DNA Methylation Analysis and Age Prediction

Yucheng Wang

A thesis submitted for the degree of Doctor of Philosophy

in Computer Science

School of Computer Science and Electronic Engineering

University of Essex

June 2023

Abstract

DNA methylation microarrays have been the most cost-effective choice for large cohort studies aimed to investigate associations between methylome changes and diseases or environmental exposures. The findings of many CpG sites across the genome whose methylation changes are highly correlated with age have led to the construction of various interesting epigenetic age estimation models, also known as epigenetic clocks. However, there is still largely unclear regarding the mechanisms that drive age associate methylation changes. In this thesis, the first two chapters describe two novel bioinformatic tools for analyzing DNA methylation microarray data respectively. After that, the existing claim that cerebellums age slowly is re-examined.

Many samples on the Gene Expression Omnibus frequently lack a sex annotation or are incorrectly labelled. Considering the influence that sex imposes on DNA methylation patterns, it is necessary to ensure that methods for filtering poor samples and checking sex assignments are accurate and widely applicable. In the first chapter, a novel method to predict sample sex using only DNA methylation beta values is presented, which can be readily applied to almost all DNA methylation datasets of different formats. I firstly identified 4,345 CpG sites located on both 450K and EPIC arrays which are differentially methylated between females and males. A novel sex classifier was then constructed by combining the two first principal components of the DNA methylation data of sex-associated probes mapped on sex chromosomes. The proposed method was constructed using whole blood samples and exhibits good performance across a wide range of tissues. It is also demonstrated that this classifier can be used to identify samples with sex chromosome aneuploidy, this function is

validated by five Turner syndrome cases and one Klinefelter syndrome case.

Data normalization is an essential step to reduce technical variation within and between arrays. Due to the different karyotypes and the effects of X chromosome inactivation, females and males exhibit distinct methylation patterns on sex chromosomes; this poses a significant challenge to normalize sex chromosome data without introducing bias. Currently, existing methods do not provide unbiased solutions to normalize sex chromosome data, usually, they just process autosomal and sex chromosomes indiscriminately. In chapter 2, I first demonstrate that ignoring this sex difference will lead to introducing artificial sex bias, especially for thousands of autosomal CpGs. Then a novel two-step strategy (interpolatedXY) was created to address this issue, which is applicable to all quantile-based normalization methods. Employing this new strategy, the autosomal CpGs are first normalized independently by conventional methods, such as funnorm [1] or dasen[2]; then the corrected methylation values of sex chromosome-linked CpGs are estimated as the weighted average of their nearest neighbors on autosomes. The proposed two-step strategy can also be applied to other non-quantile-based normalization methods, as well as other array-based data types.

Despite different tissues having vastly different rates of proliferation, it is still largely unknown whether they age at different rates. It was previously reported that the cerebellum ages slowly, however, this claim was drawn from a single methylation clock using a small sample size and thus warrants further investigation. In chapter 3, I first collected the largest cerebellum DNAm dataset (N=752) and found their respective epigenetic ages were all severely underestimated by six representative DNAm age clocks, with the underestimation effects more pronounced in the four clocks whose training datasets did not include brain-related tissues. Then 613 age-associated CpGs are identified in the cerebellum, which accounts for only 14.5% of the number found in the middle temporal gyrus from the same population (N=404). Subsequently, I built a highly accurate age prediction model for the

cerebellum named CerebellumClock$_{\text{specific}}$ (Pearson correlation=0.941, mean absolute deviation=3.18 years). Ageing rate comparisons based on the two tissue-specific clocks constructed on the 201 overlapping age-associated CpGs support the cerebellum has younger DNAm age. Nevertheless, BrainCortexClock is constructed to prove a single DNAm clock is able to unbiasedly estimate DNAm ages of both cerebellum and cerebral cortex when they are adequately and equally represented in the training dataset. In conclusion, comparing ageing rates across tissues using DNA methylation multi-tissue clocks is flawed. The large underestimation of age prediction for cerebellum by previous clocks mainly reflects the improper usage of the age clocks. There exist strong and consistent ageing effects on the cerebellar methylome and we suggest the smaller number of age-associated CpG sites in cerebellum is largely attributed to its extremely low average cell replication rates.

In summary, the sex classifier method presented in the first chapter provides a robust and widely applicable tool to identify the sexes of DNAm methylation samples. It can be applied to make sex annotations and identify sex-mismatch samples. The second chapter presents a novel two-step strategy to bypass the issue of introducing artifactual sex bias when normalizing female samples and male samples together by conventional normalization methods. In the last chapter, the unique age-associated methylome change in the cerebellum is revealed and a cerebellum-specific clock is constructed that can accurately predict cerebellum age and it is demonstrated that the comparison of ageing rates across tissues using epigenetic clocks is flawed. These findings have wider implications for the use of ageing clocks.

List of Publications

1. **Wang, Y.**, Grant, O. A., Zhai, X., McDonald-Maier, K. D., & Schalkwyk, L. C. (2022). Recalibrating the cerebellum DNA methylation clock: implications for ageing rates comparison. *bioRxiv*. (Under Review)

2. **Wang, Y.**, Gorrie-Stone, T. J., Grant, O. A., Andrayas, A. D., Zhai, X., McDonald-Maier, K. D., & Schalkwyk, L. C. (2022). InterpolatedXY: a two-step strategy to normalize DNA methylation microarray data avoiding sex bias. *Bioinformatics*, 38(16), 3950-3957.

3. **Wang, Y.**, Hannon, E., Grant, O. A., Gorrie-Stone, T. J., Kumari, M., Mill, J., Zhai, X., McDonald-Maier, K. D., & Schalkwyk, L. C. (2021). DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy. *BMC genomics*, 22(1), 1-11.

4. Grant, O. A., **Wang, Y.**, Kumari, M., Zabet, N. R., & Schalkwyk, L. (2022). Characterising sex differences of autosomal DNA methylation in whole blood using the Illumina EPIC array. *Clinical epigenetics*, 14(1), 1-16.

5. **Wang, Y.**, Su, J., Zhai, X., Meng, F., & Liu, C. (2022). Snow coverage mapping by learning from sentinel-2 satellite multispectral images via machine learning algorithms. *Remote Sensing*, 14(3), 782.

6. Liew, B. X., Rügamer, D., Zhai, X., **Wang, Y.**, Morris, S., & Netto, K. (2021). Comparing shallow, deep, and transfer learning in predicting joint moments in running. *Journal of Biomechanics*, 129, 110820.

Acknowledgments

I would like to thank numerous people who have helped and supported me to finish my PhD course. Firstly, I would like to thank my three supervisors—Dr Xiaojun Zhai, Professor Klaus D McDonald-Maier and Professor Leonard C Schalkwyk, for they have provided me with invaluable guidance in doing research, I am now more confident than ever to pursue an academic research career.

I would also like to thank the University of Essex provided me with a faculty scholarship to support my living during my PhD course.

I appreciate the time I spent with my friends, lab mates, and colleagues—Dr Yufan Lu, Dr Jinya Su, Issam M A Boukhennoufa, Olivia A Grant, Dr Somdip Dey, Dr Cong Gao and more persons.

I thank my parents for their support and I thank my son Yiheng Wang for his accompany, he has brought me numerous happiness. Lastly, I especially want to thank my wife Ruiping for her consistent encouragement and nice emotional support throughout my studies. Without Ruiping's full support, I would not even have started my PhD study.

Abbreviations

450K: Infinium HumanMethylation450 BeadChip

5mC: 5-methylcytosine

BMIQ: Beta MIxture Quantile normalization

CBL: Cerebellum

ChrX: X chromosome

ChrY: Y chromosome

CpG: Cytosine-Guanine Dinucleotide

DMP: Differentially Methylated Position

DNA: Deoxyribonucleic acid

DNMT: DNA methyltransferase

DNAm: DNA methylation

EC: Entorhinal Cortex

EPIC: Infinium MethylationEPIC BeadChip

EWAS: Epigenome-wide Association Studies

FC: Frontal Cortex

GEO: Gene Expression Omnibus

GO: Gene Oncology

MAD: Mean Absolute Deviation

MTG: Middle Temporal Gyrus

NAD: Nicotinamide Adenine Dinucleotide

PCA: Principal Component Analysis

PCR: Polymerase Chain Reaction

PBC: Peak-based Correction

RMSE: Root Mean Squared Error

RMSD: Root Mean Squared Deviation

STG: Superior Temporal Gyrus

TL: Telomere Length

UKHLS: UK Household Longitudinal Study

WGBS: Whole Genome Bisulfite Sequencing

WB: Whole Blood

Table of Contents

List of Figures

List of Tables

Chapter1

Introduction

1.1 Epigenetics

1.1.1 Epigenetics concept

The concept of epigenetics has evolved significantly since its first brought up by Waddington in 1942 [3]. Today, epigenetics is widely accepted as the study of heritable changes that cause gene expression alterations but are independent of changes in DNA sequence [4, 5, 6]. Typical epigenetic modifications include DNA methylation, histone modifications, nucleosome positioning and etc [5]. The 'epigenome' refers to the complete description of these heritable changes [7]. Human beings and other multicellular organisms, all develop from a single cell, i.e. zygote. The descendent cells of the zygote all share the same set of DNA sequences while having distinct morphologies and providing various functions in different organs and tissues. The distinct epigenome inside each cell type mediates the same set of DNA sequences to act differently.

1.1.2 What is DNA methylation

DNA methylation is the most well-studied epigenetic modification, not only because it plays important role in mammal development by participating in various biological processes, such as repressing gene expression [8], silencing transposable elements [9, 10], female X chromosome random inactivation [11] and genomic imprinting [12], but equally important is its relatively stable and easy to be quantitatively measured characteristics.

DNA methylation is a chemical modification to the DNA molecule, typically, a methyl group (CH3) is covalently attached to the fifth carbon (5C) of cytosine residue to form 5-methylcytosine (5mC) [13]. 5mC was first found in bacteria by Johnson and Coghill in 1925 [14], thereafter, 5mC has been revealed to exist in all domains of life, including bacteria, plants and animals. In mammals, 5mC is predominantly found within the cytosine-guanine dinucleotide (CpG) context. The CpG distribution is nonrandom and the majority of the genome is CpG-poor. The majority (80%) of CpG sites across the genome of mammal somatic cells are methylated [15]. The unmethylated CpGs are predominately located in CpG islands, which are defined as stretches of DNA sequence (around 1000 bp) with a high density of CpG dinucleotides [16], gene promoters are generally associated with CpG islands, with around 70% of gene promoters reside within CpG islands [17]. The methylation CpG island usually leads to stable silencing of gene expression [18]. Generally, all CpGs are categorized into four classes according to their distance to CpG island, the four CpG classes include CpG islands (inside the island), CpG island shores ($\leq$ 2kb from an island) [19], CpG island shelves (2$-$4kb from an island) [20] and CpG open seas ($\geq$ 4kb from an island) [21].

1.1.3 DNA methylation dynamics

The methylation of CpG is catalyzed by a family of DNA methyltransferases (Dnmts). The de novo methylation is mediated by DNA methyltransferase 3A (Dnmt3a) and 3B (Dnmt3b) [22], they both share a similar structure and function. *Dnmt3a* is ubiquitously expressed and easily detectable in most adult tissues, whereas *Dnmt3b* is poorly expressed within most tissues, except bone marrow testis and thyroid [23]. When cell replicates, Dnmt1 binds to the newly synthesized hemimethylated DNA to replicate the original DNA methylation patterns [24].

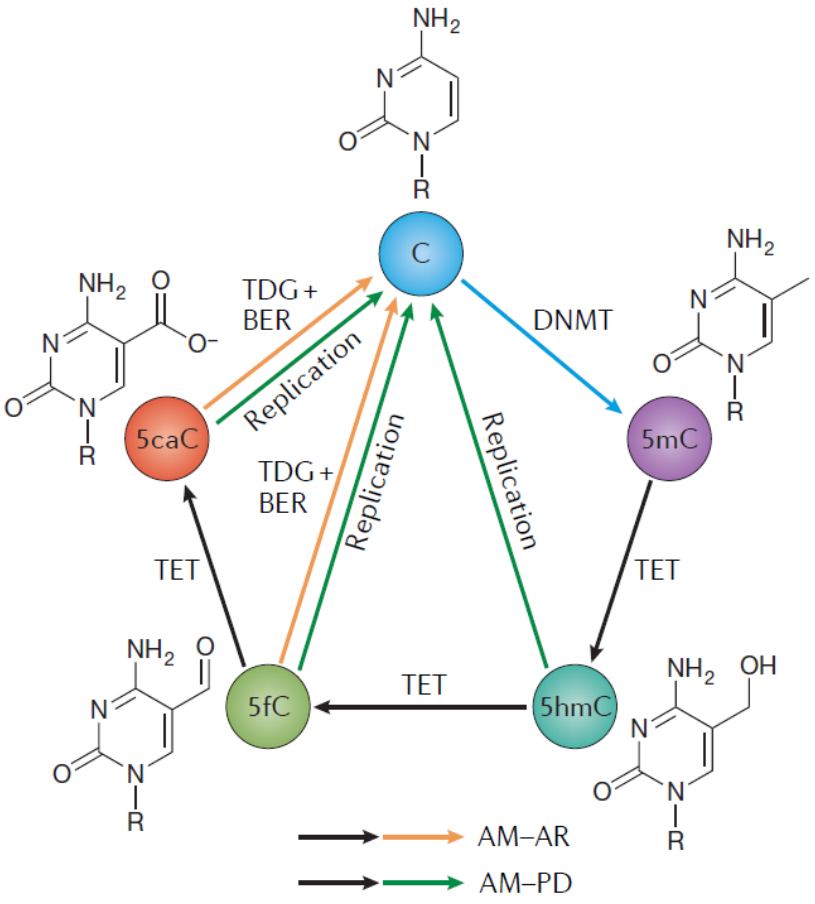

Figure 1.1: The cycle of active DNA demethylation, this diagram is adopted from [25].

After methylation patterns are established, the methylated cytosines can reverse to their

original unmethylated state by either passive dilution when DNA replicating without effective methylation maintenance machinery, or actively, even in non-replicating cells, demethylating to unmethylated cytosine via the TET-TDG pathway (Figure 1.1) [25]. The dynamic process of DNA methylation and the plasticity of the DNA methylation landscape make genes responsive to environmental exposures. Several health and lifestyle factors have been found to be associated with DNA methylation signatures, including childhood disease, tobacco smoke, drug use and poor nutrition [26, 27, 28].

1.1.4 Measure DNA methylation

The methylation status of any cytosine residue across the genome in any single cell is binary, i.e. either methylated or unmethylated. Therefore, the methylation level of a CpG for any specific tissue that is comprised of multiple cells is determined as the percentage of methylated cells for that locus. To quantitatively measure methylation levels, bisulfite conversion is usually the first step to distinguish the methylated and unmethylated cytosine. Bisulfite conversion involves deaminating the unmethylated cytosines into uracils while keeping the methylated cytosines, both 5-methylcytosine and 5-hydroxymethylcytosine, unchanged. In the following PCR steps, the methylated cytosine will be recognized as cytosine, while the uracil will be recognized as thymine.

During the past decades, Epigenome-wide Association Studies (EWAS) has been a popular technique to discover novel associations between lifestyles or environmental exposures and alterations in epigenome or methylome. While whole genome bisulfite sequencing (WGBS) is recognized as the gold standard to measure the methylation patterns across the human genome, the high cost and technical complexity still pose significant challenges that prevent application to large-scale samples [29]. DNA methylation microarrays, such as Infinium Hu-

manMethylation450 BeadChip [20] and Infinium MethylationEPIC BeadChip [30], provide cost-effective and high-throughput measurements of the methylation status for over half a million CpG sites across the genome will continue to be the first choice by most DNA methylation related large cohort studies in the near future.

1.2 Ageing and DNA methylation clocks

1.2.1 Ageing and ageing biomarker

The world's population is ageing at an ever-fast pace. According to World Health Organization, the number of people aged 60 years and older has outnumbered children younger than 5 years by 2020 [31]. The rapid increase in the elderly population is posing a socioeconomic challenge to societies all over the world. Ageing is characterized by progressive loss of cellular functions, leading to increased risk of morbidity and mortality [32]. Organismal ageing has significant importance for human health because it increases susceptibility to many diseases, such as diabetes, cardiovascular disorders and neurodegenerative diseases [33]. In recent years, many drugs and interventions [34, 35, 36], such as calorie restriction, rapamycin, metformin, NAD+ supplements and exercise, have been taken to clinical trials with the hope they may delay the ageing process or even restore young capacity. However, a significant challenge still exists in the field which is how to accurately measure the ageing status. People may age at different rates, largely influenced by genetic background, lifestyle and environmental exposures. An ideal age biomarker should thus capture this difference. Chronological age, which is the number of years a person has been alive, can serve as an objective approximate description of how well people functions. However, the nature of chronological age means it only steadily increases at the same rate for everyone, no matter

what the subject's health status is, making chronological age can not be served as a useful biomarker for anti-ageing clinical trials. Thus the ageing field urgently requires a reliable biomarker to measure a person's biological age—although there is a lack of a precise and widely accepted definition, biological age is often referred to as a quantity describing the person's true global ageing state [37]. Further investigation of ageing biomarkers will not only increase our knowledge of the mechanisms of ageing, but also facilitate monitoring the various interventions for improving human healthspan and rejuvenation experiments.

1.2.2 A variety of ageing biomarkers

In 2013, a highly influential review summarised nine important hallmarks of ageing from cellular and molecular levels, including stem cell exhaustion, cellular senescence, mitochondrial dysfunction, deregulated nutrient sensing, loss of proteostasis, telomere attrition, epigenetic alterations, genomic instability and altered intracellular communication (Figure 1.2) [32]. In the last decade, a variety of ageing biomarkers that were derived from different ageing hallmarks, such as telomere attrition [38], DNA methylation changes [39, 40] and alterations in gene expression [41, 42] and metabolite concentration [43, 44], have attracted even more attention and were used to build age estimators or age clocks attempt to measure the biological age [45].

Relative leukocyte telomere length is one of the first biological phenomena that showed promising potential as biomarkers of biological ageing [46, 38]. Telomeres are repetitive DNA-protein complexes located at the ends of chromosomes. They shorten every time cells divide, when their telomere length reaches a critical length, cells stop dividing or die. However, telomere length and chronological age are only loosely correlated, with Pearson's correlation coefficient usually under 0.5 [47], thus it can not provide accurate and reliable age estimations

Figure 1.2: The nine hallmarks of ageing, this diagram is adopted from [32].

on an individual level. Studies have reported thousands of genes differentially expressed with chronological age [48, 49], as a result of this, several transcriptome age clocks were developed and reported the deviations between the transcriptomic age and the chronological age are associated with several clinical features [48, 41, 50, 42]. However, the transcriptome age predictors also suffer from an accuracy issue, with mean absolute errors usually greater than 5 years [45]. Similarly, there are age estimation models trained on metabolomic data, the so-called metabolomic ages also showed strong chronological age correlation and the residuals of the predicted metabolomic age also demonstrated to be associated with several clinical

phenotypes [43, 44]. In contrast, age clocks based on DNAm changes, also called epigenetic clocks, were demonstrated to be the most accurate and robust age estimators, they are the most promising ageing biomarker that can be applied to individuals [51].

1.2.3 DNAm clocks

The discovery of CpG methylation as a potential epigenetic inheritance mechanism has led to speculations that it may be involved in the ageing process [52]. Early studies relying on techniques measuring total 5-mC content reported a subtle and gradual loss of methylation in different animals and occurs in several different tissues, such as the brain, liver, heart and T cell [53, 54, 55, 56]. With modern technology development, especially microarray and sequencing-based technology, massive measuring base-level methylation statuses are widely adopted to discover phenotype-associated methylation changes in CpG sites across the genome. Since then, age-related DNA methylation changes have been found to be widespread across the genome, throughout the life course [57, 39, 58, 59] and exist in a wide variety of tissues [40, 60]. Depending on the gene and the tissue, methylation changes with age can be either positive or negative, i.e. hypermethylated or hypomethylated. A well-reported hypermethylated example is the CpG site targeted by the probe of cg16867657, locates at the promoter of *ELOVL2* gene, and consistently gains methylation with age across many different tissue types [60]. It was also reported as the top age-associated CpG in the blood by different studies [61, 62], with its methylation level increasing from around 20% to nearly 90% when the person grows from birth to 100 years old. Whereas the locus targeted by the probe of cg10501210, located upstream of *miR-29b-2* gene, with its methylation level in blood close to 100% in newborns and gradually decreases to about 30% when the subject grew to 100 years old [63]. Generally, CpG sites exhibiting similar age-associated methylation changes across multiple tissues are more likely to be hypermethylated, whereas those hypomethylated

age-associated CpG sites are mostly tissue specific.

The discovery of many highly age-associated CpGs has led researchers to work on building chronological age prediction models by simply weighted averaging the methylation values of a group of age-associated CpG sites. As a first attempt, Sven Bocklandt and colleagues built a first DNAm-based age prediction model by including only two CpG sites and they reported the model explains 73% of the variance in age within a small dataset [64]. Since 2013, many more DNAm-based age models have been published and those later models achieved much higher age prediction accuracy by including much more informative CpGs, from dozens to nearly a thousand, and were trained on much larger datasets, with sample sizes ranging from several hundred to ten thousand. In this field, all kinds of age prediction models are also called epigenetic age clocks. Among them, Hannumn's clock [39] and Horvath's clock [40] are the two most widely known DNAm clocks. Hannum et al. built the first multivariate DNAm age model which included 71 CpG sites and the model was trained on microarray samples of the whole blood from 656 adult human individuals [39]. In the same year 2013, Steve Horvath trained an age model on 8,000 microarray samples comprising 51 healthy tissues and cell types, the finalised Horvath multi-tissue clock demonstrated a relatively accurate age prediction (median absolute error of 3.6 years) is possible for a broad range of human tissues and cell types via a single linear model that includes only a small number of CpGs (353 CpG sites) [40]. Since then, many more age clocks have been published [65, 66, 67]. Different tissue types not only have different methylation profiles, they may also have distinct age-related DNA methylation change patterns, therefore many tissue-specific clocks have been developed and demonstrated better age prediction performance than a single multi-tissue clock. Until now, tissue-specific clocks have been developed for skeletal muscle [68], buccal cells [69], brain cortex [70], skin [71] and so on.

So many different clocks were developed not just to estimate the subject's chronological

age accurately, even though this may be useful for forensic scientists to infer an individual's age only from a biological sample, many researchers are more interested in evaluating the ageing status of the tissue they studied by applying the epigenetic clocks. Many researchers believed the epigenetic age residues—the difference between the epigenetic age estimated by DNAm clocks and the actual chronological age, are not just errors derived from the under-performance of poor models, but represent a significant proportion of biological meaningful ageing signals. The epigenetic age residue is better known as epigenetic age acceleration in this field, a straightforward interpretation is, that the epigenetic age is a kind of approximation for biological age, thus a positive age acceleration means the person is biologically older than his actual age, while a negative age acceleration indicates the person is biologically younger than his age. Motivated by this premise, several studies have reported significant associations between epigenetic age acceleration and disease or health-related issues, such as obesity [72], HIV infection [73], Down syndrome [74], Huntington's disease [75] and Werner syndrome [76]. A more striking finding is from Marioni and colleagues, who reported a higher age acceleration derived from both Hannum's clock and Horvath's clock is significantly associated with a higher all-cause mortality risk [77]. However, with the development of an ever-accurate age prediction model for blood samples by leveraging the largest ever training dataset, Zhang et al. demonstrated that the association between epigenetic age acceleration and mortality risk decreased to non-significant by applying epigenetic age models with improved accuracy of chronological age prediction [78]. A recent study also reported no significant associations were found between longitudinal functional capacity assessments and age accelerations derived from several epigenetic clocks [79].

All the above-mentioned age clocks were built by using chronological age as the only dependent variable within the training regression algorithms, however, when further improving the precision of chronological age prediction, the application of using the estimated epigenetic age as a biomarker of biological age inevitably runs into the well-known "paradox

of biomarkers" as recomposed by Hochschild: "A hypothetical biomarker that approaches perfect correlation with chronological age could be replaced by chronological age and would be insensitive to differences in ageing among individuals." [80, 37]. As a way to walk around the paradox and also to disentangle biological age from chronological age, some researchers turned to use mortality risk score rather than chronological age as the dependent variable to regress, when training clocks. In 2018, Morgan et al. incorporated nine mortality risk-associated clinical biomarkers and chronological age into a mathematic function as an estimator of phenotypic age, then the phonotypic age was regressed on over 20,000 CpGs across the genome, the finalized model also named PhenoAge comprised of 513 CpGs in which their weighted average of methylation levels are taken as estimation of the phenotypic age [81]. Subsequently, Lu et al. developed GrimAge, a DNAm-based clock aimed to better predict mortality risk, which was built by regressing time-to-death due to all-cause mortality on eight DNAm-based estimators for seven plasma protein levels and smoking pack years, chronological age and sex [82]. Both PhenoAge and GrimAge were reported to better predict ageing-related outcomes, especially all-cause mortality than previous clocks regressed on chronological age [83, 84]. Despite many DNAm-based clocks have been developed and applied to predict ageing-related issues, it is still largely unknown whether the methylation changes of CpGs used to build the clocks are a reflection of other underlying molecular or cellular processes, or whether they themselves are involved in the ageing process [85, 86].

