castanho, A. Dahir, J. Mill and E. Hannon (June 2021). 'Characterizing the properties of bisulfite sequencing data: maximizing power and sensitivity to identify between-group differences in DNA methylation'. In: *BMC Genomics 2021 22:1* 22.1, pp. 1–16.

G. C. Sharp, D. A. Lawlor, R. C. Richmond, A. Fraser, A. Simpkin, M. Suderman, H. A. Shihab, O. Lyttleton, W. McArdle, S. M. Ring et al. (Aug. 2015). 'Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: findings from the Avon Longitudinal Study of Parents and Children'. In: *International Journal of Epidemiology* 44.4, pp. 1288–1304.

Y. Shi, P. Kirwan and F. J. Livesey (Sept. 2012). 'Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks'. In: *Nature Protocols 2012 7:10* 7.10, pp. 1836–1846.

G. L. Shireby, J. P. Davies, P. T. Francis, J. Burrage, E. M. Walker, G. W. A. Neilson, A. Dahir, A. J. Thomas, S. Love, R. G. Smith et al. (Dec. 2020). 'Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex'. In: *Brain* 143.12, pp. 3763–3775.

M. W. Simmen (July 2008). 'Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals'. In: *Genomics* 92.1, pp. 33–40.

M. K. Skinner and C. Guerrero-Bosagna (Aug. 2014). 'Role of CpG deserts in the epigenetic transgenerational inheritance of differential DNA methylation regions'. In: *BMC Genomics* 15.1.

K. Skvortsova, E. Zotenko, P.-L. Luu, C. M. Gould, S. S. Nair, S. J. Clark and C. Stirzaker (Apr. 2017). 'Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA'. In: *Epigenetics & Chromatin 2017 10:1* 10.1, pp. 1–20.

M. Z. Smith, M. M. Esiri, L. Barnetson, E. King and Z. Nagy (2001). *Constructional Apraxia in Alzheimer's Disease: Association with Occipital Lobe Pathology and Accelerated Cognitive Decline*. Tech. rep., pp. 281–288.

R. G. Smith, E. Pishva, G. Shireby, A. R. Smith, J. A. Y. Roubroeks, E. Hannon, G. Wheildon, D. Mastroeni, G. Gasparoni, M. Riemenschneider et al. (June 2021). 'A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex'. In: *Nature Communications 2021 12:1* 12.1, pp. 1–13.

R. G. Smith, E. Pishva, G. Shireby, A. R. Smith, J. A. Y. Roubroeks, E. Hannon, G. Wheildon, D. Mastroeni, G. Gasparoni, M. Riemenschneider et al. (2020). 'Meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights 220 differentially methylated loci across cortex'. In:

Z. D. Smith and A. Meissner (Feb. 2013). 'DNA methylation: roles in mammalian development'. In: *Nature Reviews Genetics 2013 14:3* 14.3, pp. 204–220.

Z. D. Smith, H. Gu, C. Bock, A. Gnirke and A. Meissner (July 2009). 'High-throughput bisulfite sequencing in mammalian genomes.' In: *Methods (San Diego, Calif.)* 48.3, pp. 226–32.

H. Spiers, E. Hannon, L. C. Schalkwyk, N. J. Bray and J. Mill (Sept. 2017). '5-hydroxymethylcytosine is highly dynamic across human fetal brain development'. In: *BMC Genomics 2017 18:1* 18.1, pp. 1–14.

H. Spiers, E. Hannon, L. C. Schalkwyk, R. Smith, C. C. Y. Wong, M. C. O'Donovan, N. J. Bray and J. Mill (Mar. 2015). 'Methylomic trajectories across human fetal brain development.' In: *Genome research* 25.3, pp. 338–52.

M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Schöler, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari et al. (Dec. 2011). 'DNA-binding factors shape the mouse methylome at distal regulatory regions'. In: *Nature* 480.7378, pp. 490–495.

M. Stefan, W. Zhang, E. Concepcion, Z. Yi and Y. Tomer (May 2014). 'DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology'. In: *Journal of Autoimmunity* 50, pp. 33–37.

L. C. Steg, G. L. Shireby, J. Imm, J. P. Davies, A. Franklin, R. Flynn, S. C. Namboori, A. Bhinge, A. R. Jeffries, J. Burrage et al. (June 2021). 'Novel epigenetic clock for fetal brain development predicts prenatal age for cellular stem cell models and derived neurons'. In: *Molecular Brain 2021 14:1* 14.1, pp. 1–11.