1.2.4 Motivations

Although many DNAm-based clocks reported high accuracy in predicting epigenetic age in terms of the high Pearson's correlation coefficients between estimated DNAm age and chronological age, the lack of robustness of existing DNAm clocks means they may make volatile predictions for samples from even technical replicates [87, 88, 89]. The lack of

robustness may be due to the performances of the probes used to construct the clocks being easily affected by technical variances, such as sample preparation procedures, probe hybridization issues and batch effects [90, 91, 88]. Thus, the large technical noise existing in epigenetic age predictions by existing DNAm clocks means they can only be instructive and meaningful at population levels. To increase the reliability of DNAm clocks, Higgins-Chen and colleagues tried to remove those with low signal-to-noise ratio CpGs in the training model, but only achieved modest improvements [89]. Recently, they proposed to use principal components (PCs) of methylation values of 78,464 CpGs as input to train clocks and reported high reliability of the PC-based clocks [89]. Initially, I was thinking to combine big data and deep learning algorithms to greatly improve the accuracy and also reliability of DNAm clocks. Most previously published clocks adopted penalised linear regression algorithms, such as Elastic Net, to select a group of informative CpGs and then weighted averaging their methylation values as prediction results. However, the linear regression algorithms can not fully take advantage of a large number of informative CpGs whose methylation levels are nonlinearly correlated with age. Such as Vershinina et al. reported that methylation levels of 15% of age-associated CpGs are nonlinearly changing with ageing [92]. In contrast, neural networks based on deep learning algorithms have proved to be able to fit into any complex nonlinear patterns provided with enough neurons and hidden layers [93]. Most previous clocks were trained on several hundred to several thousand samples which are far less than the available age informative CpGs (more than 40,000 [89]) measured by 450k array or EPIC array, thus greatly increasing training sample size will alleviate the over-fitting issues in the training process. To this end, I collected more than twenty thousand DNA methylation microarray samples from public repositories.

The collected samples were produced from different laboratories and in different years, it thus poses a challenge to confidently integrated them into any downstream analysis. Sex has previously been reported to have a strong impact on DNA methylation variation [94, 95,

96]. Females have higher life expectancy than males across countries worldwide with females generally living 3 to 7 years longer than males [97], it has also shown that males have faster epigenetic ageing rates compared to females [39, 98]. It is generally suggested practice to include sex as a covariant in EWAS analysis. However, many DNAm datasets deposited in public repositories do not include sex annotations. It thus promoted me to construct a robust sex classifier to estimate sex for DNAm samples. After assigning sex to unannotated samples and removing mismatched samples, the DNAm samples from different datasets have to be normalized to remove most of the technical noises. When I was choosing which method to normalize the DNAm datasets, I found existing quantile-based normalization methods all introducing artifactual sex bias into normalized data, when processing female samples and male samples together, therefore, I was motivated to address this issue by creating a new normalization strategy, the detail solution is documented in Chapter 3. Afterwards, clean and normalized datasets have been generated, tissue or cell types have to be considered when choosing DNAm samples to construct epigenetic age clocks, DNA methylation plays a vital role in helping differentiated cell types maintain their distinct identities. It was previously reported that the cerebellum ages slower in comparison to other tissue types, however, this claim was drawn from a single clock using a small sample size and so warrants further investigation. In recent years, many more cerebellum DNA methylation samples have become publicly available and many diverse DNAm age clocks have also been developed [86]. In Chapter 4, the claim that the cerebellum ages slowly is thoroughly examined, and the mechanisms are explored.

1.3 Aims of this Thesis

1. Chapter 2 describes a new sex classifier to accurately estimate the sex of DNA methylation samples, this method can be used to make sex annotations for unlabelled samples and also used to identify wrong-labeled samples when encountering sex mismatches.

2. Chapter 3 presents a novel two-step strategy called interpolatedXY to normalize DNA methylation microarray data avoiding sex bias. In addition, it also demonstrates how artifactual sex bias is introduced into normalized data by traditional methods and further justifies the benefits of applying between-array normalization methods.

3. Chapter 4 reexamines the claim that the cerebellum ages slowly. It presents ageing rate comparisons for the cerebellum by epigenetic clocks and reveals how the epigenetic ages of cerebellum samples are severely underestimated.

Chapter2

DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy

The work presented in this chapter has been published in *BMC Genomics* [99].

Statement of Contribution: Yucheng Wang, the author of the thesis, originally conceived and developed the method. Yucheng Wang wrote the codes and performed all the analyses. Yucheng Wang wrote the manuscript. Xiaojun Zhai, Klaus D. McDonald-Maier and Leonard C. Schalkwyk advised and oversaw the work. Xiaojun Zhai, Klaus D. McDonald-Maier, Leonard C. Schalkwyk, Eilis Hannon, Olivia A. Grant, Tyler J. Gorrie-Stone, Meena Kumari and Jonathan Mill provided insights into writing the manuscript and interpreting the results.

2.1 Introduction

Epigenome-wide Association Studies (EWAS) are a powerful way to study the relationships between epigenetic variation and human diseases [100]. Apart from sex chromosomes, thousands of CpG sites on autosomes also show very different DNA methylation patterns between males and females [101, 102]. As a result of this, sex has been considered an important covariate, when undertaking methylation and phenotype association studies.

Many researchers have submitted their methylation microarray datasets to the Gene Expression Omnibus (GEO). Currently, there are over 100,000 HM450k samples and over 18,000 EPIC samples which are publicly available. Most of these have phenotype annotations accompanying them, thus they can be used by other researchers to perform meta-analyses or as independent references to validate their hypothesis. However, many mismatches have been found between annotations and samples, Toker et al. discovered widespread mislabelling in transcriptomics datasets of GEO [103], Heiss et al. found 25% of the datasets they studied contained sex-mismatched samples, particularly in three datasets, more than 30% of the samples were identified as being mislabelled [104]. A large portion of these discrepancies may stem from data entry errors. Researchers should deal with these sex-mismatched samples carefully; the safest way is to remove them directly before downstream analysis.

McCarthy and colleagues performed meta-analysis of sex-specific methylation patterns and demonstrated that the first two principal components of X chromosome methylation data on 27k arrays can differentiate between sexes [105]. Currently, there are several methods which can be used to predict the sex of samples from DNA methylation data. The 'getSex' function of *minfi* package estimates sex based on the median values of measurements on the X and Y chromosomes respectively [106]; the 'estimateSex' method of *sEst* package groups

beta values and detection p-values of probes mapped on sex chromosomes into different intervals and achieved sex prediction by looking at the different distribution patterns of these intervals from two sexes. [107]; The 'check_sex' method within the *ewastools* package predict sex based on normalized average signal intensity values on the sex chromosomes [104].

We propose a novel method to predict the sex of samples using solely DNA methylation beta values. We identify a set of significant sex-associated CpG sites, perform principal component analysis (PCA) on these sites to obtain a sex classifier, and evaluate our method's performance across a wide range of human tissues. The proposed sex classifier allows users to attribute sex to unannotated samples on public databases, and also identify samples with sex aneuploidy.

2.2 Methods

2.2.1 Data collection and preprocessing

We downloaded publicly available methylation microarray datasets from GEO (https://www.ncbi.nlm.nih.gov/geo/), for those datasets in which raw IDAT files were not available, such as GSE78874 and GSE137884, the intensity values of methylated and unmethylated signals were extracted from raw intensity text files. While for most of the datasets in which raw IDAT were provided, we used the function 'iadd2' from bigmelon package [2] to read and load intensity values from IDAT files. After that, beta values are calculated as:

$$\beta = \frac{M}{M + U + 100}$$

where β is beta value, M denotes methylated densities and U represents unmethylated densities. Beta values are ranged between 0 and 1, beta value close to 1 means high-level methylation and a near-zero beta value represents low-level methylation. With manual inspection, those samples with apparent abnormal beta value density distributions were removed prior to downstream analysis. Also, those samples with more than 10% missing data were excluded.

There are 453,152 probes that exist in both 450k array and EPIC array, therefore, we only keep the shared 453,152 probes for downstream analysis. For each sample, the missing values of each probe were replaced by their corresponding means across all samples. Then, Z-score normalization was applied to each sample separately to reduce technical variance, which means all beta values were transformed to their Z-score values by subtracting the mean of all autosomal beta values and then divided by the standard deviation of all autosomal beta values within a sample. Z-score transformed beta values were used to construct PCA models and were used to make sex predictions.

2.2.2 Model construction

GSE105018 was used to screen for sex-associated CpGs, it includes 1658 whole blood DNA methylation samples from participants in the Environmental Risk Longitudinal Twin Study, there are 826 female samples and 832 male samples in this dataset, with all participants aged at 18, among them, 1468 participants who were members of complete twin pairs (430 MZ pairs and 304 DZ pairs).

To identify sex-associated probes, T-test was applied to raw beta values of each of the 453,152 probes for the two sex groups, after Bonferroni multiple comparison correction, those

probes with p-value less than 0.01 and absolute beta value difference between sexes greater than 0.2 were selected as significant sex-associated probes.

In order to have equal ratios of sexes, we randomly selected 800 females and 800 males from GSE105018, the Z-score transformed beta values of the identified sex-associated probes mapped on sex chromosomes were used as input data. To be specific, the Z-score transformed beta values of the sex-associated probes mapped on X chromosomes were processed by PCA, and the coefficients of the first principal component were used in the final model to distinguish whether a sample contains one copy X chromosome or two copy X chromosomes. Similarly, the Z-score transformed beta values of the sex-associated probes mapped on Y chromosomes were processed by another PCA, and the coefficients of the result first principal component were used in the final model to distinguish whether a sample has Y chromosomes or not. As a result, the final model includes two sets of coefficients from two first principal components of two separate PCAs. Finally, the proposed sex classifier was tested by the UKHLS dataset, with the labelled sexes as true sex annotations.

2.3 Availability of data and materials

All the DNA methylation datasets except for the validation set analysed during the current study are publicly available and were obtained from the GEO public repository. The training set is from GSE105018[108] which includes 832 male and 826 female whole blood samples, the validation set which includes 1175 whole blood samples is available from the European Genome-phenome Archive under accession EGAS00001002836 (https://www.ebi.ac.uk/ega/home). Other datasets: purified blood cell types (GSE103541 [109]), buccal cells (GSE137884 [110]), brain cells (GSE112179 [111]), saliva (GSE78874 [112]), liver (GSE119100 [113]), placenta

(GSE100197 [114]), sperms (GSE64096 [115]). The one Klinefelter syndrome positive sample is available upon request. More details about these datasets are shown in Table 2.1.

2.3.1 Sofware

All the analyses were conducted in R (version 3.6.0) [116] under a Linux environment. The proposed method has been integrated into the *wateRmelon* Bioconductor package, which is freely and easily accessible by calling the 'estimateSex' function.

Table 2.1: Summary of datasets used in this study.

Dataset	Source	Platform	Number	Male/Female	Age(years)	Reference
GSE105018	Whole blood	450k	1658	832/826	18 - 18	[108]
UKHLS	Whole blood	EPIC	1175	489/686	28 - 98	[117]
GSE103541	Purified blood cells	EPIC	145	NA	NA	[109]
GSE137884	Buccal cells	450k	89	51/38	3 - 6	[110]
GSE112179	Brain cells	EPIC	100	75/25	23 - 77	[111]
GSE78874	Saliva	450k	259	146/113	36 - 88	[112]
GSE119100	Liver	EPIC	108	46/62	25 - 71	[113]
GSE100197	Placenta	450k	102	NA	NA	[114]
GSE64096	Sperms	450k	40	NA	NA	[115]
GSE51032	Buffy coat	450k	845	188/657	34 - 72	[118]

* GSM3562874 and GSM3667736 refer to the same case.

2.4 Results

2.4.1 Identifying sex-associated CpG loci

To make our method compatible with both 450K and EPIC, we only included 453,152 probes that are present on both arrays. Two-sample T-tests were applied to GSE105018 [108] to identify differentially methylated CpG sites between sexes, after Bonferroni multiple comparison correction, those with p-value less than 0.01 and absolute beta value difference between sexes greater than 0.2 were selected as the most significant sex associated CpG sites. As a result of this, we obtain 4345 significantly sex-associated sites. In this study, we have chosen a relatively strict threshold, as we aim to capture those most robust features which methylate differently and consistently between the two sex groups across various datasets. As expected, most of the sex-associated sites belong to sex chromosomes, with the majority (4047, 93%)located on the X chromosome (ChrX), and with a total of 284 (6.5%) CpG sites located on the Y chromosome (ChrY).

As shown in Figure 2.1a, these sex-associated CpG sites on ChrX are distributed throughout the whole chromosome, and most of them (3781, 93.4%) are associated with higher methylation levels in females compared to males, this is mainly because one X chromosome of the female is inactivated and highly methylated. However, we also observed a small portion of CpG sites (266, 6.6%) on ChrX that have higher methylation levels in males compared to females, this could attribute to the fact that around 15% of X-chromosome genes often escape from XCI and another fifteen percentage shows variable degree of 'escape' [119]. For example, four out of the 266 probes mapped to *Xist* which is an escape gene with known exclusive expression from the inactivated X chromosome [119].

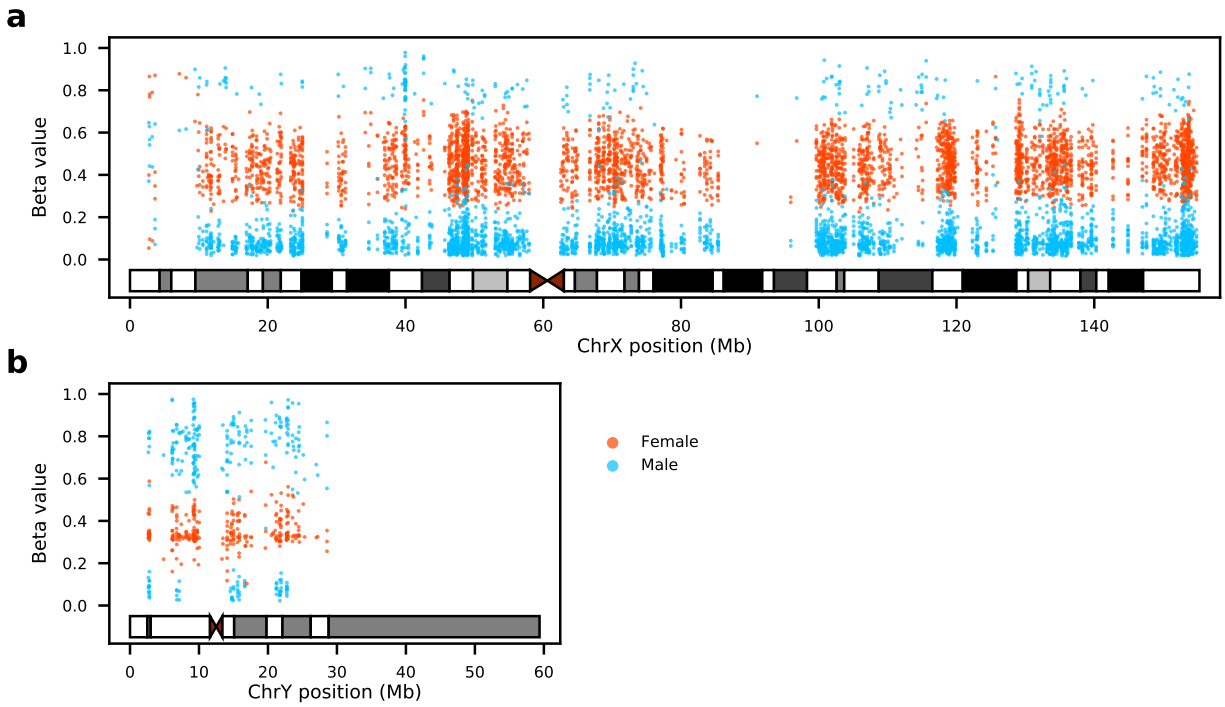

Figure 2.1: Females and males exhibit distinct methylation patterns at sex-associated CpG sites on the two sex chromosomes **a**: The X chromosome: most sex-associated CpG sites from females have beta values range between 0.2 and 0.8; most of these sites from males are less methylated (beta values less than 0.2). **b**: The Y chromosome: the identified sex-associated CpG sites of males are highly methylated with beta values greater than 0.6 whereas females exhibited low methylation signals.

Among the 284 sex-associated CpG sites on ChrY, 211 CpG sites have higher methylation levels in male samples (Figure 2.1b). Females do not carry Y chromosomes, thus most of the intensity signals of ChrY we observed from females may due to background noise and non-specific hybridisation, nevertheless, the mean raw signal intensities of the 284 probes in females are only around 11% of that in males. Interestingly, 70 of the 284 probes are on McCartney's list of 67,609 potential non-specific probes of EPIC array [120], however, 69 of them are hypermethylated in males (mean=0.73, sd=0.11), while hypomethylated in females (mean=0.35, sd=0.07). The raw signal intensities of the 70 probes in females are also only around 10% of that in males, suggesting they were less affected by the non-specific hybridisation issue.

2.4.2 Sex classifier based on sex-associated CpG sites

Since we have obtained a large group of CpG sites which show a significant difference ($p < 0.01$) in methylation levels between males and females, we are able to construct a sex classifier. To begin with, the DNA methylation values of the 4047 sex-associated CpG sites on ChrX from the same training samples are processed using PCA. PCA takes a linear approach to generate reduced dimensions by maximizing the captured residual variance in each further dimension[121]. As shown in Figure 2.2a, the first principal component, which explained 98% of the total variance, has captured the most sex differences among all training samples. Thus, we could use this first component to separate samples into two categories: 1) with two copies of X chromosomes and 2) with only one copy of X chromosome.

Similarly, a PCA is performed using the 284 CpG sites of ChrY, and as that of ChrX, the first principal component accounted for the most variances can make a good separation between male and female samples (Figure 2.2b). As a result of this, the first component can

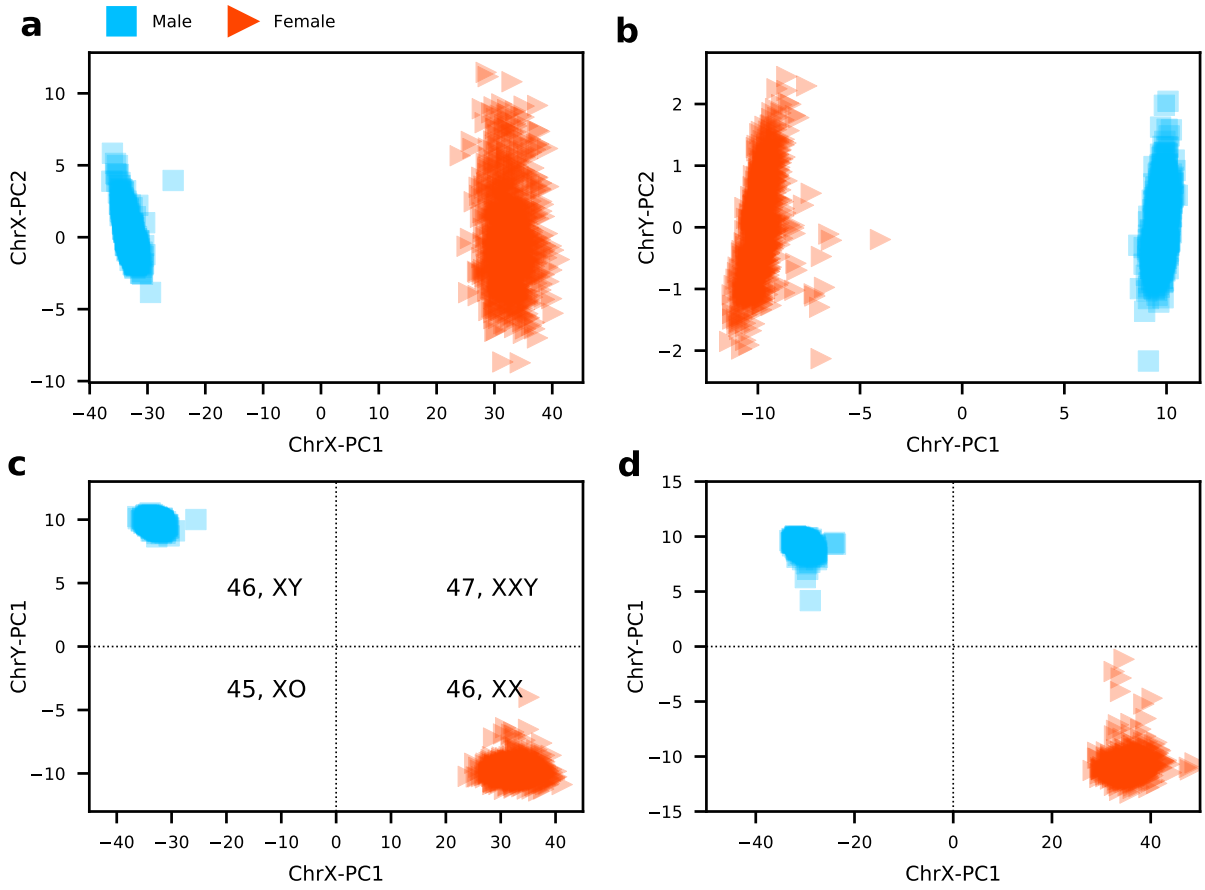

Figure 2.2: A sex classifier is constructed by applying two PCAs on two sex chromosomes separately. **a**: The first two components on ChrX. **b**: The first two components on ChrY. Results of **c** training set and **d** validation set produced by the sex classifier, all samples are classified into four categories: 46XY, 46XX, 47XXY, and 45XO.

be used to divide samples into two categories: 1) with Y and 2) without Y.

Finally, the two first principal components of the two PCAs which both explained the most sex differences are utilized to build the sex classifier. Normal females have two copies of X chromosomes and normal males have one copy of X chromosome and one copy of Y chromosome. By our sex classifer, male samples with 46,XY should locate in the top left area and female samples with 46,XX should distribute at the bottom right area (Figure 2.2c). It is reasonable to suggest that this model can be applied to identify samples with sex aneuploidy: samples with 45,XO will be placed at the bottom left corner, and samples with 47,XXY should be distributed at the top right corner.

The codes for the proposed sex classifier are listed below, this function is also available by calling 'estimateSex' from the *waterMelon* bioconductor R package (https://github.com/schalkwyk/wateRmelon/tree/master).

```
1
2 #' Predict sex by using robust sex-related CpG sites on ChrX and
     ChrY
3 #'
4 #' @param betas
5 #' A matrix with sample IDs as column names, and probe names as row
     names,
6 #' ideally: beta = M / (M + U + 100). Take a look at an example
     betas with:
7 #' "data(melon); print(betas(melon)[1:10, 1:3])".
8 #' @param do_plot logical. Should plot the predicted results?
     Default: FALSE
9 #'
```

```
#' @return dataframe contains predicted sex information.
#' @export
#' @author
#' Wang, Yucheng, et al. "DNA methylation-based sex classifier to predict sex
#' and identify sex chromosome aneuploidy." BMC genomics 22.1 (2021): 1-11.
#'
#' @examples
#' data(melon)
#' pred_XY <- estimateSex(betas(melon), do_plot=TRUE)
estimateSex <- function(betas, do_plot=FALSE){
  betas <- as.matrix(betas)
  single_sample <- FALSE
  if(ncol(betas) == 1) {
    betas <- cbind(betas, betas)
    single_sample <- TRUE
  }
  # predict sex by two PCAs on X and Y chromosomes
  data("sexCoef")
  # Z score normalization
  betas <- betas[rownames(betas) %in% sex_coef$IlmnID, ]
  message('Normalize beta values by Z score...')
  autosomes <- sex_coef$IlmnID[!(sex_coef$CHR %in% c('X', 'Y'))]
  auto_betas <- betas[rownames(betas) %in% autosomes, ]
  d_mean <- colMeans(auto_betas, na.rm=TRUE)
  d_sd <- colSds(auto_betas, na.rm=TRUE)
  z_beta <- (t(betas) - d_mean) / d_sd
  message('Fishished Zscore normalization.')
```

```
37
38  # Sex prediction
39  pred_XY <- list()
40  for(chr in c('X', 'Y')){
41    coefs <- sex_coef[sex_coef$pca == chr,]
42    miss_probes <- setdiff(coefs$IlmnID, colnames(z_beta))
43    if(length(miss_probes) > 0){
44      warning('Missing ', length(miss_probes), ' probes!\n',
          paste(c(miss_probes), collapse=", "))
45      coefs <- coefs[!(coefs$IlmnID %in% miss_probes), ]
46    }
47    chr_beta <- z_beta[, coefs$IlmnID]
48    chr_beta[is.na(chr_beta)] <- 0
49    pred_chr <- t(t(chr_beta) - coefs$mean) %*% coefs$coeff
50    pred_XY[[chr]] <- pred_chr
51  }
52  pred_XY <- data.frame(pred_XY)
53
54  pred_XY$'predicted_sex' <- 'Female'
55  pred_XY$'predicted_sex'[(pred_XY$X < 0) & (pred_XY$Y > 0)] <-
      'Male'
56  pred_XY$'predicted_sex'[(pred_XY$X > 0) & (pred_XY$Y > 0)] <-
      '47,XXY'
57  pred_XY$'predicted_sex'[(pred_XY$X < 0) & (pred_XY$Y < 0)] <-
      '45,XO'
58  if(single_sample){
59    pred_XY <- pred_XY[1, ]
60  }
61
```

```
62    if(do_plot){
63      plot_predicted_sex(pred_XY)
64    }else{
65      message('You can visualize the predicted results by set
          "do_plot=TRUE".\n')
66    }
67    return(pred_XY)
68  }
69
70  plot_predicted_sex <- function(pred_XY){
71    # visualization of predicted sex
72    plot(Y~X, data=pred_XY, pch=1, xlab='ChrX-PC1', ylab='ChrY-PC1')
73    abline(v=0, lty='dashed')
74    abline(h=0, lty='dashed')
75    abnormls <- pred_XY[!(pred_XY$'predicted_sex' %in% c('Male',
        'Female')),]
76    if(nrow(abnormls) > 0){
77      points(Y~X, data=abnormls, pch=2, col='red')
78      for(i in 1:nrow(abnormls)){
79        text(abnormls$X[i], abnormls$Y[i], rownames(abnormls)[i],
            pos=3, col='red', cex=0.5)
80      }
81    }
82    text(-10, 2, '46,XY', cex=1.2, col='blue')
83    text(-10, -2, '45,XO', cex=1.2, col='blue')
84    text(10, -2, '46,XX', cex=1.2, col='blue')
85    text(10, 2, '47,XXY', cex=1.2, col='blue')
86  }
```

2.4.3 Comparison with other tools

To compare the proposed sex classifier with three other existing sex prediction classifiers for DNA methylation microarray data taken from the R packages (see Table 2.2), *minfi* [106], *ewastools* [104] and *sEst* [107], we take GSE51032 [118] as a benchmark dataset, as it was used in developing *ewastools* and *sEst*. GSE51032 includes 857 samples (188 men and 657 women) and their source tissue are all from buffy coat. Figure 2.3 shows the results generated by the four methods, as we can see, there are eight samples (four males and four females) displaying mismatches between predicted sex and labelled sex, and the mismatches are consistent in the results from four methods, thus we have high confidence that the eight samples are mislabelled. Two samples (marked by black circles) are identified by our classifier as 47,XXY, *sEst* also identified the two outliers. However, only one of the two samples appears as an outlier from *minfi* and *ewastools*, and the other one stays close to the main male cluster.