T. Strachan and A. Read (2018). *Human Molecular Genetics*. 5th ed. CRC Press LLC.

B. D. Strahl and C. D. Allis (Jan. 2000). *The language of covalent histone modifications*.

T. M. Stubbs, M. J. Bonder, A.-K. Stark, F. Krueger, F. von Meyenn, O. Stegle and W. Reik (Dec. 2017). 'Multi-tissue DNA methylation age predictor in mouse'. In: *Genome Biology* 18.1, p. 68.

M. Suderman, J. R. Staley, R. French, R. Arathimos, A. Simpkin and K. Tilling (n.d.). 'dmrff: identifying differentially methylated regions efficiently with power and control'. In: ().

M. Suelves, E. Carrió, Y. Núñez-Álvarez and M. A. Peinado (Nov. 2016). 'DNA methylation dynamics in cellular commitment and differentiation'. In: *Briefings in Functional Genomics* 15.6, pp. 443–453.

Z. Sun, H. S. Chai, Y. Wu, W. M. White, K. V. Donkena, C. J. Klein, V. D. Garovic, T. M. Therneau and J.-P. A. Kocher (Dec. 2011). 'Batch effect correction for genome-wide methylation data with Illumina Infinium platform'. In: *BMC Medical Genomics 2011 4:1* 4.1, pp. 1–12.

R. P. Talens, K. Christensen, H. Putter, G. Willemsen, L. Christiansen, D. Kremer, H. E. D. Suchiman, P. E. Slagboom, D. I. Boomsma and B. T. Heijmans (Aug. 2012). 'Epigenetic variation during the adult lifespan: cross-sectional and longitudinal data on monozygotic twin pairs'. In: *Aging Cell* 11.4, pp. 694–703.

X. Tang, S. Ruan, Q. Tang, J. Huang, X. Luo, X. Li, S. Luo, Z. Gao, Q. Kang, Y. Tang et al. (2018). *Original Article DNA methylation profiles in the hippocampus of an Alzheimer's disease mouse model at mid-stage neurodegeneration*. Tech. rep. 11. Department of Pathology, The Affiliated Hospital of Southwest Medical University, Luzhou 646000, Sichuan, China, pp. 11876–11888.

Y. Tang, M. Horikoshi and W. Li (2016). 'ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages'. In: *The R Journal* 8.2.

A. E. Teschendorff and S. C. Zheng (May 2017). 'Cell-type deconvolution in epigenomewide association studies: a review and recommendations'. In: *Epigenomics* 9.5, pp. 757–768.

A. E. Teschendorff, J. Zhuang and M. Widschwendter (June 2011). 'Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies'. In: *Bioinformatics* 27.11, pp. 1496–1505.

A. E. Teschendorff, C. E. Breeze, S. C. Zheng and S. Beck (n.d.). 'A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies'. In: ().

A. Thompson, M. R. May, B. R. Moore and A. Kopp (n.d.). 'A Hierarchical Bayesian Mixture Model for Inferring the Expression State of Genes in Transcriptomes'. In: ().

A. J. Titus, R. M. Gallimore, L. A. Salas and B. C. Christensen (2017). 'Cell-type deconvolution from DNA methylation: a review of recent applications.' In: *Human molecular genetics* 26.R2, R216–R224.

J. Tulloch, L. Leong, Z. Thomson, S. Chen, E. G. Lee, C. D. Keene, S. P. Millard and C. E. Yu (Nov. 2018). 'Glia-specific APOE epigenetic changes in the Alzheimer's disease brain'. In: *Brain Research* 1698, pp. 179–186.

P. Turko, K. Groberman, F. Browa, S. Cobb and I. Vida (Mar. 2019). 'Differential Dependence of GABAergic and Glutamatergic Neurons on Glia for the Establishment of Synaptic Transmission'. In: *Cerebral Cortex* 29.3, pp. 1230–1243.

N. A. Uranova, V. M. Vostrikov, D. D. Orlovskaya and V. I. Rachmanova (2004). 'Oligodendroglial density in the prefrontal cortex in schizophrenia and mood disorders: A study from the Stanley Neuropathology Consortium'. In: *Schizophrenia Research*.

M. A. Urich, J. R. Nery, R. Lister, R. J. Schmitz and J. R. Ecker (Mar. 2015). 'MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing'. In: *Nature Protocols* 10.3, pp. 475–483.