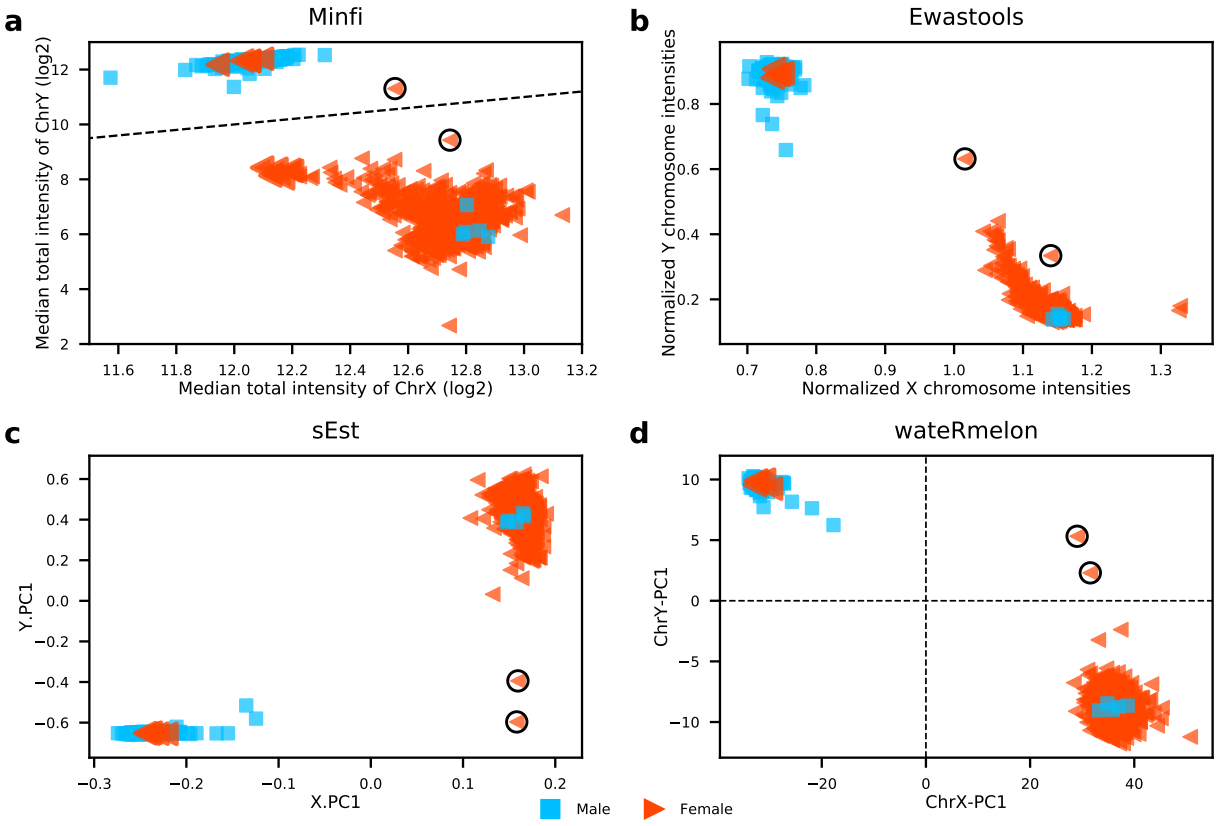

Figure 2.3: Comparisons of sex prediction ability between four tools. **a**. minfi, **b**. ewastools, **c**. sEst, **d**. our classifier in wateRmelon. Two outlier samples are marked by black circles, blue square represents male and red triangle denotes female.

Table 2.2: Summary of four sex prediction tools for DNA methylation samples.

Package	Function name	Input requirements	Mechanism	Performance on clustering females and males	SCA detection
Minfi	getSex	IDATs	Compare the log2 transformed median total intensity of probes mapped on ChrX and ChrY.	Good in clustering males and less well in clustering females	Not provided
Ewastools	check_sex	IDATs	Compare the normalized average signal intensity of probes mapped on sex chromosomes	Excellent in clustering males and good in clustering females	Not provided
sEst	estimateSex	Beta values and detection p-values	Group beta values and detection p-values into defined intervals and PCAs on the distribution patterns of these intervals.	Excellent in clustering males and females	Proposed but not validated
WaterRmelon	estimateSex	Beta values (which can be easily generated from signal intensity text files or IDATs)	PCAs on beta values of sex differently methylated CpGs on ChrX and ChrY separately.	Excellent in clustering males and females	Proposed and validated by five Turner syndrome cases and one Klinefelter syndrome case

In general, all four methods show good performance in clustering male samples, however the method from *minfi* performs much poorer in clustering female samples compare to the other three tools, as some females are not distinguishable from males along the x-axis. The female cluster produced by *ewastools* exhibits a long tail towards the male cluster; the sex prediction tools in *minfi* and *ewastools* are both based on signal intensity therefore they produce more similar results than the other two tools. Our sex classifier and the method from *sEst* are both beta value-based, although the two methods utilised beta values very differently and *sEst* requires detection *p*-values, the patterns of their results are similar. It should be noted, detection p-values are used as an index of usability for each probe but are not well defined. It is implemented as a test for signal intensity above background level in the proprietary GenomeStudio software, the detection *p*-values calculated by the minfi package are better documented but not equivalent. Overall, compared to the other three sex prediction tools, our proposed method is highly robust and shows better or similar performance in clustering females and males.

2.4.4 Performance evaluation

The DNA methylation profiles of samples from the training set and validation set are assessed by 450k array and EPIC array respectively. As we can see from the results (Figure 2.2), the proposed model has correctly classified all samples in the two datasets, proving that the proposed classifier is highly robust and compatible with both platforms.

The proposed sex classifier is trained and validated using whole blood samples. As whole blood is a heterogeneous collection of different cell types, to investigate whether our classifier is biased by blood cell types, we tested its performance on DNA methylation data derived from five purified blood cell types—B cells, CD4 T cells, CD8 T cells, monocytes and

granulocytes from 28 individuals. As shown in Figure 2.4a and Figure 2.4b, all the five cell types are clustered into two sex groups and we could not find any or very minor differences between cell types. Collectively, these results suggest that the proposed sex classifier is robust to blood cell types.

Although blood is the most studied tissue in EWAS, there are also many DNA methylation studies that use samples from other types of human tissue. To evaluate our sex classifier's range of application, we further tested its performance on several other most studied human tissues, including saliva, buccal cells, brain cells, liver, placenta, and sperm. Results from Figure 2.4c to Figure 2.4f demonstrate that the proposed classifier is robust in these vastly different types of tissues—saliva, buccal cells, brain cells, and liver. However, even though we can observe two clusters within the placenta samples, the female samples are more loosely distributed along the x-axis than that in other tissues, and all of them are more close to the zero point of x-axis, with several samples even having negative values (Figure 2.4g).

Interestingly, all sperm samples were clustered into a single group by our sex classifier, located in the bottom left region (Figure 2.4h). This area is typically recognised by our sex classifier as 45,XO. As sperm cells are a mixture of two types of haploid cells (23,X and 23,Y) this suggests that their methylation levels are lower on ChrY compared to other mature human tissues.

2.4.5 Predicting sex chromosome aneuploidy

DNA methylation has been an important way to study the various developmental symptoms caused by copy number aberrations of the sex chromosome [122]. Earlier, we proposed that our classifier can be applied to identify samples with abnormal sex chromosomes, including

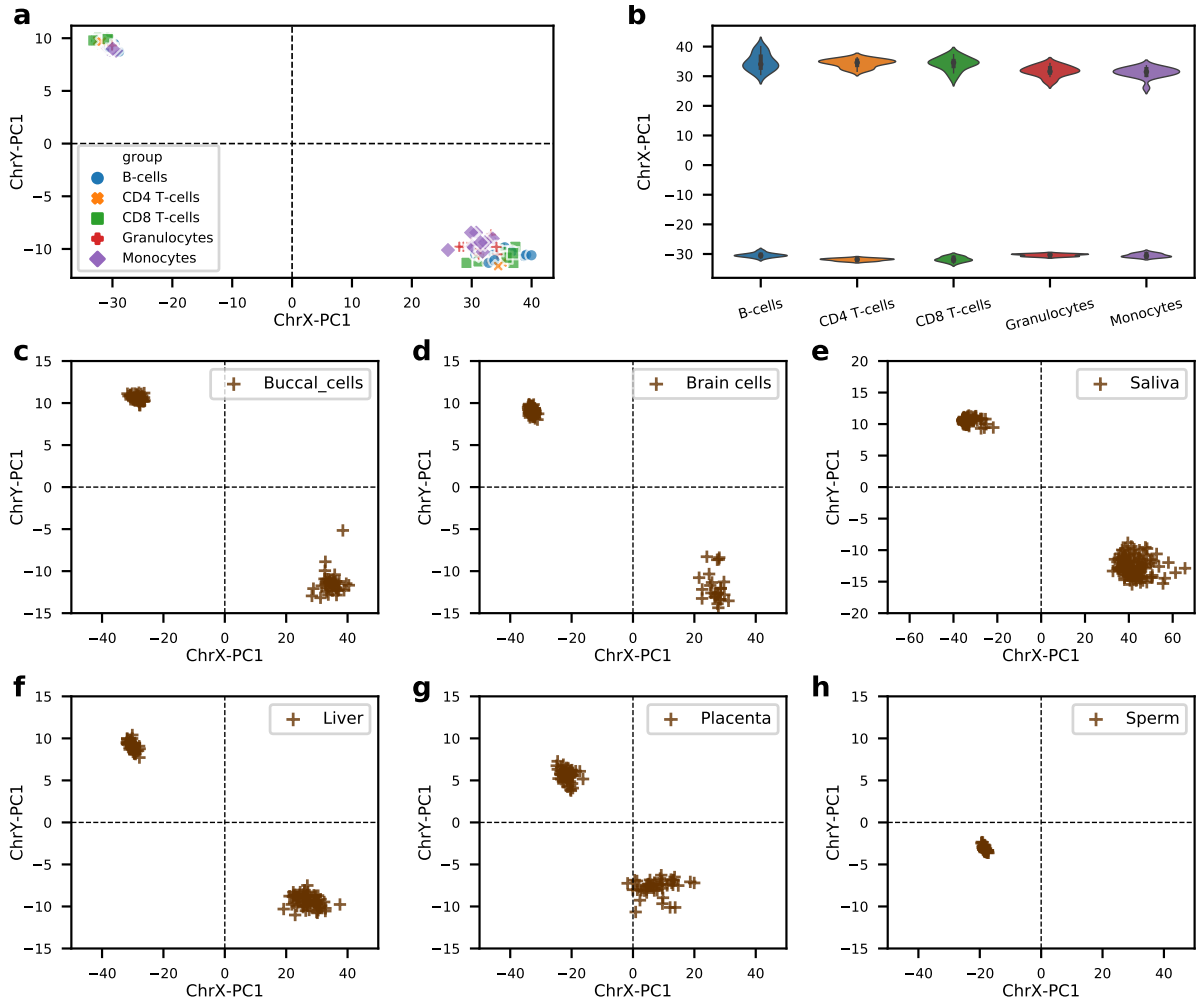

Figure 2.4: The sex classifier was evaluated across five blood cell types (**a** and **b**) and six other human tissues (**c**-**h**). **a**. Scatter plot showing results from five blood cell types: B cells, CD4 T cells, CD8 T cells, monocytes and granulocytes. **b**. On X chromosome, the five blood cell types showed similar results. **c**. Buccal cells; **d**. Brain cells; **e**. Saliva; **f**. Liver; **g**. Placenta; **h**. Sperms.

45,XO and 47,XXY. To further validate its ability, we searched the public repositories for positive samples with clinical diagnoses. As a result, we obtained five cases (Table 2) diagnosed with Turner syndrome from two studies [123, 124]. As hoped, they are all clearly classified as 45,XO by our model (Figure 2.5), proving our classifier's ability to predict females with only one X chromosome.

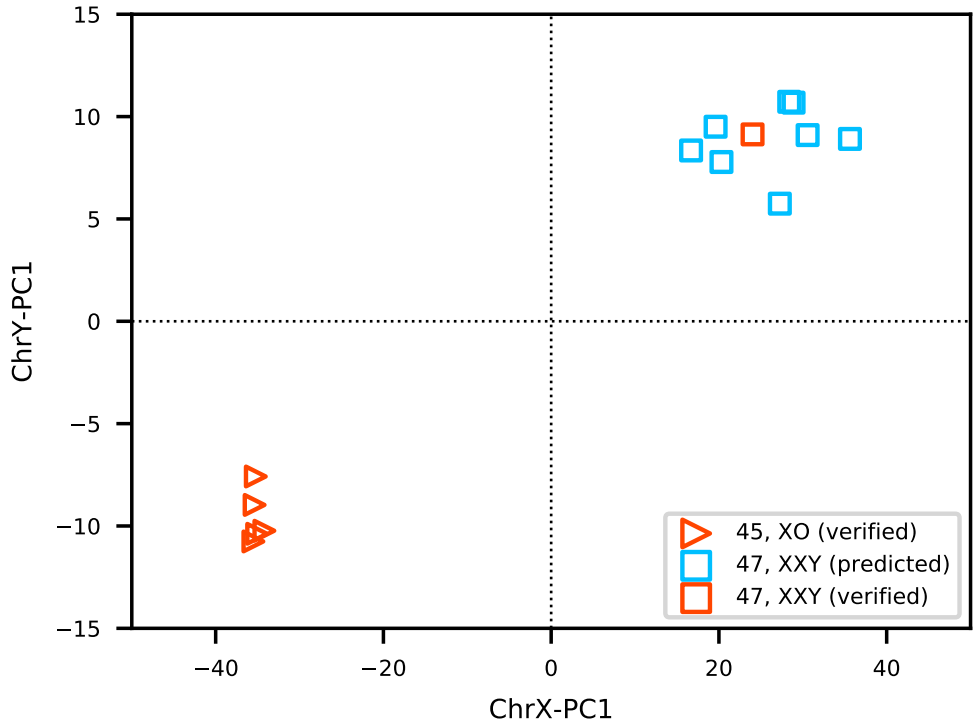

Figure 2.5: The proposed classifier is verified its ability to predict sex chromosome aneuploidy in five Turner syndrome samples and one Klinefelter syndrome case, it also predicted eight potential 47,XXY cases from GEO.

Viana et al. reported a male with schizophrenia carrying an extra X chromosome [125] which is also clearly classified as 47,XXY by our method (Figure 2.5). Unfortunately, we did not find any publicly available DNA methylation samples from those diagnosed with Klinefelter syndrome. Unlike Turner syndrome, most patients with Klinefelter syndrome have only mild symptoms and are never diagnosed. It is interesting to check if there are any samples in GEO having a karyotype of 47,XXY but not linked to a diagnosis. By applying our classifier to scan the GEO datasets, we find a total of eight samples (Table 2.3) which are

highly likely to be 47,XXY (Figure 2.5). It should be noted that we only include these samples sourced from blood or brain cells related tissues and their DNA methylation level are assessed by 450K or EPIC arrays; we also do not include those samples located near the boundaries which may be low-level sex chromosome mosaics (46,XX/47,XXY). It is interesting that two of the eight suspected abnormal samples were diagnosed with schizophrenia. Martin et al. found that Klinefelter patients have nearly a four times higher risk of schizophrenia [126], which may explain why we have predicted more 47,XXYs with schizophrenia. Studying the methylation patterns of these syndromes will provide more insights into these diseases.

Table 2.3: Samples with verified or suspect abnormal karyotypes from GEO.

Sample ID	Karyotype	Verified karyotype?	Source tissue	Disease status	Reference
GSM1566904	45,XO	Yes	Peripheral Blood	Turner syndrome	[123]
GSM1566905	45,XO	Yes	Peripheral Blood	Turner syndrome	[123]
GSM1566906	45,XO	Yes	Peripheral Blood	Turner syndrome	[123]
GSM1566907	45,XO	Yes	Peripheral Blood	Turner syndrome	[123]
GSM1572595	45,XO	Yes	Whole Blood	Turner syndrome	[124]
3999215192_R06C02	47,XXY	Yes	Prefrontal cortex	Schizophrenia and Klinefelters syndrome	[125]
GSM3562874 (GSM3667736)*	47,XXY	No	Whole blood		[127]
GSM1649023	47,XXY	No	Whole blood		[128]
GSM1946555	47,XXY	No	Whole Blood	Post-traumatic stress disorder	[129]
GSM3662121	47,XXY	No	Blood	Lynch-like syndrome	NA
GSM1344329	47,XXY	No	Peripheral blood		[130]
GSM2336820	47,XXY	No	CD8+ T-cells	Ulcerative colitis	[131]
GSM3680912	47,XXY	No	Frontal cortex	Schizophrenia	[132]
GSM1496810	47,XXY	No	Frontal cortex	Schizophrenia	[133]

* GSM3562874 and GSM3667736 refer to the same case.

2.5 Discussion

There are two principal reasons to require a good and simple sex classifier based on methylation data. First, there are still many samples in GEO that do not have sex annotations, thus an accurate classifier can provide reliable sex information. Second, due to data entry errors, there are non-negligible proportions of mislabelled samples in the public database. A mismatch between reported sex and predicted sex would be a clear indication of a wrong annotation and introduces doubt on the accuracy of the rest of the phenotype information for that sample, hence it is reasonable to remove these mislabelled samples before downstream analyses. We would recommend sex checking to be a standard part of all DNA methylation QC pipelines. Here in this study, the proposed sex classifier is straightforward and the outcomes are highly intuitive.

In this study, we first obtained a group of significant sex-associated CpG sites. 90% of these located on the X chromosome are more methylated in females than that in males, this is mainly due to the effect of X-chromosome inactivation: one of the two X chromosomes in females is randomly chosen for inactivation (highly methylated) to balance the extra gene expression dosage [134, 135]. This also justified that our classifier was built on blood samples that could work well across a wide range of other tissue types.

The proposed sex classifier shows robust performance across a wide range of tissue types despite it being built upon whole blood samples. We choose blood samples because they are easily accessible and are the most widely used tissue for measuring DNA methylation and have been adopted in most large cohort studies. However, whole blood is a heterogeneous collection of different cells, and their cell composition changes across age [136]. Different cell types can have distinct methylation profiles even though they share identical genetic

makeup [137]. Here as our results have shown that the proposed model is not biased among different blood cell types; we also demonstrated the proposed classifier performs well across a wide range of human tissues, including saliva, buccal cells, brain cells and liver. These results suggest that our model is not driven by blood-specific sex differences, but it has captured the more general sex-associated differences across human tissues and cell types. However, we have also found some tissues such as placenta (Figure 2.4h) showing an ambiguous boundary between the two sexes. The placenta is a fetal-maternal endocrine organ responsible for ensuring proper fetal development throughout pregnancy [138]. The fetal part of the placenta has the same genetic composition as the fetus, whereas it exhibits apparent different DNA methylation patterns. Our results demonstrate placenta samples are less distinguishable between the two sex groups, showing both ChrX in female placentas and ChrY in male placentas are less methylated than that in other normal tissues. During the early development of the human embryo, sperm cells are highly methylated and then become hypomethylated after fertilization [139]. Our results have shown that those sex-associated CpG on X chromosomes of sperm cells exhibited similar methylation patterns with other normal male tissues, however, the Y chromosomes are much less methylated. Collectively, our method can also be used to compare the methylation level of the two sex chromosomes in different tissues.

Our method can be readily applied to almost all DNA methylation datasets in GEO. Nearly half of the DNA methylation datasets uploaded to GEO are not in IDAT format, which is a prerequisite by using *minfi* and *ewastools*, many of these datasets only include intensity values of the methylated and unmethylated signals. Our sex classifier developed in this paper is based on beta values of those differently methylated CpG loci between the two sexes, users are only required to feed the whole beta value matrix, which can be easily computed from the signal intensity text files to the 'estimateSex' function in *wateRmelon* to obtain final sex predictions.

The underlying mechanism of our sex classifier is very intuitive: females have higher levels of methylation on ChrX, on the contrary, males are less methylated on ChrX and show strong methylation signals on ChrY. We have also demonstrated that the proposed classifier can be applied on both 450K and EPIC arrays. Compared to signal density-based methods such as *minfi* and *ewastools*, the methylation ratio-based method from our sex classifier and *sEst* provide better separation between the two sexes (Figure 2.3). In addition, both *minfi* and *ewastools* require at least one female and one male in the input samples to make correct sex predictions, however, our method and *sEst* do not have such limitation. Lastly, our method has a much higher advantage over *sEst* on running speed and this is especially the case when applied to a large sample size, for example, our method is more than four times faster than *sEst* when the number of input samples exceeds 1,000. Our speed advantage lies in that we saved the pre-trained weights for the sex-associated CpGs and only matrix multiplication is required to make sex classification, however, *sEst* requires performing two separate PCAs which are very time-consuming.

We have provided a powerful tool that can identify sex chromosome aneuploidies (45,XO and 47,XXY) from DNA methylation data. This function has been verified in five Turner syndrome samples and one Klinefelter syndrome case, we should acknowledge that we need many more positive cases to testify its sensitivity and specificity. It is a pity that we did not find any DNA methylation samples labelled as Klinefelter syndrome in the public repositories. Nevertheless, we found eight cases in the GEO database with great potential to be 47,XXY by applying our classifier, with the knowledge that most patients with Klinefelter syndrome have only mild symptoms and are never diagnosed. Those eight suspect Klinefelter syndrome cases can be good candidates to study the various developmental symptoms caused by copy number aberrations of sex chromosomes.

2.6 Conclusion

In this chapter, we constructed a very biologically intuitive sex classifier, simply based on the most robust CpG sites on the sex chromosomes, which not only can be used for sex predictions but also applied to identify samples with sex chromosome aneuploidy. Our classifier has been integrated into the *wateRmelon* Bioconductor package, which is freely and easily accessible by calling the 'estimateSex' function.

After constructing the new sex estimation tool for DNA methylation data and thoroughly demonstrating its accuracy and robustness. I then started to preprocess the DNA methylation datasets that were collected from public repositories. Firstly, the sex of sex-unknown samples were annotated as their estimated sex and these samples which have mismatches between their reported and estimated sex are discarded. Subsequently, the DNA methylation samples need to be normalized to reduce technical variations, the next chapter will explain why I need to create a new normalisation method and how it is created.

Chapter3

InterpolatedXY: a two-step strategy to normalise DNA methylation microarray data avoiding sex bias

The work presented in this chapter has been published in *Bioinformatics* [140].

Statement of Contribution: Yucheng Wang, the author of the thesis, originally conceived and developed the method. Yucheng Wang wrote the codes and performed all the analyses. Yucheng Wang wrote the manuscript. Xiaojun Zhai, Klaus D. McDonald-Maier and Leonard C. Schalkwyk advised and oversaw the work. Xiaojun Zhai, Klaus D. McDonald-Maier, Leonard C. Schalkwyk, Tyler J. Gorrie-Stone, Olivia A. Grant and Alexandria D Andrayas provided insights into writing the manuscript and interpreting the results.

3.1 Introduction

DNA methylation microarrays, such as Infinium HumanMethylation450 BeadChip [20] and Infinium MethylationEPIC BeadChip [30], provide cost-effective and high-throughput measurements of the methylation status over half a million CpG sites across the genome will continue to be the first choice by most DNA methylation related large cohort studies in the near future. Although whole genome bisulfite sequencing (WGBS) is recognized as the gold standard to measure the methylation patterns across the human genome, the high costs and technical complexity still pose significant challenges that prevent application to large-scale samples [29]. Data normalisation is an important prerequisite step to reduce unwanted technical variation. Currently, several normalisation methods are available for DNA methylation microarray samples. Among them, peak-based correction (PBC) [141], Beta MIxture Quantile normalization (BMIQ) [142] and noob [143] are all within-array normalization methods however they do not reduce between-array variation. By contrast, dasen [2] and funnorm [1] are the two most widely used between-array normalisation methods, which were reported to be able to effectively reduce the variation between samples. Dasen in the wateRmelon package utilises quantile normalisation to normalise methylated and unmethylated intensities separately, and also addresses the two types of probes, i.e. Infinium I and Infinium II probes, separately. Prior to the normalisation steps, there are linear regression procedures in dasen to reduce the density distribution difference between Type I and Type II probes [2]. The functional normalisation employed by funnorm is also an extension to quantile normalisation that removes variation explained by a set of selected covariates. In funnorm, the covariates are set as the first two principal components of the control probes, and linear regression is used to determine the proportion of variation explained by the covariates [1].

Females have two copies of the X chromosome, while males have one X chromosome and

one Y chromosome. To compensate for the different dosages of the X chromosome genes, one X chromosome in female cells is randomly subjected to inactivation in each cell lineage, with most parts of the inactive X being highly methylated [134, 135, 144]. As a result of this, the mean methylation values of the X chromosomes between sexes are very different [105, 99, 96]. The distinct methylation patterns of sex chromosomes between females and males raise a great challenge to unbiasedly normalising sex chromosome data. The existing between-array normalisation methods do not provide good solutions for normalising sex chromosome data. For example, dasen ignores this issue and normalises autosomes and sex chromosomes together, while funnorm is designed to normalise male samples and female samples separately for X chromosomes and Y chromosomes. Some DNA methylation related studies simply remove those probes mapped to the X and Y chromosomes prior to the normalisation step and do not include them in the downstream analysis. All these strategies come with their own drawbacks, either through losing some potentially interesting and biologically relevant signals from sex chromosomes or by introducing systematic technical differences between sexes.

Here we first demonstrate that the existing normalisation methods used to handle probes mapped on the X and Y chromosomes lead to introducing artificial sex bias into the normalised data. Then, we present a novel two-step strategy, which is designed to unbiasedly normalise both autosome data and sex chromosome data, is applicable to all quantile-based normalisation methods.

3.2 Materials and methods

3.2.1 Datasets

Two main datasets (Table 3.1) were used in this study. The first dataset includes 1195 individuals from the Understanding Society: UK Household Longitudinal Survey (UKHLS). Details about this UKHLS dataset are described by Gorrie-Stone et al. [145]. In brief, DNA methylation levels in whole blood within 489 male and 686 female healthy individuals were measured by EPIC array. The UKHLS dataset is available under request from the European Genome-phenome Archive under accession EGAS00001002836 (https://www.ebi.ac.uk). Since funnorm was developed and tested on 450k array samples, in this study we produce subsets from GSE142512 [146] to evaluate funnorm. GSE142512 includes 87 individuals with type 1 diabetes (T1D) and 87 individuals without T1D. The peripheral blood samples were collected from the subjects between 1 and 5 time points, with DNA methylation levels measured by either 450K or EPIC array, further details were documented by Johnson et al. [146]. We randomly selected 16 450k samples (12 males and 4 females) from GSE142512 as dataset one which is used to evaluate the performance of funnorm on small size dataset, and randomly selected 48 450k samples (23 males and 25 females) as dataset two to test funnorm's performance on relatively larger size dataset. For reproducibility, the sample IDs in the two subset datasets are listed in Table 3.2. GSE142512 is publicly available from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/).

Table 3.1: Characteristics of the datasets used in this study.

Name	Array type	Samples (female/male)	Age range (years)	Source
Dataset one	450k	16 (4/12)	0.8–13.6	GSE142512
Dataset two	450k	48 (25/23)	0.8–14.1	GSE142512
UKHLS	EPIC	1195 (686/489)	28–98	UKHLS

Table 3.2: Lists of sample ID used in dataset one and dataset two.

Sample	Sex	Dataset one (n=16)	Dataset two (n=48)
GSM4230892	Female	TRUE	TRUE
GSM4230891	Male	TRUE	TRUE
GSM4230890	Male	TRUE	TRUE
GSM4230889	Female	TRUE	TRUE
GSM4230888	Female	TRUE	TRUE
GSM4230887	Male	TRUE	TRUE
GSM4230886	Male	TRUE	TRUE
GSM4230885	Female	TRUE	TRUE
GSM4230884	Male	TRUE	TRUE
GSM4230883	Male	TRUE	TRUE
GSM4230882	Male	TRUE	TRUE
GSM4230881	Male	TRUE	TRUE
GSM4230880	Male	TRUE	TRUE
GSM4230879	Male	TRUE	TRUE
GSM4230878	Male	TRUE	TRUE
GSM4230877	Male	TRUE	TRUE
GSM4230876	Male	FALSE	TRUE
GSM4230875	Male	FALSE	TRUE
GSM4230874	Female	FALSE	TRUE
GSM4230873	Female	FALSE	TRUE
GSM4230872	Female	FALSE	TRUE
GSM4230871	Female	FALSE	TRUE
GSM4230870	Female	FALSE	TRUE
GSM4230869	Female	FALSE	TRUE
GSM4230868	Female	FALSE	TRUE
GSM4230867	Female	FALSE	TRUE
GSM4230866	Female	FALSE	TRUE
GSM4230865	Female	FALSE	TRUE
GSM4230864	Female	FALSE	TRUE
GSM4230863	Female	FALSE	TRUE
GSM4230862	Female	FALSE	TRUE
GSM4230861	Female	FALSE	TRUE
GSM4230860	Female	FALSE	TRUE
GSM4230859	Female	FALSE	TRUE
GSM4230858	Female	FALSE	TRUE
GSM4230857	Female	FALSE	TRUE
GSM4230856	Female	FALSE	TRUE
GSM4230855	Female	FALSE	TRUE
GSM4230854	Female	FALSE	TRUE
GSM4230853	Male	FALSE	TRUE
GSM4230852	Male	FALSE	TRUE
GSM4230851	Male	FALSE	TRUE
GSM4230850	Male	FALSE	TRUE
GSM4230845	Male	FALSE	TRUE
GSM4230844	Male	FALSE	TRUE
GSM4230843	Male	FALSE	TRUE
GSM4230842	Male	FALSE	TRUE
GSM4230841	Male	FALSE	TRUE

3.2.2 DNA methylation data processing

The DNA methylation raw data (IDAT files) were read into R by either using *iadd2* function in bigmelon or *read.metharray.exp* function in minfi. The methylation level of any given CpG locus is measured by its beta value which is defined as $\beta = (M) / (M + U + 100)$, where M is methylated intensity and U is unmethylated intensity for a given CpG locus. Basic quality control steps were performed to identify outliers, as recommended by Gorrie-Stone et al. [145]. Further, the reported sexes of samples were checked against the predicted sexes from DNA methylation data by using the *estimateSex* function in watermelon package [2], which predicts sex by comparing the methylation levels on sex chromosomes [99]. The original dasen normalisation is performed by calling the *dasen* function with default settings in the watermelon package, the original funnorm normalisation is performed by calling the *preprocessFunnorm* with default settings in the minfi package [147], which is actually applies *noob* method [143] as a first step for background correction and then perform the functional normalisation.

All analyses were performed using R 3.6.0 under Linux environment.

3.2.3 A two-step strategy to unbiasedly normalise DNA methylation samples

The framework of the interpolatedXY strategy is illustrated in Figure 3.1. The explicit procedures of the proposed new strategy to unbiasedly normalise both autosomal CpGs and sex chromosome-linked CpGs are as follows:

1. Step one: normalise the autosomal CpGs by one of the conventional normalisation methods, such as funnorm or dasen. It should be noted, that the probes mapped to sex chromosomes should not be included in this step to avoid potential influence.

2. Step two: infer the corrected values of sex chromosome-linked CpGs by looking for their nearest neighbours on autosomes, this is achieved by linear interpolation, here is the very efficient implementation:

   (a) Sort the corrected values of autosomal CpGs and build a function F which reflects correspondence of the rank of a CpG to its corrected value: $Corrected\_value_i = F(rank_i)$.

   (b) Sort and get the ranks of autosomal CpGs based on their raw values.

   (c) Estimate the ranks of sex chromosome-linked CpGs by linear interpolation on the rank distribution from procedure b.

   (d) Put the inferred ranks of sex chromosome-linked CpGs into the function F to get their final corrected values.

The above steps are ideally performed on raw signal intensities (M and U) and on each probe type (IGrn, IRed and II in funnorm, I and II in dasen) individually. After that, the normalised intensities can be converted into beta values as: $\beta = (M) / (M + U + 100)$. We name this strategy interpolatedXY. When dasen is used to normalise autosomal CpGs in the first step, we call this new normalisation method as "interpolatedXY adjusted dasen", the codes for this function are listed in Appendix codes A1. Similarly, "interpolatedXY adjusted funnorm" refers to another new normalisation method in which functional normalisation is applied in the first step, codes are listed in Appendix codes A2.

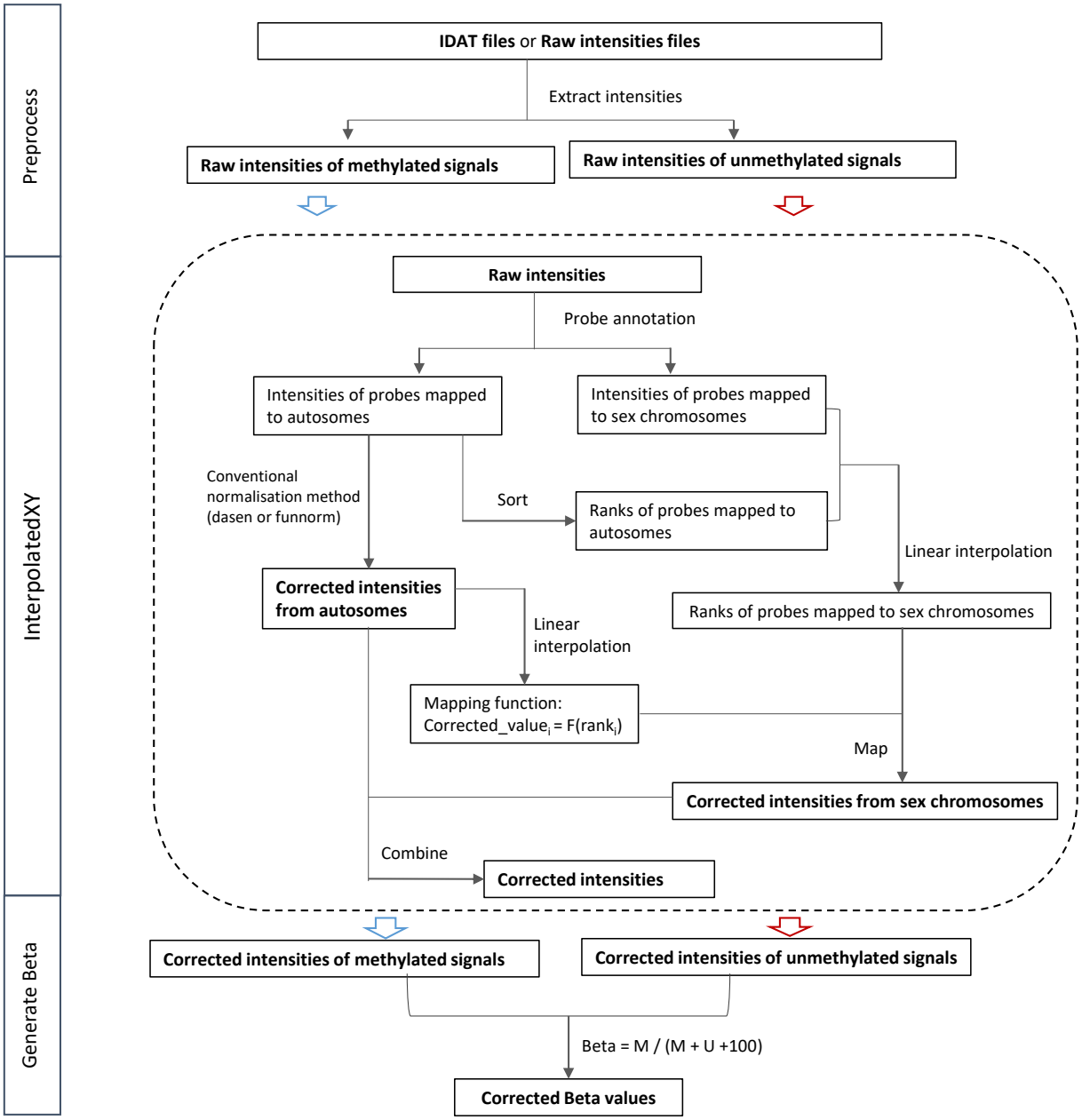

Figure 3.1: Overview of the interpolatedXY framework. Raw intensities are extracted from IDAT files or intensity text files, and then, the raw intensities of methylated and unmethylated signals are processed separately by the interpolatedXY procedure. Above all, chromosome annotation is performed on all probes to separate the raw input intensities into autosome-linked signals and sex chromosome-linked signals. These autosomes linked intensity signals are then normalised by a conventional normalisation method, such as dasen or funnorm. These sex chromosomes linked intensity signals are corrected as approximations of their nearest neighbours on autosomes, this is achieved by: 1) obtaining their approximate rankings by linear interpolation on the raw intensity distribution of autosomes mapped probes; 2) constructing a mapping function which deduces the corrected intensity value from its intensity rank by linear interpolation on the corrected intensities of autosome mapped probes. Finally, the corrected beta values are deduced from the corrected intensities signals

3.2.4 Performance evaluation for the interpolation approach

The proposed new approach infers the corrected values of sex chromosome-linked CpGs by linear interpolation on autosomal CpGs. To investigate whether the inferred data is accurate and reliable, we need a gold standard to evaluate the estimation accuracy. Females and males have very different methylation patterns on sex chromosomes, that is the main reason that we avoid normalising female samples and male samples together, with autosomes and sex chromosomes treated indiscriminately. However, when the targeted dataset includes only unisexual samples (only females or only males), then the sex chromosomes should be normalised together with other autosomes.

Inspired by this, we designed single sex groups: one that includes only female samples and the second that consists of only male samples. Firstly, the two groups are both normalised by conventional methods (e.g. dasen and funnorm) with the sex chromosomes being treated as general autosomes, thus the corrected values of those sex chromosome-linked CpGs could serve as the golden references (i.e. expected values). Secondly, based on our proposed interpolation approach, we infer the corrected values of sex chromosome-linked CpGs by interpolating on the normalised values of the autosomal CpGs. Lastly, the interpolated values are compared with their corresponding reference values. Root mean squared error (RMSE), which is sensitive to outliers, is used here to measure the deviations from the inferred values to their expected values:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\beta_i - \widehat{\beta_i}\right)^2} \tag{3.1}$$

where β_i is the methylation beta value of the i^{th} CpG, $\widehat{\beta_i}$ represents the expected methylation beta value of the i^{th} CpG, m represents the total number of CpGs studied.

3.2.5 Evaluation of the technical sex biases

The original dasen performs quantile normalisation with autosomal CpGs and sex chromosome CpGs processed together even when the dataset to be normalised is composed of both females and males. To investigate whether such an approach would introduce artificial sex biases, we compared the normalisation results of the UKHLS dataset generated by the original dasen and the interpolatedXY adjusted dasen.

The human methylome is not constant but responsive to many internal and external factors, such as genetic backgrounds and environmental factors [148]. As a result, the overall variance of the measured methylation values across all the CpG sites in the studied population can be described as:

$$V_{total} = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\beta_{ij} - \overline{\beta_j}\right)^2 \tag{3.2}$$

Where V_{total} represents the total variance of the studied samples, n is the total number of all samples, m is the total number of studied CpGs, β_{ij} represents the methylation beta value of the j^{th} CpG in the i^{th} sample, $\overline{\beta_j}$ represents the mean methylation beta value of the j^{th} CpG across all samples. Theoretically, we can then split the overall variance into the following two parts:

$$V_{total} = V_{biological} + V_{technical} = \frac{1}{n}\sum_{i=1}^{n}(V_i) \tag{3.3}$$

The first part $V_{biological}$ represents variance caused by meaningful biological reasons, such as cell types, age, gender, health status and other reasonable factors. The second part $V_{technical}$ represents variance resulting from technical issues, such as batch effect, random fluctuation and other unknown issues.

$$V_{biological} = V_{cell_type} + V_{age} + V_{sex} + V_{others} \tag{3.4}$$

$$V_{technical} = V_{batch} + V_{random} + V_{unknown} \tag{3.5}$$

Sex is one of the major biological factors which influences the methylation status of many autosomal CpGs, as a result, hundreds of autosomal CpGs have been reported showing significant different methylation levels between sexes [105, 102, 96]. The fraction of variances which are explained by sex can be deduced as follows:

$$\begin{aligned} F_{sex} &= \frac{V_{sex}}{V_{total}} \\ &= 1 - \frac{n_{females} V_{total\_in\_females} + n_{males} V_{total\_in\_males}}{(n_{females} + n_{males})\, V_{total}} \end{aligned} \tag{3.6}$$

Ideally, a good normalisation method should be able to not only greatly reduce the variances that are resulted from technical issues ($V_{technical}$), but also need to keep variances which have meaningful biological reasons ($V_{biological}$). This means, after the normalisation process, the overall variance should be reduced significantly while the sex explained fraction of variance should be increased. In this paper, to study the potential sex bias introduced by the mix normalisation method dasen, we compared the mean variance and the fraction of sex explained variances of the methylation values of CpGs after no normalisation (raw beta values), dasen normalisation and interpolatedXY adjusted dasen normalisation within the three chromosome groups (i.e. autosomes, X chromosomes and Y chromosomes).

3.2.6 Artifactual sex differences

If the conventional mixed normalisation approaches do introduce systematic artificial sex biases into the autosomal CpGs, then some autosomal CpGs could be falsely sex-associated. Epigenome-wide association studies (EWAS) are commonly used to systematically assess

the association between DNA methylation levels at genetic loci across the genome and a phenotype of interest. In this study, we apply EWAS to identify sex-associated CpG sites and then compare the EWAS results resulting from different preprocess approaches.

To perform EWASs for sex, the *champ.dmp* function in champ package [149], which utilises linear regression and F-test to identify differentially methylated positions is applied in this study to identify sex-associated CpGs. After Bonferroni multiple comparison correction, those CpG sites with p-value less than 0.05 were selected as significantly sex-associate. For simplicity and better comparison, we do not include age, cell type proportions and other covariates within the EWASs.

3.2.7 Comparison of the funnorm and the interpolatedXY adjusted funnorm

Funnorm is reported to be suitable for normalising methylation data with substantial global differences. The main difference between the original funnorm and the proposed interpolatedXY adjusted funnorm is how to normalise the methylation values of sex chromosome-linked CpGs. The original funnorm is designed to normalise X chromosomes separately and differently with Y chromosomes, as well as processes female samples and male samples separately. In contrast, the interpolatedXY adjusted funnorm does not require prior sex annotations and process both genders equally, which generates the corrected values of sex chromosome-linked CpGs by interpolation on the normalised values of autosomal CpGs.

To compare the normalisation effects on sex chromosome data between the original funnorm and the adjusted funnorm, we studied both the density distributions and the variances of the methylation values of CpG sites after no normalisation (raw beta values), funnorm

normalisation and adjusted funnorm normalisation within three chromosome groups (i.e. autosomes, X chromosomes and Y chromosomes) in two 450k datasets. The first dataset (dataset one) includes 12 male samples and 4 female samples, while the second dataset (dataset two) contains 23 male samples and 25 female samples.

3.3 Results

3.3.1 Estimation using the interpolation approach

We first investigated the performance of the interpolation approach employed by the interpolatedXY adjusted funnorm method. The deviations from the inferred values by the interpolation approach to their corresponding reference values are measured by RMSE. As it can be seen from Figure 3.2, the resulting RMSEs are all very small, especially for those in both X chromosomes and male Y chromosomes: the mean RMSE of X chromosome-linked CpGs is 1.15e-05 (sd=8.7e-06) in females and is 1.11e-05 (sd=4.8e-06) in male samples, while the mean RMSE of estimations for male Y chromosomes is 6.61e-06 (sd=3.2e-06). Though the RMSEs of Y chromosome-linked CpGs in females are slightly higher (mean=8.98e-04, sd=6.0e-04), they are still very subtle. With the knowledge that females do not carry Y chromosomes, and those observed signal intensities result from background noises and non-specific hybridization, there is no need to look much into the methylation values of female Y chromosomes. In the same way, we could observe similar performances of the interpolation approach employed by the interpolatedXY adjusted dasen method (Figure 3.3).

In summary, the above results demonstrate the proposed interpolation approach provides accurate and robust estimations for the corrected values of sex chromosome-linked CpGs.

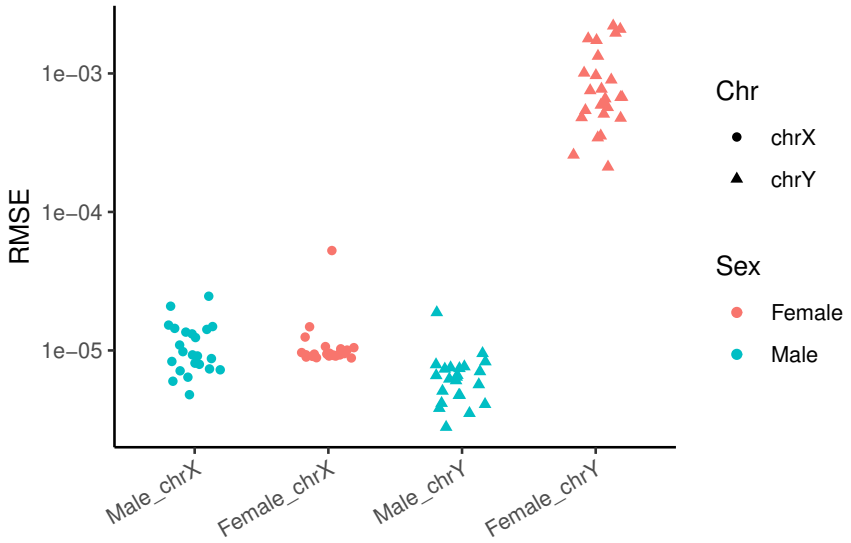

Figure 3.2: Difference between interpolated values and expected values within the adjusted funnorm. RMSEs are grouped into four categories: male X chromosomes, female X chromosomes, male Y chromosomes and female Y chromosomes. Female samples are in red colour and male samples are in blue colour. Dots represent X chromosomes, while triangles represent Y chromosomes.

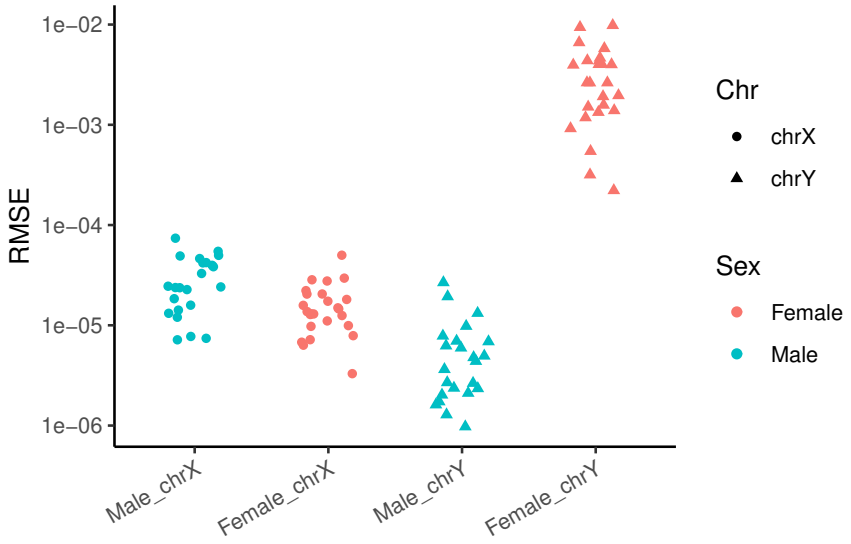

Figure 3.3: Difference between interpolated values and expected values within the adjusted dasen. RMSEs are grouped into four categories: male X chromosomes, female X chromosomes, male Y chromosomes and female Y chromosomes. Female samples are in red colour and male samples are in blue colour. Dots represent X chromosomes, while triangles represent Y chromosomes.

Table 3.3: The fraction of variance explained by sex in the UKHLS dataset with no normalisation (raw), dasen normalisation, interpolatedXY adjusted dasen normalisation and interpolatedXY adjusted funnorm normalisation.

Fraction of variance explained by sex (%)	Raw	Dasen	Adjusted dasen	Adjusted funnorm
Autosomes	0.34	**0.57**	0.45	**0.46**
X chromosome	73.18	77.24	**77.57**	76.93
Y chromosome	85.34	87.64	87.50	**88.82**

3.3.2 Artificial sex biases are introduced into autosomal CpGs by the conventional mixed normalisation method

The first round of the UKHLS dataset [145] includes 1175 whole blood samples whose DNA methylation levels were measured using the EPIC array. After quality control, 685 female samples and 486 male samples were kept for this analysis. To study the normalisation effects, the variance of beta values with three different pre-processing methods (no-normalisation, dasen and interpolatedXY adjusted dasen) are compared within three different chromosome groups (i.e. autosomes, X chromosomes and Y chromosomes) separately. As shown in Figure 3.5, both dasen and adjusted dasen significantly (Wilcoxon signed-rank test, p-value less than 2.2e-16) reduce the variance in all three chromosome groups. For instance, the mean variance of autosomes in both sexes decreased from around 0.0025 in non-normalised beta values to about 0.0018 after either dasen or adjusted dasen normalisation. The beta values density plots also demonstrate that both dasen and adjusted dasen greatly reduce the distribution variation (Figure 3.4). However, the difference in normalisation effects between dasen and adjusted dasen is not significant from the variance level.