K. E. Varley, J. Gertz, K. M. Bowling, S. L. Parker, T. E. Reddy, F. Pauli-Behn, M. K. Cross, B. A. Williams, J. A. Stamatoyannopoulos, G. E. Crawford et al. (Mar. 2013). 'Dynamic DNA methylation across diverse human cell lines and tissues'. In: *Genome Research* 23.3, pp. 555–567.

S. Vazire (Jan. 2017). *Quality uncertainty erodes trust in science*.

C. H. Waddington (1957). *The Strategy of the Genes*.

S. Wahl, A. Drong, B. Lehne, M. Loh, W. R. Scott, S. Kunze, P.-C. Tsai, J. S. Ried, W. Zhang, Y. Yang et al. (Dec. 2016). 'Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity'. In: *Nature 2016 541:7635* 541.7635, pp. 81–86.

E. Walton, J. B. Pingault, C. A. M. Cecil, T. R. Gaunt, C. L. Relton, J. Mill and E. D. Barker (May 2016). ‘Epigenetic profiling of ADHD symptoms trajectories: a prospective, methylome-wide study’. In: *Molecular Psychiatry 2017 22:2* 22.2, pp. 250–256.

L. Wang, Y. Zhou, L. Xu, R. Xiao, X. Lu, L. Chen, J. Chong, H. Li, C. He, X.-D. Fu et al. (June 2015). ‘Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex’. In: *Nature 2015 523:7562* 523.7562, pp. 621–625.

R. Y. Wang, C. W. Gehrke and M. Ehrlich (Oct. 1980). ‘Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues’. In: *Nucleic Acids Research* 8.20, pp. 4777–4790.

R. A. Waterland (Nov. 2006). ‘Epigenetic mechanisms and gastrointestinal development’. In: *Journal of Pediatrics* 149.5 SUPPL. S137–S142.

H. Wickham (2007). ‘Reshaping Data with the reshape Package’. In: *Journal of Statistical Software* 21.12, pp. 1–20.

H. Wickham (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

H. Wickham (2020). *forcats: Tools for Working with Categorical Variables (Factors)*.

H. Wickham, R. François, L. Henry and K. Müller (2020). *dplyr: A Grammar of Data Manipulation*.

H. Wickham and L. Henry (2020). *tidyr: Tidy Messy Data*.

H. Wickham and D. Seidel (2020). *scales: Scale Functions for Visualization*.

C. O. Wilke (n.d.). *Cowplot R package*.

L. F. Wockner, E. P. Noble, B. R. Lawford, R. M. D. Young, C. P. Morris, V. L. Whitehall and J. Voisey (Jan. 2014). ‘Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients’. In: *Translational Psychiatry 2014 4:1* 4.1, e339–e339.

C. C. Y. Wong, R. G. Smith, E. Hannon, G. Ramaswami, N. N. Parikshak, E. Assary, C. Troakes, J. Poschmann, L. C. Schalkwyk, W. Sun et al. (2019). ‘Genome-wide DNA methylation profiling identifies convergent molecular signatures associated with idiopathic and syndromic autism in post-mortem human brain tissue'. In: *Human Molecular Genetics* 28.13, pp. 2201–2211.

X. Wu and Y. Zhang (May 2017). 'TET-mediated active DNA demethylation: mechanism, function and beyond'. In: *Nature Reviews Genetics 2017 18:9* 18.9, pp. 517–534.

H. R. Xie, L. S. Hu and G. Y. Li (2010). 'SH-SY5Y human neuroblastoma cell line: In vitro cell model of dopaminergic neurons in Parkinson's disease'. In: *Chinese Medical Journal* 123.8, pp. 1086–1092.

W. Xie, M. D. Schultz, R. Lister, Z. Hou, N. Rajagopal, P. Ray, J. W. Whitaker, S. Tian, R. D. Hawkins, D. Leung et al. (May 2013). 'Epigenomic analysis of multilineage differentiation of human embryonic stem cells'. In: *Cell* 153.5, pp. 1134–1148.

Y. Yin, E. Morgunova, A. Jolma, E. Kaasinen, B. Sahu, S. Khund-Sayeed, P. K. Das, T. Kivioja, K. Dave, F. Zhong et al. (May 2017). 'Impact of cytosine methylation on DNA binding specificities of human transcription factors'. In: *Science (New York, N.Y.)* 356.6337.

P. Yousefi, K. Huen, V. Davé, L. Barcellos, B. Eskenazi and N. Holland (Nov. 2015). 'Sex differences in DNA methylation assessed by 450K BeadChip in newborns'. In: *BMC Genomics* 16.1, p. 911.