Table 3.3 describes the sex explained fraction of variance between three methods in three chromosome categories. We can see that the sex explained variance in sex chromosomes by the three methods all exceeds 70%, while it accounts to only around 0.5% in autosomes. That is in line with our expectation, as sex is a dominant factor causing difference in methy-

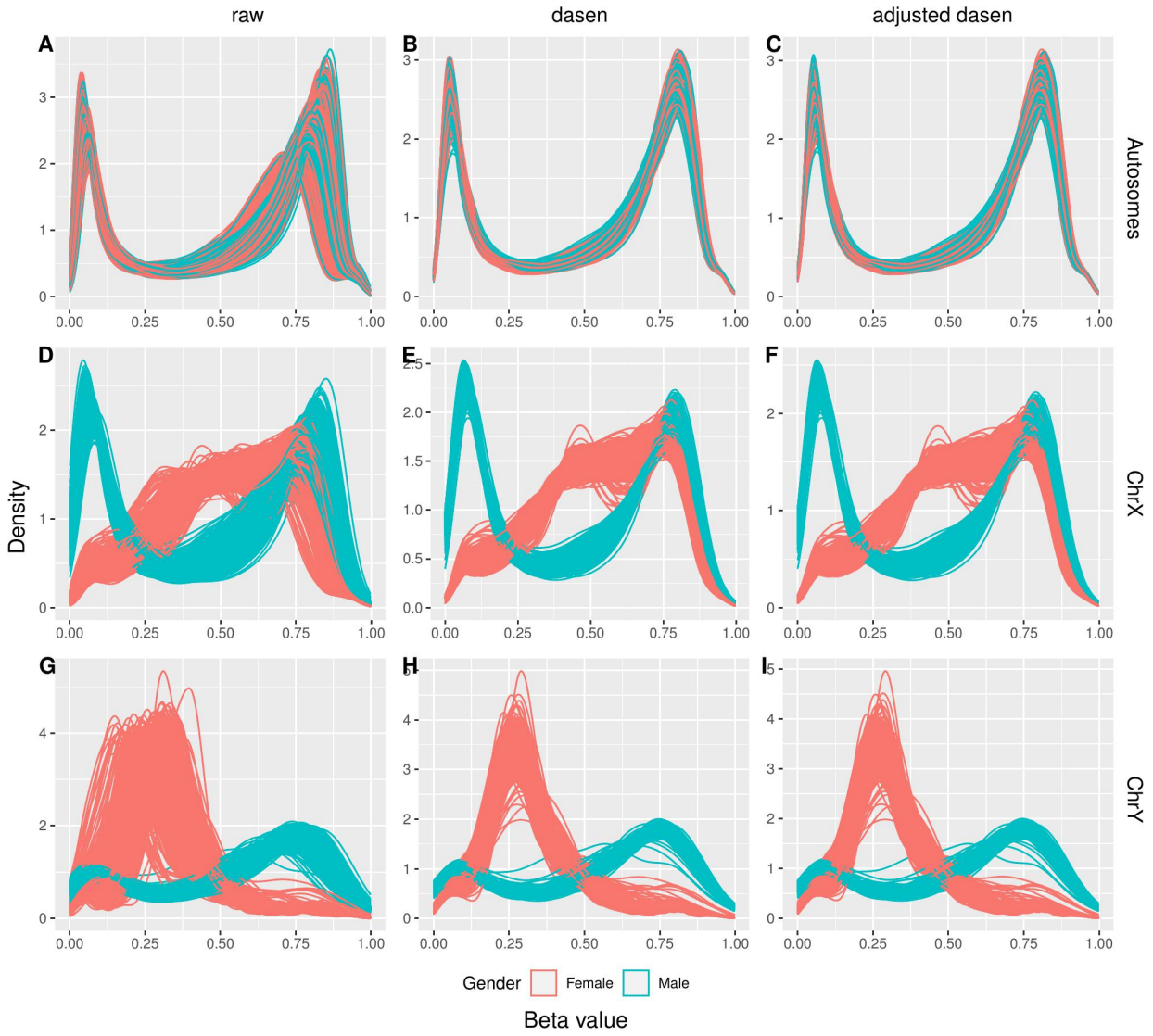

Figure 3.4: Comparisons in methylation beta value density distributions for UKHLS dataset. The three columns illustrate results from raw data (left column), funnorm normalised data (middle column) and the adjusted funnorm normalised data (right column). The three rows show density distributions of autosomal CpGs (first row), X chromosome linked CpGs (second row) and Y chromosome linked CpGs (third row). Red lines represent females and blue lines represent males.

lation levels of sex chromosomes, while the majority of autosomal CpGs are not influenced by sex. Interestingly, the sex explained fraction of variance of raw beta values in autosomes is 0.34%, it rises to 0.45% after normalising by the adjusted dasen, indicating the adjusted dasen method retained the meaningful biological difference when reducing technical variances (Figure 3.5A). However, the sex explained variance is much higher (0.57%) by normalising with the original dasen, can we conclude that the original dasen is better than the adjusted dasen to retain meaningful biological difference? On the contrary, these results indicate the original dasen has introduced artificial sex bias into to the normalised data. Combining the facts that only autosomal CpGs were included to compute the variance, and the difference in normalising the autosomal CpGs between the two methods is that the correction of autosomal CpGs is affected by the enrolling of sex chromosome data within the original dasen procedures, but not influenced within the adjusted dasen method. We can conclude that the observed higher fraction (sex explained fraction of variance in autosomes) with the original dasen normalisation is partly driven by the involvement of sex chromosome data, and this higher figure (i.e. than the adjusted dasen) indicates that technical sex biases have been introduced into to autosomal CpGs by the original dasen.

3.3.3 Confirmation of the introduced sex biases

We performed EWASs of sex based on autosomal beta values of UKHLS samples with three different pre-processing: no normalisation, dasen normalisation and interpolatedXY adjusted dasen normalisation. The identified number of sex significant (Bonferroni p-value less than 0.05) differentially methylated positions (saDMPs) are shown in Figure 3.6.

As illustrated in the Venn diagram (Figure 3.6A), there are 10,778 CpG sites been identified as saDMPs in the raw data, with 96.7% of them (10,427) also been captured after ad-

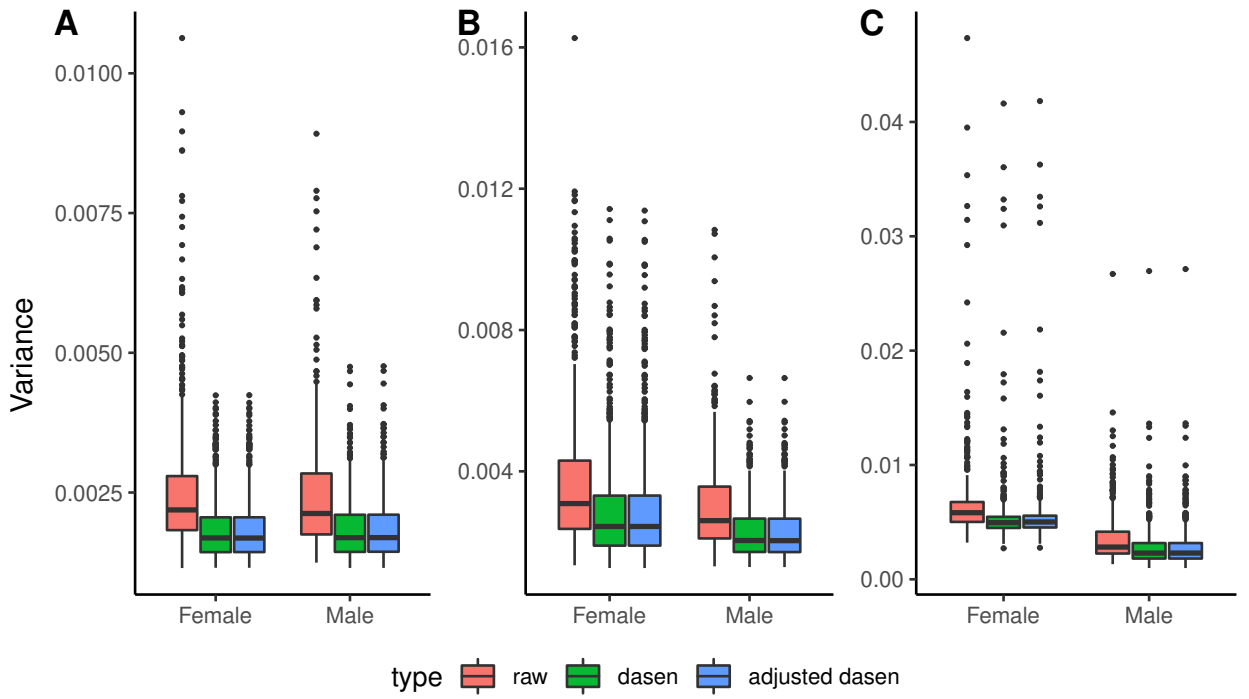

Figure 3.5: Variance comparisons in the UKHLS dataset. Boxplots comparing the variance of methylation beta values with three different pre-processing methods (i.e. no normalisation, dasen normalisation and adjusted dasen normalisation) in autosomes (A), X chromosomes (B) and Y chromosomes (C). Females and males are dealt with separately.

justed dasen normalisation. In addition, compared to raw data, the adjusted dasen approach enables the identification of another 4,201 saDMPs. Once again, these results demonstrate that while the adjusted dasen greatly reduces the variation of beta values (Figure 3.5A), it preserves the meaningful biological differences.

We found a total of 32,929 saDMPs after the original dasen normalization, which is more than three times the number with no normalisation or 2.25 times the number with adjusted dasen normalisation. Even so, 1,600 CpGs which are identified by both no normalisation and adjusted dasen normalization, are missed by the original dasen method. When comparing the dasen and adjusted dasen (Figure 3.6B), there are 12,021 saDMPs shared between the two methods. Interestingly, among the 20,908 dasen specific saDMPs, 96.0% of them (20,070) have higher methylation values in males than that in females. On the contrary, 2,318 out of

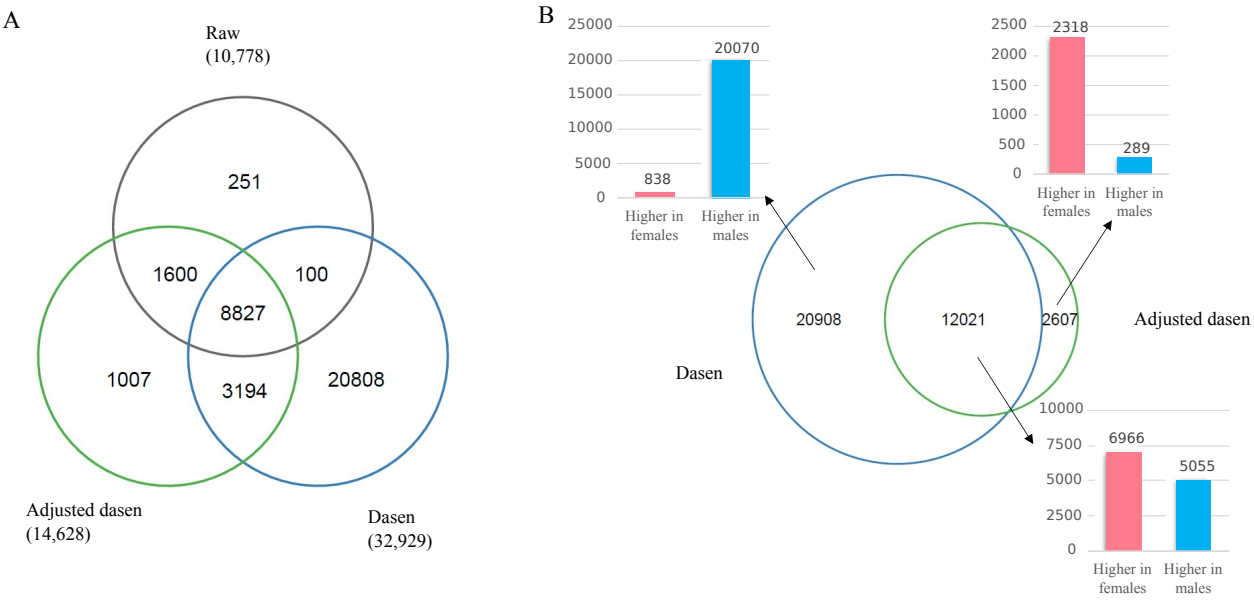

Figure 3.6: EWAS results of UKHLS dataset. A. The Venn diagram shows the number of unique and shared saDMPs between three approaches: no normalisation (raw), dasen normalisation and adjusted dasen normalisation. B. The Euler diagram describes the number of unique and shared saDMPs between dasen normalisation and adjusted dasen normalisation, with the three bar plots showing the number of CpGs which have higher methylation values in females (red) or males (blue) in three categories separately.

the 2,607 adjusted dasen specific saDMPs (88.9%) show higher methylation values in females than males. Again, with the fact that the interpolatedXY adjusted dasen only differs from the original dasen by not enrolling sex chromosome data when normalising the autosomal data, the above results suggest the original dasen did introduce artificial sex biases into autosomal CpGs by making the methylation values of many CpGs slightly higher in male samples and lower in female samples. This explains why nearly all the dasen specific saDMPs have higher methylation values in male samples, and there are more than two thousand CpG sites which have higher methylation values in female samples that were identified as significant saDMPs by the adjusted dasen approach but missed by the original dasen.

3.3.4 InterpolatedXY adjusted funnorm provides better normalisation results for sex chromosome-linked CpGs than the original funnorm

Since the original funnorm has two different designs to deal with datasets with different sizes, we compared the normalisation effects between the original funnorm and the interpolatedXY adjusted funnorm in two datasets. The adjusted funnorm does not differ from the original funnorm in normalising the autosomal CpGs, so the corrected values of autosome data from the two methods are the same, we can thus observe identical results for autosomal CpGs by the two methods (Figure 3.7B and 3.7C, Table 3.4, Figure 3.8B and Figure3.8C, Table 3.5).

Table 3.4: The fraction of variance explained by sex in dataset one (n=16) with no normalization (raw), funnorm normalization and interpolatedXY adjusted funnorm normalization.

Fraction of variance explained by sex (%)	Raw	Funnorm	Adjusted funnorm
Autosomes	9.48	10.93	10.93
X chromosome	92.68	**82.82**	92.99
Y chromosome	91.48	**97.09**	93.89

For the X chromosome-linked CpGs, when applied to small datasets, whose number of female samples or male samples is less than ten, such as dataset one, funnorm is designed to normalise female X chromosomes and male X chromosomes together by functional normalisation. Compared to the non-normalised raw beta values, the density distributions of the corrected data generated by funnorm turn out to be much discordant in both female samples and male samples (e.g. Figure 3.7E). On the contrary, after the adjusted funnorm normalisation, the density distributions become more consistent in both sexes (Figure 3.7F). We can also observe the same trends from the bar plots in Figure 3.9B, the original funnorm greatly increases the variance in both sex groups, while the adjusted funnorm keeps the variance low. Furthermore, the sex explained fraction of variance was reduced to 82.8% by the original funnorm, which is 92.7% in raw data and 93.0% after the adjusted funnorm nor-

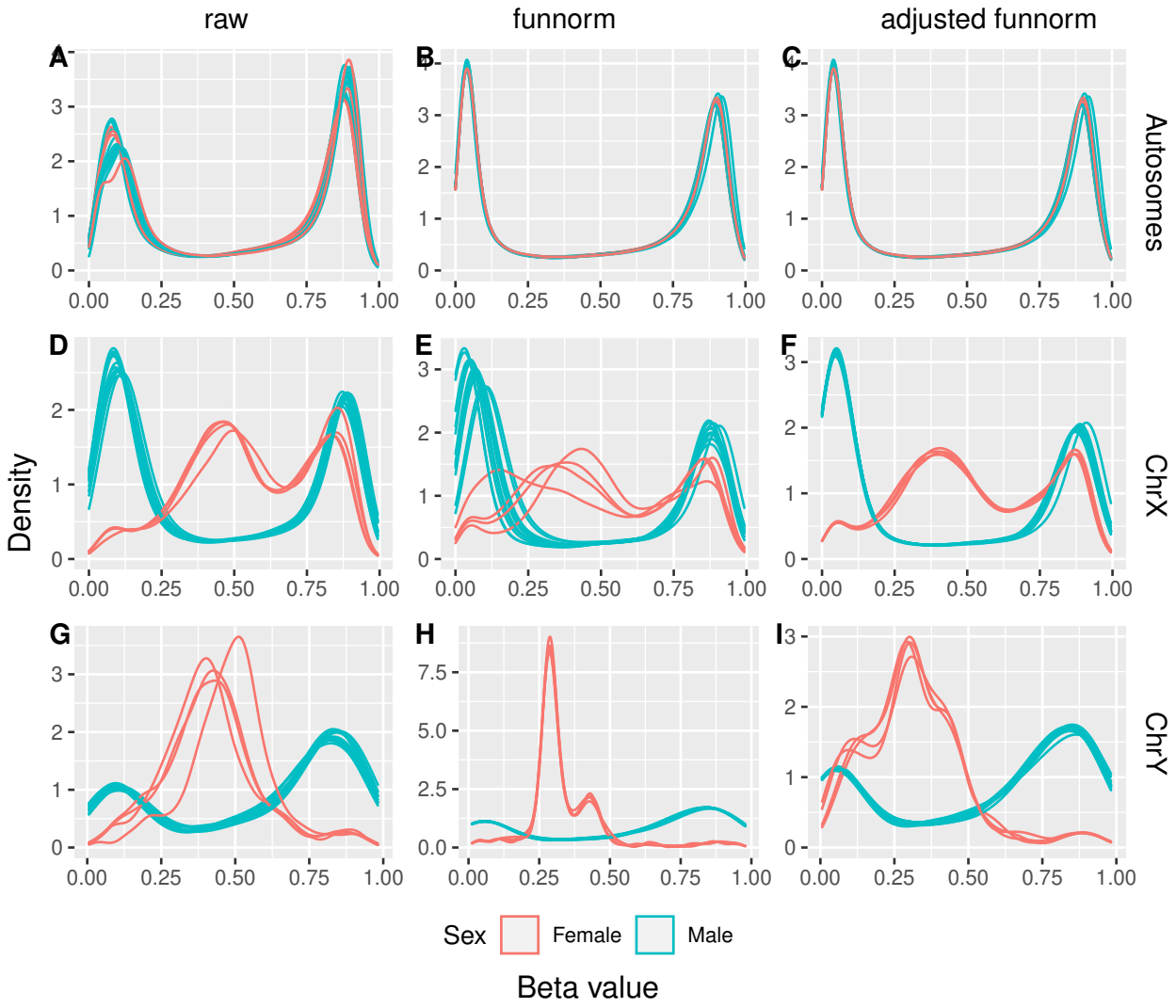

Figure 3.7: Comparisons in methylation beta value density distributions for dataset one. The three columns list results from raw data (left column), funnorm normalised data (middle column) and the adjusted funnorm normalised data (right column). The three rows show density distributions of autosomal CpGs (first row), X chromosome-linked CpGs (second row) and Y chromosome-linked CpGs (third row). Red lines represent females and blue lines represent males.

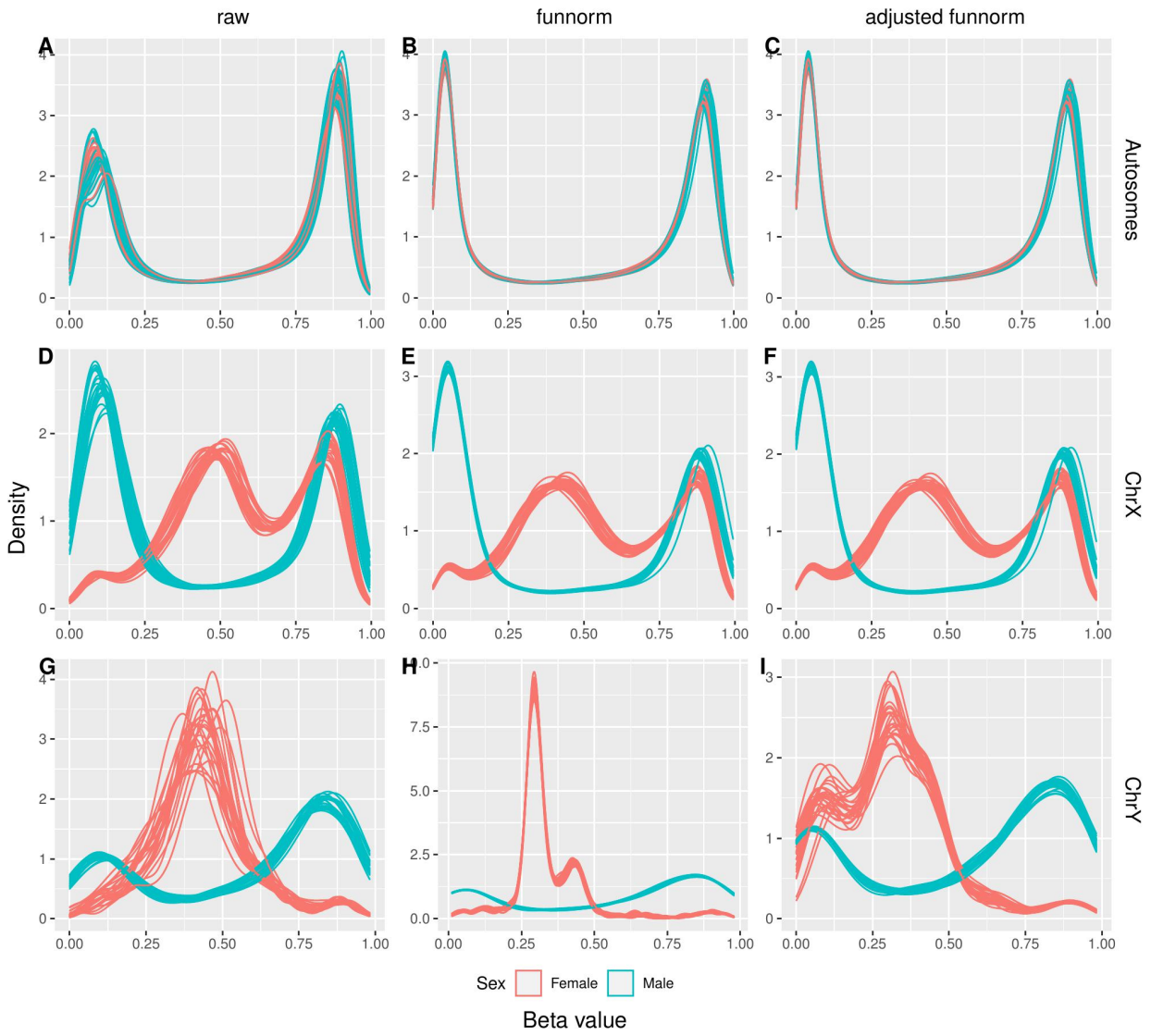

Figure 3.8: Comparisons in methylation beta value density distributions for dataset two. The three columns illustrate results from raw data (left column), normalised data (middle column) and the adjusted funnorm normalised data (right column). The three rows show density distributions of autosomal CpGs (first row), X chromosome linked CpGs (second row) and Y chromosome linked CpGs (third row). Red lines represent females and blue lines represent males.

Table 3.5: The fraction of variance explained by sex in the dataset two (n=48) with no normalisation, funnorm normalisation and interpolatedXY adjusted funnorm normalisation.

Fraction of variance explained by sex (%)	raw	funnorm	interpolatedXY adjusted funnorm
Autosomes	6.39	7.05	7.05
X chromosome	93.37	93.33	93.21
Y chromosome	88.45	97.75	89.12

malisation (Table 3.4). Taken together, the above results indicate that the original funnorm is actually adding technical variation into the methylation data of X chromosomes for those small sample size datasets.

When applied to larger datasets, such as in the case of dataset two, funnorm performs separate functional normalisations on female X chromosomes and male X chromosomes, with the underlying consideration that females and males have very different methylation patterns on X chromosomes. When comparing the normalisation effects between the original funnorm and the adjusted funnorm based on dataset two, we did not observe any significant differences in the methylation profiles of X chromosomes (Figure 3.8, Figure 3.10 and Table 3.5).

For the Y chromosome-linked CpGs, the original funnorm does not use functional normalisation as it does on other chromosomes, such as autosomes. Instead, only quantile normalisation is employed by the original funnorm to normalise the Y chromosome data, with female samples and male samples processed separately. This may explain why the sex explained variance within the original funnorm is much higher (i.e. 97.75%) than that in the raw data (i.e. 88.45%) and adjusted funnorm (i.e. 89.12%) (Table 3.5). We can also observe a similar trend from Table 3.4. These results suggest the separate normalisation strategy employed by the original funnorm will increase the difference between the two sex groups, and thus introduce artificial technical bias.

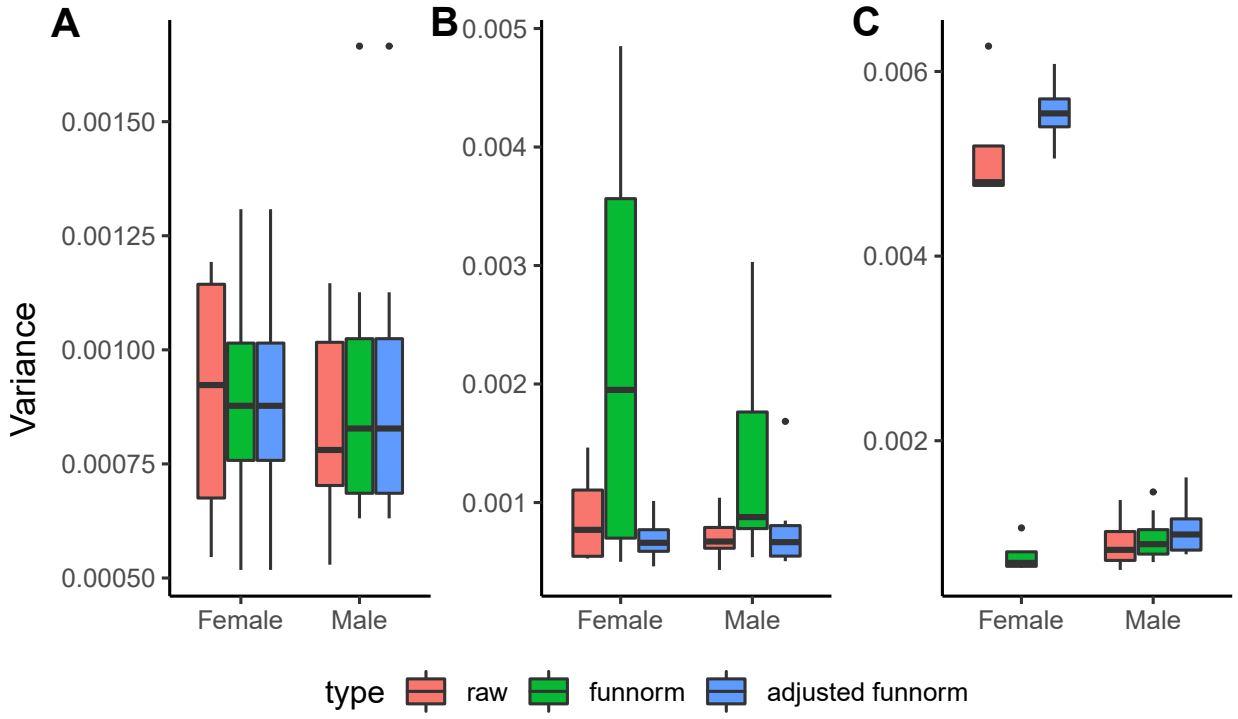

Figure 3.9: Variance comparisons in dataset one. Boxplots comparing the variance of methylation beta values with three different pre-processing methods (i.e. no normalisation, dasen normalisation and adjusted dasen normalisation) in autosomes (A), X chromosomes (B) and Y chromosomes (C). Females and males are dealt with separately.

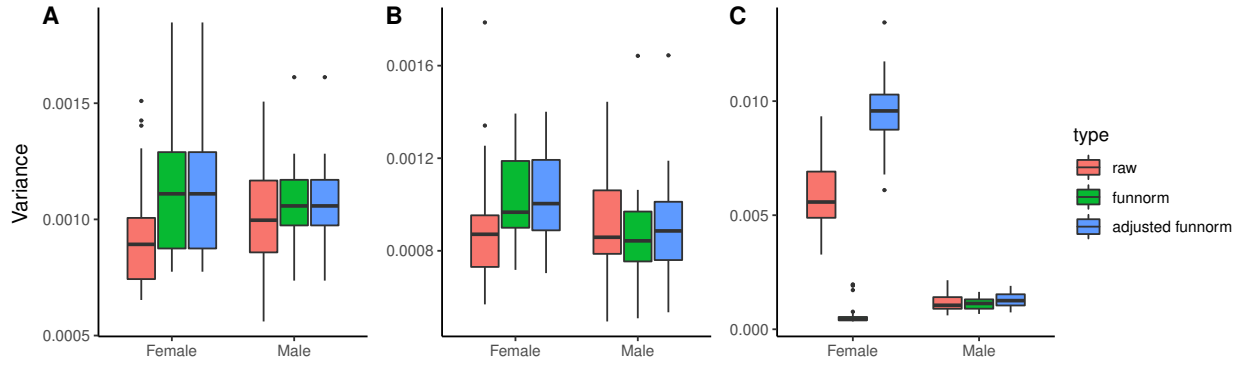

Figure 3.10: Variance comparisons in the dataset two. Boxplots comparing the variance of methylation beta values with three different pre-processing methods (i.e. no normalisation, dasen normalisation and adjusted dasen normalisation) in autosomes (A), X chromosomes (B) and Y chromosomes (C). Females and males are dealt with separately.

3.3.5 Comparison between the interpolatedXY adjusted funnorm and interpolatedXY adjusted dasen

We have demonstrated that the fraction of variance explained by sex is very useful to measure the normalisation effects for different methods and have also shown that the adjusted the dasen and the adjusted funnorm are both superior to their original versions. Then we compared their normalisation effects on a large healthy population: the UKHLS dataset (n=1171). The results are shown in Table 3.3, the first obvious observation is that both the adjusted dasen and the adjusted funnorm clearly increased the fraction of variance explained by sex in all chromosome groups (i.e. autosomes, X chromosome and Y chromosome) than the raw data, demonstrating that the use of either normalisation method is beneficial and worthwhile. As compared to the two adjusted normalisation methods, we can see their effects are comparable in the studied dataset (Table 3.3): the adjusted funnorm marginally outperforms the adjusted dasen in normalising the autosome data (0.46% vs. 0.45%) and Y chromosome data (88.82% vs. 87.5%), while the adjusted dasen is slightly better in normalising the X chromosome data (77.57% vs. 76.93%).

3.4 Discussion

We have described a two-step sex-unbiased data normalisation strategy for normalising DNA methylation microarray samples, which can be applied to almost all quantile-based normalisation methods, such as dasen and funnorm. By this strategy, the autosomal CpGs are normalised independently and separately from the sex chromosome CpGs, while the corrected values of sex chromosomes CpGs are estimated as the weighted average of the corrected methylation values of their nearest neighbour atusosomal CpGs.

The two steps are necessary. Since the average methylation levels of CpGs on X chromosome in females are very different from that in males, normalising them together with the autosomal CpGs, especially by the quantile-based methods, will introduce technical biases for both autosomes and sex chromosomes. By comparing the normalisation effects of the original dasen and the interpolatedXY adjusted dasen, we confirmed that the technical sex biases were introduced into the autosomal CpGs by the mix normalisation approach (original dasen)—with the sex explained fraction of variance in autosomes rising to 0.57% from 0.44% in the adjusted dasen normalised data. We further propose a rational explanation for this (Figure 3.11): within the quantile normalisation steps in dasen, there are procedures to sort and return ranks for all the probes, as the mean methylation values of the most X chromosome-linked CpGs in females are higher than nearly half of the autosomal CpGs, whereas the methylation values of the corresponding positions in males are relatively low, thus the quantile normalisation algorithm used to make all studied samples fit into the same distribution creating a systematic negative shift for many autosomal CpGs (their methylation values are lower than most X chromosome-linked CpGs) in females and a systematic positive shift for those CpGs in males. As a result of this, when we perform EWAS to look for autosomal sex-associated CpGs, the original dasen approach identified more than two times the number as identified by the adjusted dasen or non-normalised data. Moreover, 96.0% of the dasen specific saDMPs show higher methylation values in male samples than in female samples, by contrast, the majority of the 2,607 CpGs missed by the original dasen but identified by the adjusted dasen have higher methylation values in female samples than male samples.

Estimation of the corrected values for sex chromosomes CpGs by looking at their nearest neighbours on autosomes is made both possible and reliable by the fact that DNA methylation microarrays simultaneously measure over half a million CpG sites across the genome, and only a relatively small portion (i.e. 2.3% in EPIC and 2.4% in 450K) is mapped on the

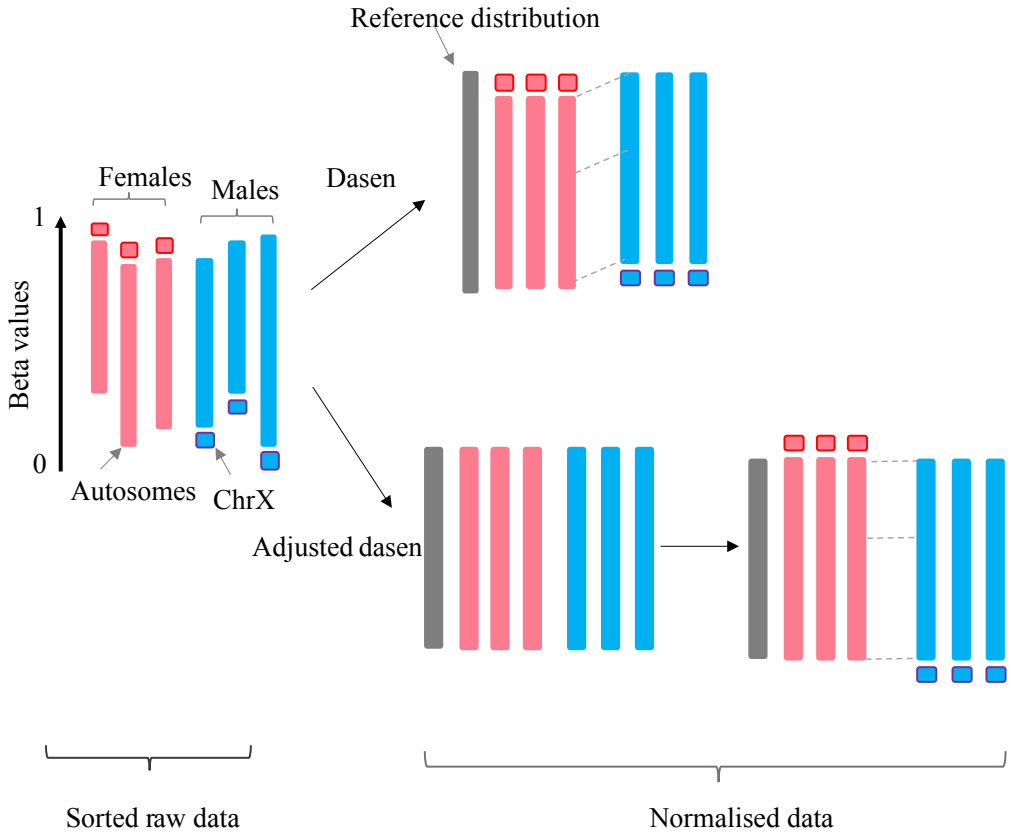

Figure 3.11: A simplified schematic diagram illustrates the difference in the normalisation process between the original dasen and the interpolatedXY adjusted dasen. The original dasen normalises autosomes and sex chromosomes together, the mean methylation values of most X chromosome-linked CpGs in females are higher than nearly half of the autosomal CpGs, whereas the values of the corresponding locus in males are relatively very low, thus the quantile normalisation algorithm employed by dasen to make all studied samples fit into a same distribution creating a systematic shift for many autosomal CpGs in two sexes. The adjusted dasen manages to avoid such an issue by doing quantile normalisation in autosomes separately and independently with sex chromosomes, and inferring the corrected values of sex chromosomes by interpolating on autosomes. Red denotes female sample and blue denotes male sample, the long bar represents sorted autosomal CpGs and the short bar represents sorted X chromosome-linked CpGs.

sex chromosomes. Here in this study, we have demonstrated that the linear interpolation approach provides both accurate and robust estimations for the sex chromosome data, with the mean RMSE less than 1.2e-5.

Funnorm is favoured for normalising methylation data with substantial global differences, such as cancer samples [1]. With the consideration that females and males have distinct methylation patterns for sex chromosomes, funnorm has very explicit rules to normalise X chromosomes and Y chromosomes differently. Within the functional normalisation in funnorm, there is a regression step to infer the explainable technical variants based on control probes. The authors may have considered the regression models would be less accurate in the circumstance of only a few samples, so funnorm is designed to perform functional normalisations on female X chromosomes and male X chromosomes together when the number of either female samples or male samples is less than ten. Our results in Section 3.3.4 have clearly shown that such a mixed normalisation approach is destructive to the methylation profiles of X chromosomes in both females and males. Though doing functional normalisation on females and males separately is a way to avoid such an issue, it may also introduce potential systematic technical bias between the two separate groups.

For the Y chromosome-linked CpGs, the original funnorm does not actually perform the functional normalisation as it does on other chromosomes, instead it performs only quantile normalisations on Y chromosomes, and processes female samples and male samples separately. As the proposed interpolatedXY adjusted funnorm could provide near-perfect estimations for corrected values generated by functional normalisation, it could be particularly useful for studies that focus on sex chromosomes DNA methylation data, especially when the methylation difference between the studied groups are known to be very different. Moreover, by the adjusted funnorm method, the corrected values of sex chromosome-linked CpGs are produced by linear interpolating on the distribution of autosomal CpGs, so in theory, they

are more comparable with the autosomal CpGs.

In this paper, we not only present a novel two-step strategy to unbiasedly normalise DNA methylation microarray samples, but also provide a useful concept—the fraction of variance explained by sex, to quantitively measure the normalisation effect. Sex is an important biological factor that not only determines the methylation status of sex chromosomes, but also influences many autosomal CpGs. A good candidate normalisation method should not only be able to greatly reduce the technical variation between samples, but also should preserve the meaningful variation that has biological reasons (e.g. sex). Even though quantile normalisation has been widely employed by several DNA methylation normalisation methods, such as SWAN [150], dasen [2] and funnorm [1]. There are still concerns about whether the use of between-array normalisation methods could bring enough benefits to counterbalance the potential impairment of data quality [151]. Here, in this study, we demonstrated that the interpolatedXY adjusted dasen and the interpolatedXY adjusted funnorm are two good normalisation method candidates, they are able to not only greatly reduce technical variation but also retain the meaningful biological difference, which will be very useful for large cohort EWAS projects.

We believe that the proposed novel two-step strategy may have wider application outside of DNA methylation microarrays and could even be applied in broader technologies such as RNA-Seq.

3.5 Conclusion

The proposed two-step strategy of interpolatedXY allows for the normalisation of autosomal data and sex chromosome data without bias. The two steps are necessary and reliable, the interpolatedXY approach infers the normalised methylation beta values of sex chromosome-linked CpGs with deviations, i.e. RMSE, in around 1.15e-05 to their expected values. With the integration of the interpolatedXY, the adjusted dasen and the adjusted funnorm both show superior performance than their original versions, i.e. the adjustedDasen avoids the risk of introducing sex bias into the autosomal data when normalising mixed-sex samples compared to the original dasen; the adjustedFunnorm reduces artificial sex bias in the sex chromosome data as compared to the original funnom. In addition, the sex explained variance analysis reveals the two between-array normalisation methods, dasen and funnorm, both enable retaining the meaningful biological difference while reducing technical variation.

For the DNA methylation samples I collected from public repositories, after sex annotation and sex checking, the remaining samples are then normalised with a fixed reference by the newly developed adjusted dasen method. Then, the clean and normalised DNA methylation samples are ready for downstream analysis, such as being used to build age clocks. Even whole blood is comprised of different cell types, however, do different cell types or tissues have different ageing rates, and if so, how to confidently measure them? The next chapter will present my finding on ageing rate comparisons across tissues by using DNA methylation clocks.

Chapter4

Insights into ageing rates comparison across tissues from recalibrating cerebellum DNA methylation clocks

The work presented in this chapter has been online in *bioRxiv* [152].

Statement of Contribution: Yucheng Wang (YW), the author of the thesis, conceived the study. YW collected the data and performed all analyses except for the enrichment analyses. Olivia A. Grant performed the enrichment analyses and summarised this part of work. Leonard C. Schalkwyk and Xiaojun Zhai contributed to the interpretation of the results. YW drafted the manuscript with critical contributions from Leonard C. Schalkwyk, Klaus D. McDonald-Maier, Olivia A. Grant and Xiaojun Zhai. Xiaojun Zhai, Klaus D. McDonald-Maier and Leonard C. Schalkwyk advised and oversaw the work.

4.1 Introduction

Ageing is generally considered a gradual process that happens to the body as a whole. It is still an open question whether different organs/tissues have different ageing rates. Furthermore, how can we truthfully compare the ageing rates between different tissues? Horvath's pan-tissue clock gives excellent accuracy in estimating DNAm age for many different cells and tissues [40], which may suggest that those different cells and tissue types may have similar ageing rates. In 2015, Horvath et al. claimed that the cerebellum ages slower than many other parts of the human body based on the observations that the DNAm age of the cerebellum is much lower than other tissues based on the pan-tissue clock [153]. In addition, Horvath and his colleagues also claimed that women's breast tissues have a relatively higher DNAm ageing rate [40, 154]. If it is true that some tissues have significantly different DNAm ageing rates than other tissues, then we can go further to identify what drives the difference. This is a very important angle to understand the mechanisms of age-associated DNA methylation changes. Even though there have been reported many strong age-associated CpG sites, there is still very little known about the underlying mechanisms that drive age-associated DNA methylation changes [155, 156, 157].

In recent years, many more cerebellum DNA methylation samples have become publicly available and many diverse DNAm age clocks have also been developed [86]. We set out to examine the claim that the cerebellum ages slowly within a much larger size dataset and find out the mechanisms. To achieve that, we first collect the largest cerebellum DNA methylation sample dataset, then compare their estimated epigenetic ages from six representative DNAm age clocks. After that, we perform age EWAS for cerebellums and middle temporal gyrus, separately on the same large-size elderly population (n = 404) to reveal the distinct age-associated methylomic changes of the cerebellum. Lastly, we construct cerebellum-specific

clocks and further examine the claim that the cerebellum ages slower.

4.2 Methods

4.2.1 DNAm datasets

The DNAm samples were collected from the public data repository—GEO. The cerebellum samples are from six datasets, including GSE134379 [158], GSE59685 [159], GSE105109 [160], GSE125895 [161], GSE61431 [162] and GSE72778 [75]. They were included according to the following criteria: contain at least 20 cerebellum samples; with age annotations, and raw IDAT files or methylated and unmethylated intensity files are available. The cerebellum samples were used to reveal the underestimation issues for the cerebellum tissue by six representative clocks and were also used to train cerebellum age clocks. Apart from cerebellum samples, GSE134379 [158] also includes DNAm microarray data of the middle temporal gyrus from the same 404 individuals, thus it was used to perform age EWASs on the two brain tissues. GSE59685 [159] includes 531 DNAm samples of five tissues, i.e. cerebellum, entorhinal cortex, frontal cortex, superior temporal gyrus and whole blood, from donors (N = 122) archived in the MRC London Brainbank for Neurodegenerative Disease. GSE59685 and GSE134379 were also used to compare the DNAm ages of different tissues which were estimated by our trained cerebellum clocks. The DNAm samples of the non-cerebellar brain tissues in four datasets, i.e. GSE134379 [158], GSE74193 [163], GSE80970 [164] and GSE61431 [162], were used to train the CerebralCortexClock$_{\text{common}}$ clock, the four datasets were selected to ensure a relatively equal sample distribution across all age groups in the adult population.

4.2.2 Data preprocessing

For all the DNAm datasets, after downloading from the GEO, they were read into R by using the *iadd2* function from the 'bigmelon' package [165] when raw IDAT files were available. For those datasets in which only text-formatted intensity files exist, the methylated and unmethylated intensities were extracted and read into R directly. Then the raw methylation beta values is calculated as: $\beta = \frac{M}{M+U+100}$, where M denotes methylated intensities and U denotes unmethylated intensities. For all those samples, we estimated their sex by using the *estimateSex* function [166] from the wateRmelon package [2], any samples with mismatches between its reported sex and the estimated sex from the DNAm data were excluded for downstream analysis. Also, the beta value density distributions of samples within each dataset were manually checked to remove any samples with abnormal distribution profiles.

4.2.3 DNA methylation age prediction

The DNAm age prediction of the six representative clocks, i.e. Hannum2013 [39], Horvath2013 [40], Horvath2018 [155], Levine2018 [81], Zhang2019 [78] and Shireby2020 [70], was completed by using the *methyAge* function from the 'dnaMethyAge' R package [167]. Only methylation beta values are required to feed into the *methyAge* function. Note, when calculating the DNAm age of Horvath2013, the raw beta values are firstly normalized with an adjusted BMIQ which has a fixed reference, this is consistent with Horvath's original publication [40]. To calculate the DNAm age of Zhang2019, the beta values of each sample are first subjected to Z-score normalization [78]. For the remaining clocks, no normalization steps were applied. The difference between DNAm age and chronological age is measured as:

$$RMSD = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{4.1}$$

$$MAD = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{4.2}$$

where y_i represents the chronological age of the i_{th} sample, $\hat{y}_i$ represents the predicted DNAm age of the i_{th} sample, m denotes the number of all samples. RMSD: root mean squared deviation; MAD: mean absolute deviation.

4.2.4 Epigenome-wide association study

The age EWASs were performed on GSE134379 [158] which includes DNAm microarray data of two brain tissues (CBL and MTG) in every individual from a large elderly population (N=404). The CBL samples and MTG samples were normalised by the *adjustedDasen* [140] from the 'wateRmelon' package [2] separately. These probes target CpGs mapped to sex chromosomes or reported to have cross-hybridizing issues and were removed from downstream analysis [168]. To find out age-associated differentially methylated CpGs across the genome in the two brain tissues, we fitted the following linear regression model for each CpG site involved in the two tissues separately:

$$\beta_i \sim w_{1i}*Age+w_{2i}*Sex+w_{3i}*Plate+w_{4i}*Beadchip+w_{5i}*Disease\_status+intercept \tag{4.3}$$

where β_i is the methylation beta value of the i_{th} CpG, w_{1i} is the coefficient of chronological age for the i_{th} CpG. Age, sex and disease status describe the biological difference between different subjects while plate and beadchip describe potential technical differences introduce

when measuring the methylation level. The *t statistic* of the coefficient w_{1i} is checked in the *Student's t* distribution to determine the p-value. After that, the p-values of all studied CpGs were adjusted with the Benjamini & Hochberg method. A CpG is called to be significant age-associated when its adjusted p-value (or FDR) is less than 0.01.

4.2.5 The construction of DNAm clocks

Prior to any training steps, all DNAm samples were normalized by a modified version of *adjustedDasen* [140] method from the 'wateRmelon' package [2], in which the modified *adjustedDasen* is supplied with a fixed reference to reduce the batch variance between different datasets. Also, the chronological age is log-transformed.

The new clocks mentioned in this study were all trained by the penalized linear regression algorithm—Elastic net [169], which is essentially a linear combination of the L1 and L2 penalties of the lasso regression and ridge regression, L1 regularisation penalises the total of the weights' absolute values, whereas L2 regularisation penalises the total of the weights' squares. The loss function of Elastic net is defined as:

$$\frac{1}{2}\sum w_i(y_i - \beta_i^T c - c_0)^2 + \frac{1}{2}\sum \lambda\gamma_j(1-\alpha)c^2 + \alpha|c| \tag{4.4}$$

where the β_i denotes the methylation beta value of ith CpG, c is the coefficient vector of all the CpG accounted, α is the critical parameter that controls the weights of the L1 and L2 penalties and has been defined prior to the training.

We used the *cv.glmnet* function from the 'glmnet' R package [170] to train the Elastic net models. To train the $\text{CerebellumClock}_{\text{specific}}$, the input samples are the 752 cerebellum sam-

ples from six independent datasets, the input CpG set of each sample was restricted to the 613 age-associated CpGs in the cerebellum, alpha was set to 0.5, and 10-fold cross-validation was used to determine the optimal coefficient combination. We made use of leave-one(dataset)-out cross-validation to infer the age prediction performance of the CerebellumClock$_{\text{specific}}$. Specifically, we have six independent cerebellum datasets, then for each round of the total six rounds of cross-validation process, one dataset was taken out and their DNAm ages were estimated by the model trained on the remaining five datasets, after six rounds, the DNAm ages of samples from the six datasets were derived and they were not overfitted by the training process. In the same way, the CerebellumClock$_{\text{common}}$ was trained on the same 752 cerebellum samples but the input CpG sets were restricted to the 201 shared age-associated CpGs. Another difference was the alpha value was set to 0.2 to let the final model includes more CpGs from the 201 CpGs.

The training of CerebralCortexClock$_{\text{common}}$ also employed Elastic net linear regression, the training samples were those of non-cerebellar brain tissues from four independent datasets, the input CpG set of each sample was also restricted to the 201 shared age-associated CpGs and the alpha was set to 0.2. As we only have four separate datasets, and only GSE74193 [163] has a wide age range, we employed a 10-fold cross validation to measure the age prediction performance of CerebralCortexClock$_{\text{common}}$. That is to say, we first randomly separate all the training samples into equal 10 portions, for each round of the total 10 rounds of cross-validation processes, we took one portion out and their DNAm ages were then estimated by the model trained on the remaining 9 portions. After 10 rounds, the DNAm ages of samples from all ten portions were obtained and they were not overfitted by the training process.

BrainCortexClock is trained on 640 cerebellum samples, which are from GSE134379 [158], GSE105109 [160], GSE125895 [161], GSE61431 [162] and GSE72778 [75], and 720 cerebral cortex samples, which are from GSE134379 [158], GSE61431 [162], GSE80970

Table 4.1: Lists of the four new clocks constructed in this chapter

Clocks	Training involved CpGs	Training involved tissues	Training involved datasets
$\text{CerebellumClock}_{\text{specific}}$	613 age-associated CpG in CBL	Cerebellum	GSE134379, GSE59685, GSE105109, GSE125895, GSE61431, GSE72778
$\text{CerebellumClock}_{\text{common}}$	201 age-associated CpG in both CBL and MTG	Cerebellum	GSE134379, GSE59685, GSE105109, GSE125895, GSE61431, GSE72778
$\text{CerebralCortexClock}_{\text{common}}$	201 age-associated CpG in both CBL and MTG	Middle temporal gyrus, dorsolateral prefrontal cortex, prefrontal cortex, prefrontal cortex	GSE134379, GSE74193, GSE80970, GSE61431
BrainCortexClock	201 age-associated CpG in both CBL and MTG	Cerebellum, middle temporal gyrus, dorsolateral prefrontal cortex, prefrontal cortex, prefrontal cortex	GSE134379, GSE59685, GSE105109, GSE125895, GSE61431, GSE72778, GSE74193, GSE80970

[164], by Elastic net linear regression algorithm, and the input CpG sites were restricted to the 201 shared age-associated CpGs, and alpha was set to 0.2. As with the training of $\text{CerebralCortexClock}_{\text{common}}$, ten-fold cross-validation was used to measure the age prediction performance of BrainCortexClock, and it was further tested by applying to an independent dataset of GSE59685 [159].

The four new clocks constructed in this chapter are listed in Table 4.1. The coefficients of involved CpGs in each model are listed in B.5.