I. Yusipov, M. G. Bacalini, A. Kalyakulina, M. Krivonosov, C. Pirazzini, N. Gensous, F. Ravaioli, M. Milazzo, M. Vedunova, G. Fiorito et al. (2020). 'Complex sex-specific age-related changes in DNA methylation including variability, epimutations and entropy'. In: *bioRxiv*, p. 2020.01.15.905224.

S. K. Zaidi, D. W. Young, M. Montecino, A. J. Van Wijnen, J. L. Stein, J. B. Lian and G. S. Stein (May 2011). *Bookmarking the genome: Maintenance of epigenetic information.*

S. Zeilinger, B. Kühnel, N. Klopp, H. Baurecht, A. Kleinschmidt, C. Gieger, S. Weidinger, E. Lattka, J. Adamski, A. Peters et al. (May 2013). 'Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation'. In: *PLoS ONE* 8.5. Ed. by A. Chen, e63812.

A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz et al. (Mar. 2015). 'Brain structure.

Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.' In: *Science (New York, N.Y.)* 347.6226, pp. 1138–42.

H. Zeng and J. R. Sanes (Aug. 2017). 'Neuronal cell-type classification: challenges, opportunities and the path forward'. In: *Nature Reviews Neuroscience 2017 18:9* 18.9, pp. 530–546.

G. E. Zentner and S. Henikoff (Mar. 2013). 'Regulation of nucleosome dynamics by histone modifications'. In: *Nature Structural & Molecular Biology 2013 20:3* 20.3, pp. 259–266.

F. F. Zhang, R. Cardarelli, J. Carroll, K. G. Fulda, M. Kaur, K. Gonzalez, J. K. Vishwanatha, R. M. Santella and A. Morabia (May 2011). 'Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood'. In: *Epigenetics* 6.5, pp. 623–629.

W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell and B. E. Engelhardt (Jan. 2015). 'Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements'. In: *Genome Biology* 16.1, pp. 1–20.

W. Zhang, H. Wu and Z. Li (May 2021). 'Complete deconvolution of DNA methylation signals from complex tissues: a geometric approach'. In: *Bioinformatics* 37.8, pp. 1052–1059.

X. Zheng, S. Gogarten, M. Lawrence, A. Stilp, M. Conomos, B. Weir, C. Laurie and D. Levine (2017). 'SeqArray – A storage-efficient high-performance data format for WGS variant calls'. In: *Bioinformatics*.

X. Zheng, D. Levine, J. Shen, S. Gogarten, C. Laurie and B. Weir (2012). 'A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data'. In: *Bioinformatics* 28.24, pp. 3326–3328.

S. Zhong, S. Zhang, X. Fan, Q. Wu, L. Yan, J. Dong, H. Zhang, L. Li, L. Sun, N. Pan et al. (Mar. 2018). 'A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex'. In: *Nature 2018 555:7697* 555.7697, pp. 524–528.

J. Zhou, R. L. Sears, X. Xing, B. Zhang, D. Li, N. B. Rockweiler, H. S. Jang, M. N. Choudhary, H. J. Lee, R. F. Lowdon et al. (Sept. 2017). 'Tissue-specific DNA methylation is conserved across human, mouse, and rat, and driven by primary sequence conservation'. In: *BMC Genomics* 18.1, pp. 1–17.

W. Zhou, P. W. Laird and H. Shen (2016). 'Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes'. In: *Nucleic Acids Research* 45.4, p. 22.

R. S. Ziffra, C. N. Kim, J. M. Ross, A. Wilfert, T. N. Turner, M. Haeussler, A. M. Casella, P. F. Przytycki, K. C. Keough, D. Shin et al. (Oct. 2021). 'Single-cell epigenomics reveals mechanisms of human cortical development'. In: *Nature 2021 598:7879* 598.7879, pp. 205–213.

M. J. Ziller, H. Gu, F. Müller, J. Donaghey, L. T. Tsai, O. Kohlbacher, P. L. De Jager, E. D. Rosen, D. A. Bennett, B. E. Bernstein et al. (Aug. 2013). 'Charting a dynamic DNA methylation landscape of the human genome'. In: *Nature* 500.7463, pp. 477–481.

M. J. Ziller, K. D. Hansen, A. Meissner and M. J. Aryee (Feb. 2015). 'Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing'. In: *Nature Methods* 12.3, pp. 230–232.

J. Zou, C. Lippert, D. Heckerman, M. Aryee and J. Listgarten (2014). 'Epigenome-wide association studies without the need for cell-type composition'. In: *Nature Methods* 11.3, pp. 309–311.