4.2.6 Software

All the analyses were conducted in R (version 3.6.0) [116] under a Linux environment. The scatter plots in Figure 4.1, 4.3, 4.6 were produced by the *getAccel* function with proper settings from the 'dnaMethyAge' R package [167]. The three constructed models of $\text{CerebellumClock}_{\text{specific}}$, $\text{CerebellumClock}_{\text{common}}$ and $\text{CerebralCortexClock}_{\text{common}}$ are read-

ily available to be applied in independent DNAm samples by calling the *methyAge* function from the 'dnaMethyAge' R package [167] with the 'clock' parameter setting as 'Cerebellum_specific', 'Cerebellum_common' and 'Cortex_common' respectively. GO analyses were conducted using the *gometh* function in the 'missMethyl' package [171] which tests gene ontology enrichment for significant CpGs while accounting

4.3 Results

4.3.1 Characteristics of the DNAm cerebellum datasets

The cerebellum is a structure of the hindbrain, which plays a vital role in motor control [172]. Unlike peripheral tissues, such as blood or saliva, that can be non-invasively and repeatedly sampled, cerebellum samples are often collected from postmortem participants, as a result, there is a very limited number of DNAm cerebellum samples available. After rigorous searching on the Gene Expression Omnibus database, where publicly available DNAm datasets are often deposited, we found a total of 6 datasets, each including more than ten cerebellum samples measured by Illumina 450k or EPIC array. After rigorous quality control (see methods), 752 cerebellum samples remained and were used for downstream analysis. The biggest contributor for the final large cerebellum dataset is from GSE134379 [158], which contains 404 cerebellum samples. As cerebellum tissues were invasively collected from postmortem subjects, 90% of the collected samples were from individuals aged above 60 years old, with the median age at 80 years old. More detailed age, sex and disease distribution information for each dataset is listed in Table 4.2. The DNAm microarray data from those datasets were originally produced to investigate disease-associated methylomic variations in the brain regions, especially for Alzheimer's disease and Schizophrenia. As a result of this, our col-

Table 4.2: Characteristics of the clean cerebellum samples from six datasets

ID	Number	Female, Male	Age: mean (range)	Disease group	Reference
GSE134379	404	200, 204	83.7 (54-103)	Alzheimer: 225 Normal: 179	[158]
GSE59685	111	64, 47	83.9 (40-105)	Alzheimer: 59 Normal: 52	[159]
GSE105109	95	41,54	81.2 (58-99)	Alzheimer: 67 Normal: 28	[160]
GSE125895	66	32, 34	67.3 (51.8-92.3)	Alzheimer: 24 Normal: 42	[161]
GSE61431	44	16, 28	61.6 (25-96)	Schizophrenia: 21 Normal: 23	[162]
GSE72778	32	21, 11	83.2 (15-114)	Alzheimer: 23 Normal: 9	[75]

lected cerebellum samples include 333 samples with normal health status, 398 samples with Alzheimer's disease and 21 with Schizophrenia. The cerebellum is a relatively protected region, unlike other brain regions (such as prefrontal cortex), there generally are no significant AD-associated differences in the cerebellum [173, 159, 161]. Therefore, we included all these cerebellum samples, even those with disease diagnosis, for downstream analysis and also the following cerebellum DNAm age clock construction.

4.3.2 Severe age underestimation

for cerebellum samples by various DNAm age clocks

Since 2013, many specialized and robust DNAm-based clocks have been reported. As recently suggested by Liu et al., those different clocks may have captured different biological processes of ageing considering their overall weak associations in the estimated DNAm age deviations [174]. Inspired by this, we investigated the DNAm ages of our collected 752 cerebellum samples predicted by six representative clocks: Hannum's whole blood clock (Hannum2013) [39] and Horvath's pan-tissue clock (Horvath2013)[40] are the two most widely

used DNAm age clocks and especially Horvath2013 is reported to work well across many different tissue and cell types; Horvath's blood&skin clock (Horvath2018) [155] is another multi-tissue clock and was reported to outperform Horvath2013 in epigenetic age prediction across several tissues; Levine's PhenoAge clock (Levine2018) [81] was not directly regressing on chronological age and reported better prediction performance for all-cause mortality than other chronological age regressed clocks; Zhang's blood clock (Zhang2019) [78] is reported the most accurate and robust age prediction model for blood samples; Shireby's brain cortex clock (Shireby2020) [70] is a brain cortex specific clock and provides much better age predictions than other clocks in brain cortex tissues.

As shown in Figure 4.1, for almost all of the cerebellum samples, their ages are severely underestimated—they are all distributed below the diagonal lines. Hannum2013, Levine2018 and Zhang2019 are three age clocks trained almost exclusively on blood samples, the root-mean-square deviations (RMSDs) of their predictions are all very large (above 40 years), with Pearson correlations (r) ranging from 0.182 in Levine2018 and 0.56 in Zhang2019 (Figure 4.1a-c); Horvath2018 is a multi-tissue clock that was trained on eight different tissues cell types but not including brain-related tissues, it produced a similar prediction trend (Figure 4.1d) for cerebellum samples as the three blood clocks—large deviations (RMSD=66.9 years) and low correlation (r=0.452). In contrast, the age underestimation effect is less apparent for Horvath2013 and Shireby2020 (Figure 4.1e and 4.1f), their RMSDs are just above 20 years and the Pearson correlation coefficient reached 0.699 by Shireby2020 and 0.694 by Horvath2013. We speculate the smaller age underestimation effects by the two clocks are due to their training datasets having included a small ratio of cerebellum samples or entirely on brain cortex tissues. Specifically, Horvath2013 was trained on 8000 samples from 51 different tissues and cell types which include several different brain-related tissues including 282 cerebellum samples [40], while Shireby2020 was trained exclusively on brain cerebral cortices despite cerebellums not being involved [70]. It is worth noting, the regression lines

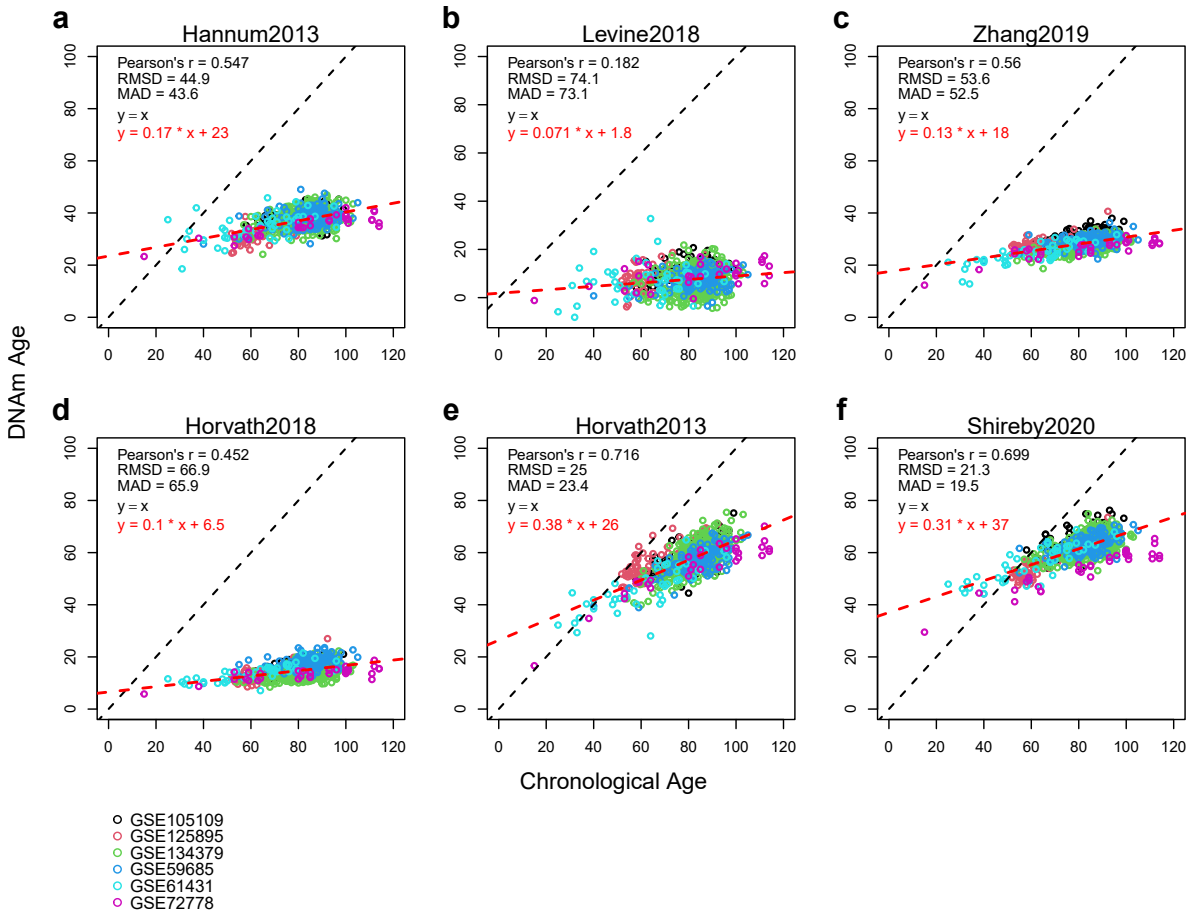

Figure 4.1: The cerebellum samples are severely underestimated by the six representative DNAm clocks. Each subplot illustrates results from different clocks: (**a**) Hannum2013, (**b**) Levine2018, (**c**) Zhang2019, (**d**) Horvath2018, (**e**) Horvath2013 and (**f**) Shireby2020. The colorful dots represent 752 cerebellum samples from six independent datasets, with different colors representing different datasets. The x-axis is chronological age and the y-axis is the estimated DNAm age. The black dashed line represents the identical diagonal line between chronological age and DNAm age, the red dashed line represents the regression line derived from regressing the DNAm age against the chronological age. RMSD: root mean squared deviation; MAD: mean absolute deviation.

of the estimated DNAm age against the chronological age by Horvath2013 and Shireby2020 both indicate that the cerebellum samples from young individuals aged below 30 years old are very likely to be overestimated (Figure 4.1e and 4.1f).

4.3.3 Smaller number of age-associated CpGs in the cerebellum methylome

We went further to investigate the underlying reasons why the cerebellum is systematically underestimated by the six age clocks. We hypothesized that, if the cerebellum truly ages slower than most other brain tissues, then due to a smaller ageing effect, there would be a much smaller number of CpGs passing the same cutoff to be identified as age-associated and even those captured age-associated CpGs would mostly exhibit a smaller rate of methylation level changes with age. Inspired by this, we carried out two epigenome-wide association studies (EWAS) on age for the cerebellum (CBL) and the middle temporal gyrus (MTG) separately, based on the same dataset GSE134379 [158] which includes DNA methylation microarray samples of the two brain regions for every subject from a large elderly population (n=404).

We identified a total of 613 significant (Bonferroni-corrected P-value ≤ 0.01) age-associated CpGs in CBL, in contrast, 4,213 CpGs were found to be age-associated in MTG (Figures 4.2a, 4.2b and Supplementary Tables 1). The top three age-associated CpGs in CBL are cg24079702, cg22454769, and cg06639320, which are all mapped to the *FHL2* gene, whereas, the three loci exhibited similar age effect sizes though were less significant in MTG. When the age-associated CpGs in the two tissues were compared, only 32.8 % (201) of the CpGs in CBL were also identified as age-associated in MTG (Figure 4.2d). More interestingly, when looking at the direction of ageing effect, CBL and MTG showed very different patterns

in their age-associated CpGs. The CBL-only group has almost equal numbers of positive and negative age associations, in contrast, more than three-quarters (76%) of the MTG-only CpGs gain methylation with ageing. Moreover, the majority (94%) of the age-associated CpGs shared in the two tissues increase methylation levels with ageing (Figure 4.2d), this is not very unexpected, as it has been shown that CpG sites exhibiting age-association in multiple tissues are more likely to gain methylation with age [60].

If the cerebellum ages slower, then it is reasonable to expect that the age-associated CpGs in the cerebellum would also have smaller ageing effect sizes. We then compared the ageing effects of age-associated CpGs between CBL and MTG (Figure 4.2b). Indeed, as shown in Figure 4.2c, the ageing effect size of positive age-associated CpGs in CBL is generally smaller than that in the MTG (Wilcoxon test, Bonferroni-corrected $p = 0.01$), though their difference is not significant in the negative age-associated CpGs ($p = 0.11$). As regards the 201 shared age-associated CpGs (Figure 4.2e), the difference in ageing effect size between CBL and MTG—lower in CBL than MTG, is much more significant (Pairwise Wilcoxon test, Bonferroni-corrected p <1.5e-15).

Gene ontology analyses showed several enriched terms for the MTG-specific CpGs and Cerebellum-specific CpGs (Table B.2 and B.1) which included terms related to chromatin such as DNA binding, nucleosome assembly and negative regulation of transcription by RNA polymerase II. The MTG-specific CpGs were enriched for pathways such as telomere organization, noradrenergic neuron differentiation and dopaminergic neuron differentiation. Telomere shortening and neuron differentiation are both characteristics of cell mitotic divisions in cerebral cortex, thus it suggests the MTG has a higher cell replication rate than cerebellum. In addition, the enriched GO terms for the cerebellum-specific CpGs were related to molecular functions such as DNA binding activity.

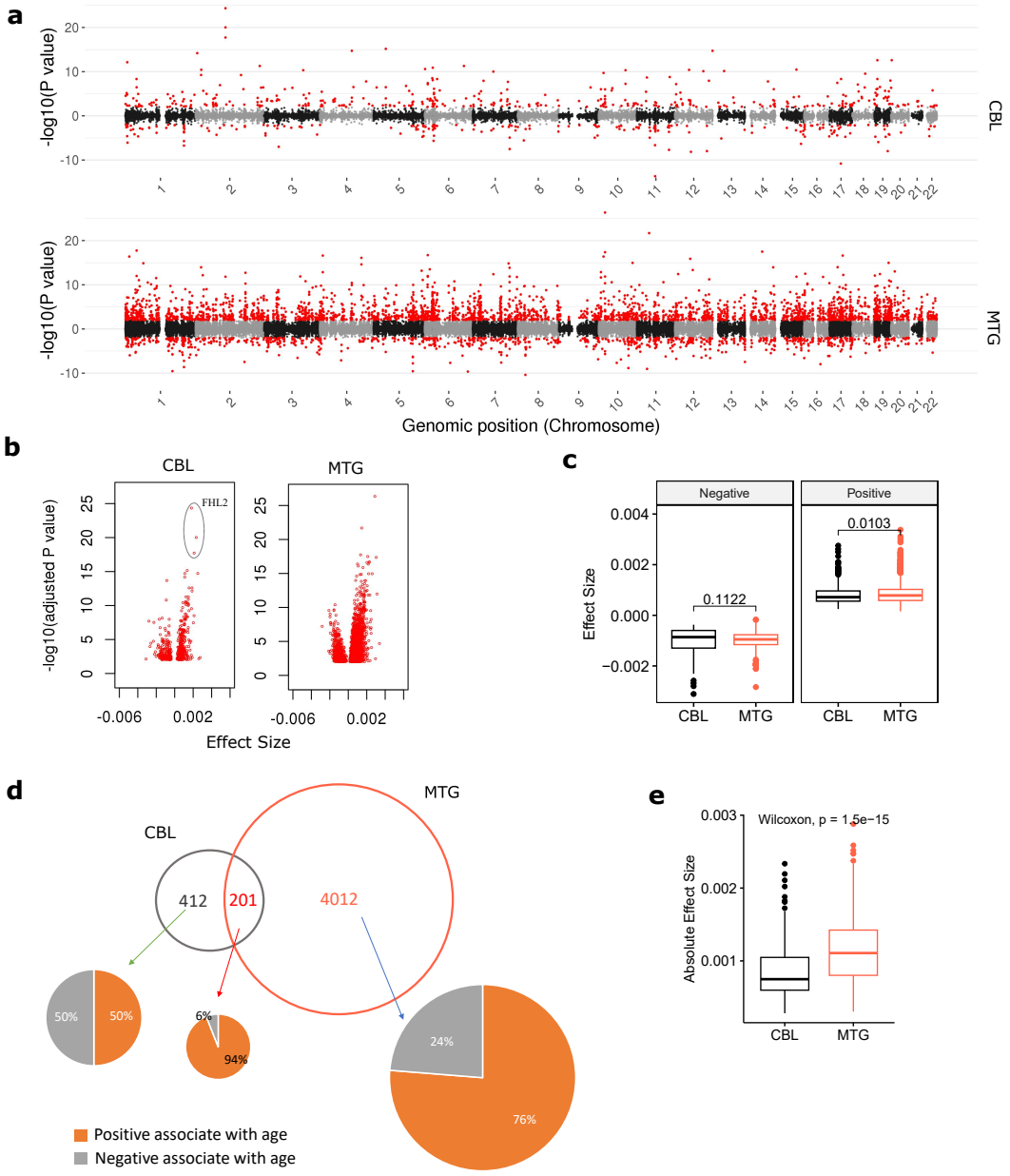

Figure 4.2: Comparison of age-associated methylation change between the cerebellum (CBL) and the middle temporal gyrus (MTG). (**a**) Manhattan plots illustrate the age EWASs results of CBL and MTG. Red dots denote significant age-associated CpGs (adjusted P-value ≤ 0.01). (**b**) Two volcano plots show the effect size distribution of significant age-associated CpGs in CBL and MTG. (**c**) Boxplots comparing the age effect size in CBL and MTG for positive and negative age-associated CpGs, Wilcoxon Tests were performed and Bonferroni-corrected P-values are displayed. (**d**) Venn plot shows the unique and shared number of the top 613 most significant age-associated CpGs in CBL and MTG. The three pie charts illustrate the proportions of CpGs gain methylation (positive associate with age) or lose methylation (negative associate with age) with age in three categories. (**e**) Boxplots comparing the absolute values of age effect sizes in CBL and MTG for the 201 shared age-associated CpG sites, Pairwise Wilcoxon Tests were performed and Bonferroni-corrected P-value is displayed.

4.3.4 Constructing DNAm age clocks for the cerebellum

4.3.4.1 Training the cerebellum specific DNAm clock

Our analyses in the previous section have clearly demonstrated that the six representative DNAm age clocks, including the pan-tissue clock and the cerebral cortex clock, all severely underestimated epigenetic ages of cerebellum samples. In addition, we have shown that the cerebellum has a much smaller number of age-associated CpGs. Then we went further to find out whether it is possible to build an accurate age prediction model for the cerebellum.

We trained a cerebellum-specific age model, named $\text{CerebellumClock}_{\text{specific}}$, by regressing the methylation beta values of the 613 age-associated CpGs from the 752 clean cerebellum samples against their corresponding chronological ages via the Elastic Net penalized linear regression algorithm [169]. The prediction performance of this model was measured by leave-one(dataset)-out cross validation (see Methods). As shown in Figure 4.3a, the cross-validation results demonstrate that the trained cerebellum age models yield accurate age predictions for nearly all cerebellum datasets, except that most of the elderly subjects in GSE72778 were relatively underestimated. The overall Pearson correlation is above 0.94, with RMSD at 4.26 years and MAD is 3.18 years. The accurate age prediction performance of the cerebellum age model demonstrates that there is a persistent and significant ageing process undergoing in the cerebellum tissues.

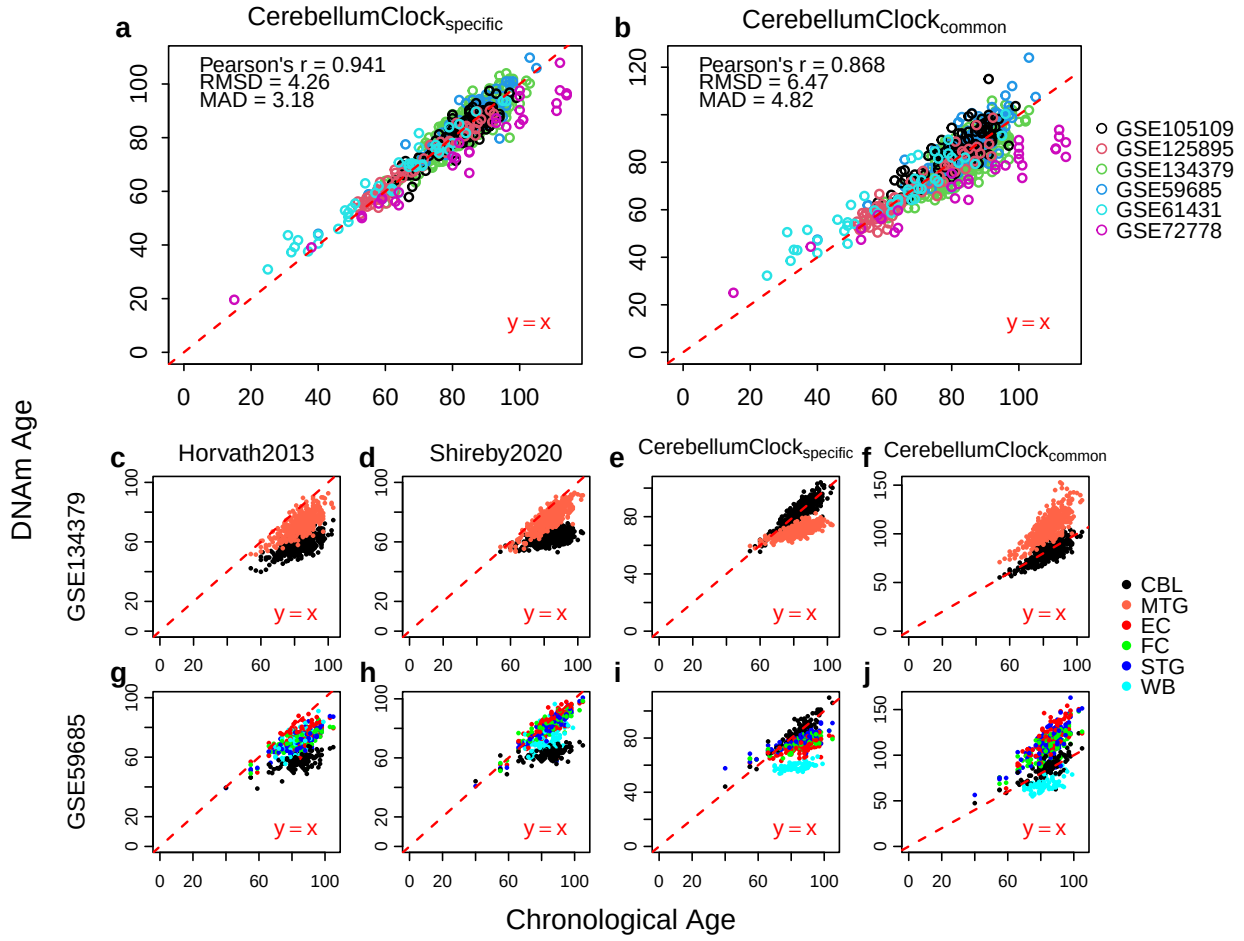

Figure 4.3: The cerebellum age clocks and their applications in other tissues. The leave-one(dataset)-out cross validation evaluates the age prediction performance of (**a**) $\text{CerebellumClock}_{\text{specific}}$ and (**b**) $\text{CerebellumClock}_{\text{common}}$ in cerebellum samples. Subplots (**c**), (**d**), (**e**) and (**f**) compare the DNAm age of CBL and MTG estimated by Horvath2013, Shireby2020, $\text{CerebellumClock}_{\text{specific}}$ and $\text{CerebellumClock}_{\text{common}}$ respectively. Similarly, Subplots (**g**), (**h**), (**i**) and (**j**) compare the DNAm age of five different tissues estimated by the same four clocks. CBL: cerebellum, MTG: middle temporal gyrus, EC: Entorhinal Cortex, FC: Frontal Cortex, STG: Superior Temporal Gyrus, WB: Whole Blood.

4.3.4.2 Applying the cerebellum clocks in other tissues

To further examine the claim that cerebellum ages slower, we made another hypothesis: other tissues, including cerebral cortex and blood, would be significantly overestimated for their DNAm ages when measured by the cerebellum clock. To test this hypothesis, we then applied $\text{CerebellumClock}_{\text{specific}}$ along with Horvath2013 and Shireby2020 in two separate datasets: GSE134379 and GSE59685, which both include cerebellum samples and samples of other tissues from the same subject. As expected, the cerebellum samples were apparently underestimated compared to other tissues by Horvath2013 and Shireby2020 in both GSE134379 and GSE59685 (Figure 4.3). Interestingly, even though blood was also not included in the training set of Shireby2020, the predicated DNAm ages of blood samples in GSE59685 are still much higher than their counterparts in the cerebellum tissue (Figure 4.3h).

However, when estimated by the cerebellum clock, the non-cerebellar samples were actually underestimated rather than overestimated compared to the cerebellum samples (Figure 4.3i). This finding counters our previous expectation, we suggest the underestimation effect for other tissues by the cerebellum clock may rather imply that this age model is working poorly in non-cerebellar tissues. In addition, we discovered that the cerebellum clock tends to overestimate the ages of non-cerebellar samples under 60 years old (Figures 4.3e and 4.3i). This is further confirmed by looking at the overestimation facts for cortex tissues from young subjects by the cerebellum clock (4.4). The penalized regression algorithm selected 275 age-associated CpGs from the 613 age-associated CpGs in the cerebellum, where the majority of them (73%) are on the CBL-only list, meaning they do not exhibit significant age correlations in MTG. Thus the observed apparent underestimation effect for those non-cerebellar samples is not biologically meaningful, instead indicating artefacts resulting from improper usage of the age model $\text{CerebellumClock}_{\text{specific}}$.

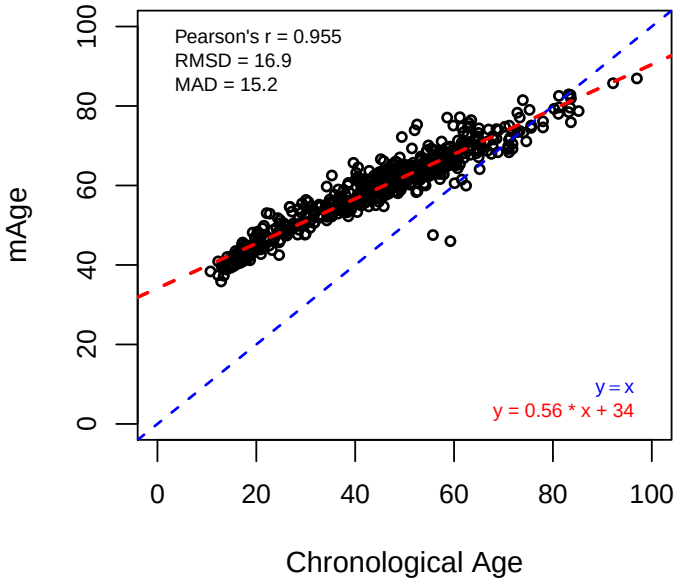

Figure 4.4: Young cortex tissues are overestimated by CerebellumClock$_{\text{specific}}$. The cortex samples are from GSE74193.

4.3.5 Slower ageing rate in cerebellum according to two oppositely designed models

The above model CerebellumClock$_{\text{specific}}$ thus captures cerebellum-specific age-related changes. In order to make more fair ageing rate comparisons, we then trained another cerebellum age model with the same regression algorithm and the same training samples except the input CpG set is restricted to the 201 CpGs that are age-associated in both CBL and MTG. The leave-one(dataset)-out cross validation demonstrated that the new cerebellum clock, named CerebellumClock$_{\text{common}}$, still gives very good age predictions for those cerebellum samples (Figure 4.3b). Notably, CerebellumClock$_{\text{common}}$ substantially overestimated the ages of brain cerebral cortices in both GSE134379 and GSE59685, though the ages of blood samples in GSE59685 were still underestimated (Figures 4.3f and 4.3j).

To further confirm the overestimation effect for non-cerebellar brain tissues by the new cerebellum clock, we applied the $\text{CerebellumClock}_{\text{common}}$ to two other independent datasets which, combined, include a large number of samples from three parts of cerebral cortex with a wide age range (20∼100 years old). The results shown in Figure 4.6a demonstrate that $\text{CerebellumClock}_{\text{common}}$ substantially overestimates the whole age range of non-cerebellar brain tissues. The overall Pearson correlation coefficient reached 0.951, indicating the new cerebellum clock has also captured the strong ageing effect on the methylome of those tissues. More importantly, the slope of the regression line obtained from regressing the predicted DNAm age against the chronological age is greater than 1 (Slope = 1.2), indicating that these non-cerebellar tissues have higher ageing tick rates than the cerebellum.

Likewise, we constructed another cerebral cortical clock, in which the training dataset includes samples from different parts of cerebral cortex, and the input CpG set was limited to 201 shared age-related CpGs. The resulting model, named $\text{CerebralCortexClock}_{\text{common}}$, performed well for samples from tissues that have been included in the training dataset (Figure 4.5). We then applied it to the clean cerebellum dataset (n=752) we collected. As expected, all the cerebellum samples were largely underestimated by $\text{CerebralCortexClock}_{\text{common}}$. Furthermore, the increasing deviations of the estimated DNAm ages from their chronological ages and the lower than 1 slope value of the regression line (Slope = 0.54) indicate that the cerebellum ticks at a slower rate than other brain cortex tissues.

Altogether, we arrive at the same conclusion from the two different analyses—the cerebellum has a smaller ageing tick rate when measured by models constructed by the same set of CpGs which were selected given they are age-associated in both CBL and MTG.

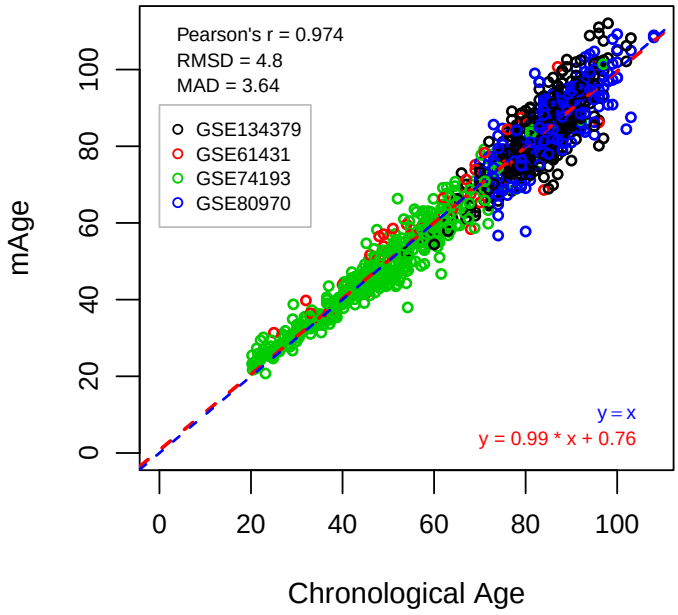

Figure 4.5: The leave-one(fold)-out cross validation evaluate the age prediction performance of $\text{CortexClock}_{\text{common}}$. Colors represent samples from different datasets.

4.3.6 Why does the cerebellum appear to age slowly

We then sought to understand the underlying reasons why the cerebellum clock ($\text{CerebellumClock}_{\text{common}}$) overestimated the ages of non-cerebellar brain tissues and the cerebral cortex clock ($\text{CerebralCortexClock}_{\text{common}}$) underestimated the cerebellum tissue. Comparing the overall methylation levels, the cerebellum has an apparent lower median methylation level than MTG (Figure 4.7a) and it also has the lowest median methylation level among the five tissue types included in GSE59685 (Figure 4.7b). When grouping all CpGs into four genomic categories, i.e. island, open sea, shelf and shore, the mean (Figure 4.7c) and median (Figure 4.7d) methylation comparison analysis both agreed that the cerebellum is less methylated in the island and the shore. Thus, it is reasonable to conclude that the overall lower methylation level in the cerebellum mainly originated from its lower

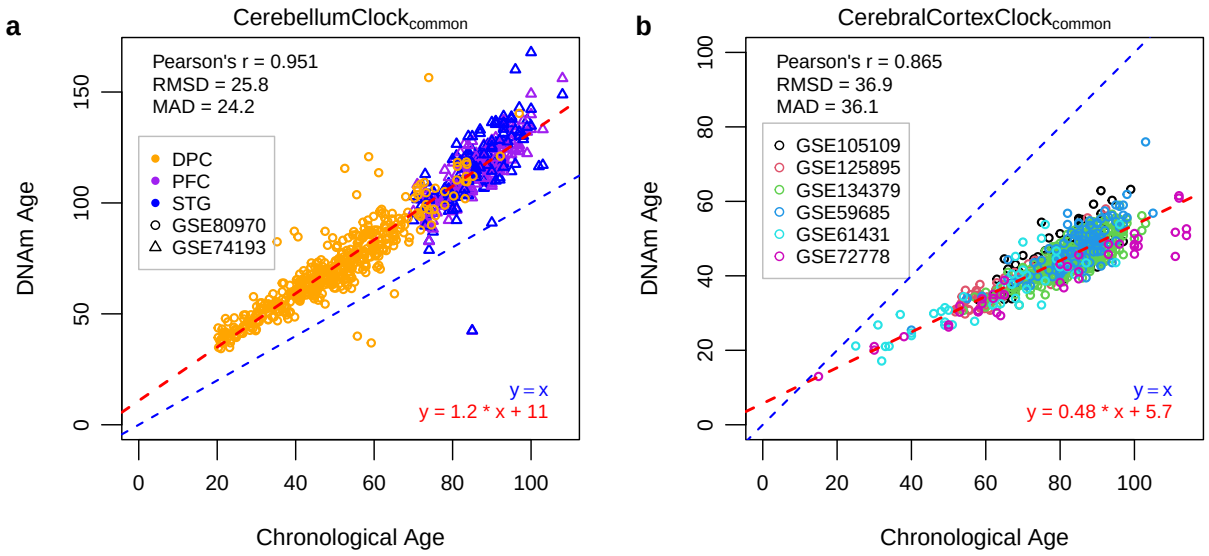

Figure 4.6: (a) The DNAm ages of samples of three different parts of human cerebral cortex (DPC: Dorsolateral Prefrontal Cortex, PC: Prefrontal Cortex, STG: Superior Temporal Gyrus) in two datasets are systematically overestimated by $\text{CerebellumClock}_{\text{common}}$. (**b**) the cerebellum samples are all severely underestimated by the $\text{CerebralCortexClock}_{\text{common}}$. The $\text{CerebellumClock}_{\text{common}}$ and $\text{CerebralCortexClock}_{\text{common}}$ were trained from the same set of CpGs (n=201) but in different tissues. The blue dashed line represents the identical diagonal line between chronological age and DNAm age, the red dashed line represents the regression line derived from regressing the DNAm age against the chronological age.

methylation level in the CpG island and the shore. It should be noted we did not detect any significant correlations between mean methylation level change with age in any tissue types or the four genomic categories (Figure 4.8 and 4.9), indicating the overall lower methylation level in the cerebellum is not due to a different ageing rate.

Next, we focused on the 201 common age-associated CpGs that were used to build $\text{CerebellumClock}_{\text{common}}$ and $\text{CerebralCortexClock}_{\text{common}}$. Firstly, we have shown that 94% of them gained methylation with age. Secondly, there are 140 CpGs on the island and 48 CpGs on the shore, they accounted for 93.5% of the 201 common CpGs (Figure 4.7e). Consistent with that CBL has generally lower methylation levels in the CpG islands and shores than MTG, we found that the majority (85%, 171) of the common age-associated CpGs also have lower mean/median methylation levels in the cerebellum (Figure 4.7f). Lastly, more than three-quarters of the 201 CpGs turned out to have a smaller ageing effect size in the CBL than MTG when regressing the methylation beta values against age, sex and batches (Figure 4.7f), meaning those CpGs have higher rates of age-associated methylation change in the MTG than the cerebellum. Altogether, the lower methylation levels, the positive age associations and smaller ageing effect sizes of the majority of the common 201 CpGs in the cerebellum explain why $\text{CerebellumClock}_{\text{common}}$ not only systematically overestimated the ages of non-cerebellar brain tissues (Intercept=11) but also with overestimation effect more prominent with age (Slope=1.2). Similarly, they also explain why the cerebellum samples were systematically underestimated by the $\text{CerebralCortexClock}_{\text{common}}$.

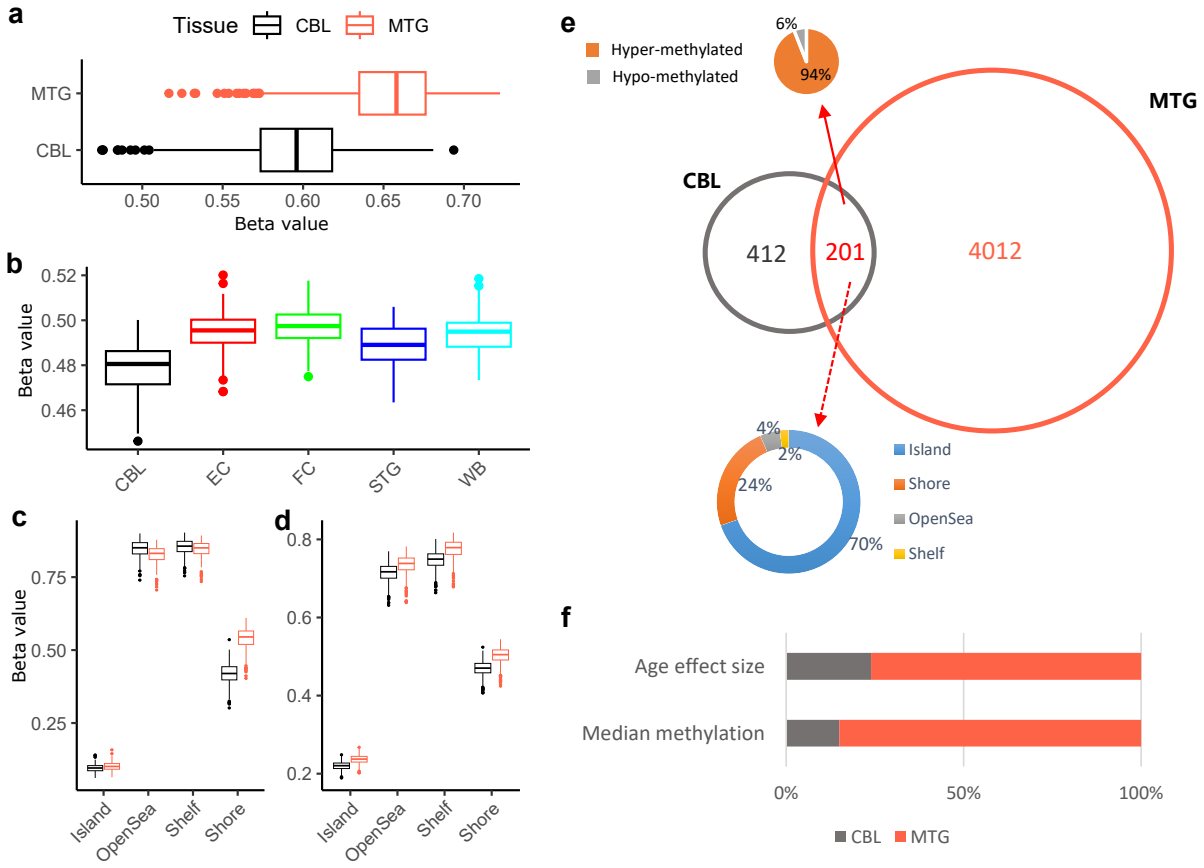

Figure 4.7: Boxplots illustrating the cerebellum (CBL) has a lower median overall methylation level than (**a**) MTG or (**b**) other four tissues (i.e. EC, FC, STG and WB). (**c**) Median and (**d**) mean methylation comparison both agree the cerebellum has a lower methylation level in the CpG island and shore. Among the 201 common age-associated CpGs, 70% of them are located on the island and 24% of them are located on the shore. The barplot in (**f**) shows the proportions of CpGs, which have larger age effect sizes or higher median methylation levels in CBL and MTG.

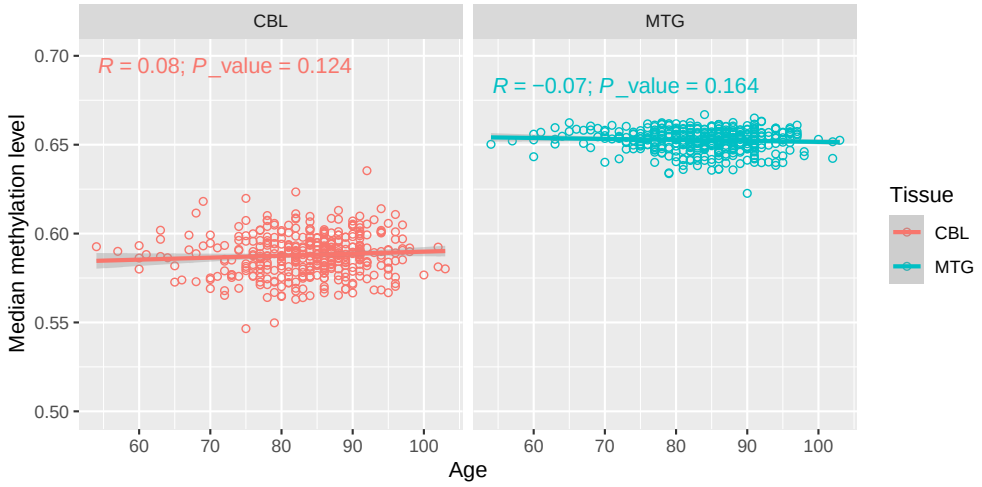

Figure 4.8: The fluctuation of the median methylation level is not correlated with chronological age either in the CBL or in the MTG.

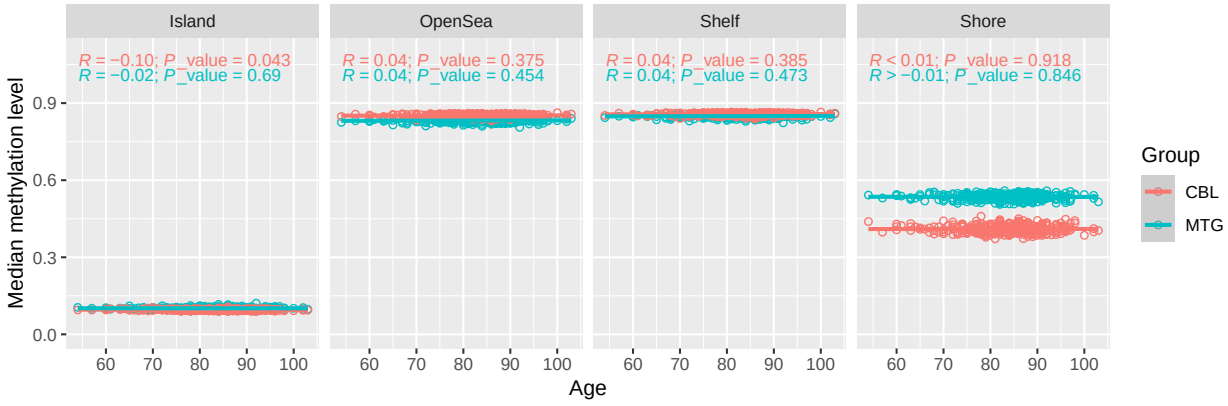

Figure 4.9: The fluctuation of the median methylation level is not correlated with chronological age in any of the four genomic regions, i.e. island, open sea, shelf and shore.Red denotes CBL sample and cyan denotes MTG sample.

4.3.7 A single clock unbiasedly estimates DNAm age of cerebellum and cerebral cortex

Though we have demonstrated that cerebellum shows different ageing patterns even on the shared 201 CpGs compared with cerebral cortex, the successful construction of $\text{CerebellumClock}_{\text{common}}$ and $\text{CerebralCortexClock}_{\text{common}}$ inspired us to investigate whether it is possible to build a single clock that works well, i.e. no systematic offset, for samples from both cerebral and cerebellar cortices.

To start with, we selected a comparable number of cerebral cortex samples to cerebellar samples, while also ensuring that they have similar age distributions (Figure 4.10a). Then the Elastic Net was applied to regress the methylation values of the 201 CpGs against the chronological age of samples from the two brain cortex tissues. Remarkably, the leave-one-fold-out cross-validation assessment showing the new brain cortex clock, named as BrainCortexClock, did accurately predict the age of samples from both cerebral and cerebellar cortices—the Pearson correlation coefficient reaches 0.906 and MAD is 3.83 years (Figures 4.10b and 4.10c). We further tested its performance on an independent dataset—GSE59685 which includes DNAm samples from multiple tissues from 122 participants. Set aside the blood samples, the evaluation matrix generated from the brain cortex tissues further confirmed BrainCortexClock's accurate age prediction performance (Figures 4.10d). The boxplots in Figure 4.10e demonstrate the cerebellum samples were not systematically underestimated and the cerebellum and cerebral cortex have similar levels of DNAm ages as estimated by this new clock, in contrast, the blood samples were apparently underestimated due to a lack of representation of this tissue in the training dataset. Furthermore, age acceleration comparisons between any two tissues from the same subjects showed the variations of age acceleration between cerebellum and other three parts of cerebral cortex were mod-

erately correlated, with Pearson's r ranging from 0.37 to 0.52 (Figure 4.10e). Altogether, BrainCortexClock provides unbiased DNAm age prediction for brain cortex tissues including cerebellum.

4.4 Discussion

In order to examine the claim that the cerebellum ages slower, we collected a large set of cerebellum samples (N=752) and assessed their DNAm ages from six representative clocks, including Horvath's multi-tissue clock, i.e. Horvath2013. The results showed that these six representative clocks severely underestimated almost all cerebellum samples. This is consistent with previous reports [74, 157]. However, we should not conclude that the cerebellum ages slower only based on these results, as the underestimations may mainly reflect the improper usage of DNAm clocks, i.e. applying DNAm clocks in tissues which do not have adequate representations in the clocks' training datasets. We found the underestimations were much more severe with the four clocks that were trained with no brain-related tissues—three clocks were trained mainly on blood tissues and Horvath2018 was trained on eight other different tissues. In contrast, the underestimations were much attenuated in Shireby2020 whose training samples comprised non-cerebellar cortex tissues and Horvath2013 which included 282 cerebellum samples in its total 8000 training samples. Different tissues may have distinct DNA methylation profiles, and the dynamic changes of their methylomes in response to ageing also vary [60]. Horvath's multi-tissue clock produces relatively accurate age predictions for many vast different tissue/cell types [40]. Still, there is no evidence or guarantee to claim that it has captured the intrinsic mechanism that drives the DNAm changes across the whole body. We do not think it is justified to compare the ageing rates of different tissues by simply comparing their DNAm ages derived from the multi-tissue clock.

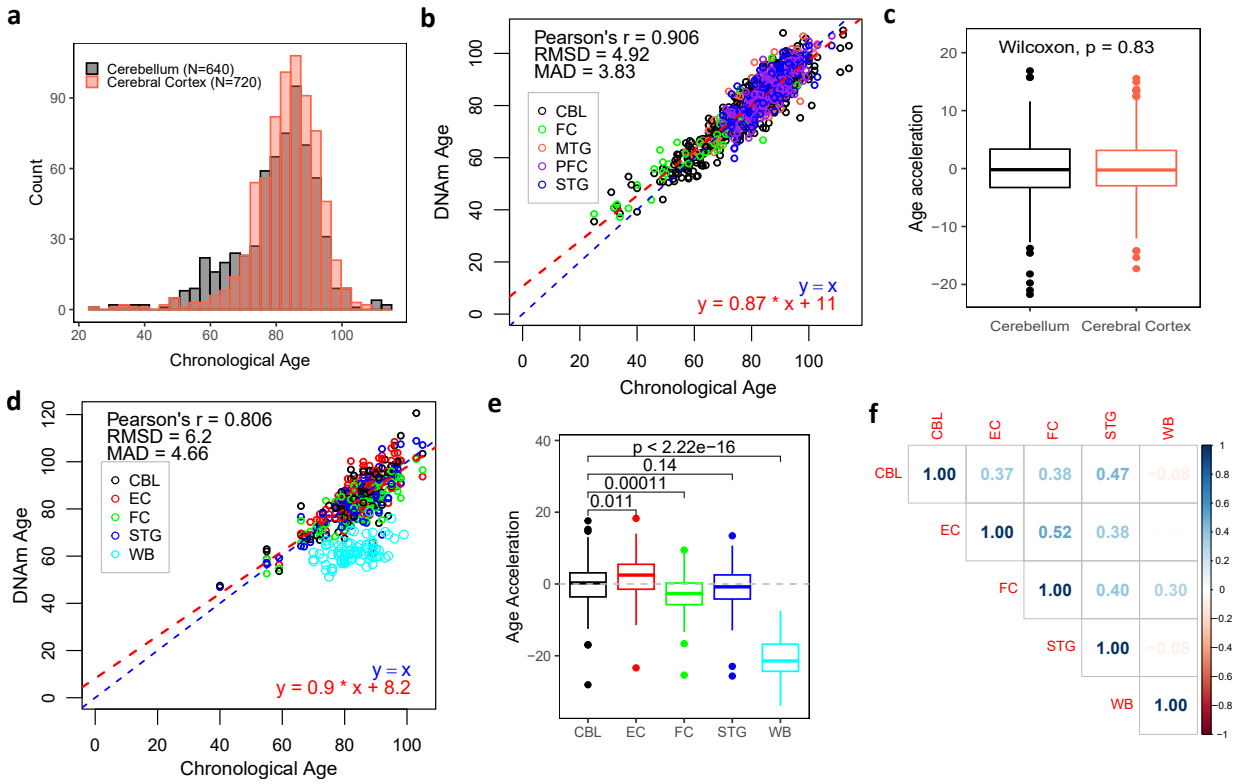

Figure 4.10: The clock of BrainCortexClock unbiasedly estimates DNAm age of cerebellum and cerebral cortex. (a) Age distributions of cerebellum samples and cerebral cortex samples in the training dataset. (b) Leave-one-fold-out cross-validation reveals the high performance of BrainCortexClock in training dataset. (c) Comparing age accelerations of samples from cerebellum and cerebral cortex in the training dataset demonstrates BrainCortexClock is not biased in the two tissues. The P-value was obtained from unpaired Wilcoxon Tests. (d) The performance of BrainCortexClock is evaluated in an independent dataset—GSE59685 which includes DNAm samples of five different tissues from 121 individuals. Note, the evaluation matrixes were drawn from samples that excluded whole blood. (e) Boxplots showing cerebellum samples are not systematically underestimated than three other parts of cerebral cortex, in contrast, whole blood is apparently underestimated. The P-values were obtained from pairwise Wilcoxon Tests. (f) Correlation matrix showing variations of age acceleration between cerebellum and other three parts of cerebral cortex were moderately correlated and all four brain tissues were poorly correlated with whole blood. CBL: cerebellum, MTG: middle temporal gyrus, EC: Entorhinal Cortex, FC: Frontal Cortex, STG: Superior Temporal Gyrus, MTG: Middle Temporal Gyrus, PFC: Prefrontal Cortex, WB: Whole Blood.

There exists a strong and consistent ageing effect on the DNA methylome of the cerebellum. By performing age EWAS on the cerebellum, we found 613 significantly age-associated CpGs from an elderly population, they were scattered across all autosomes. By taking advantage of penalised linear regression algorithm and a large training dataset, we constructed a highly accurate age clock for cerebellum ($\text{CerebellumClock}_{\text{specific}}$, r=0.941, MAE=3.18 years). As a comparison, we identified many more age-associated CpG sites in a representative cerebral cortex tissue—MTG, and we found the CBL has smaller age effect sizes than the MTG although it is only significant in the positive age-associated CpGs. We found 201 CpGs exhibiting age associations in both CBL and MTG, based on these 201 CpGs, we trained two clocks, i.e. $\text{CerebellumClock}_{\text{common}}$ and $\text{CerebralCortexClock}_{\text{common}}$, on all cerebellum samples and non-cerebellar cortex samples separately, they both performed well in age prediction for tissues that have included in their training dataset. When the two clocks are applied to samples from the cerebellum and non-cerebellar cortex tissues and the estimated DNAm ages are compared, they both agree that the cerebellum has a younger epigenetic age and a lower ageing rate. Furthermore, we have demonstrated that this is caused by 94% of the 201 CpGs gaining methylation with age, 85% are less methylated in CBL, and more than 75% have a smaller ageing effect size in CBL.

Even though our finding supports that the cerebellar methylome is more resistant to change with ageing, we should be cautious about whether this can be translated to the conclusion that the cerebellum is biologically younger than other human tissues. It should be noted that the above comparisons of ageing rates between cerebellum and MTG are based on the clocks trained on the same 201 age-associated CpGs. In fact, there is more than twice the number of CpGs found to be age-associated in the cerebellum and even more in MTG. When we apply the cerebellum-specific clock ($\text{CerebellumClock}_{\text{specific}}$), which was trained by using all age-associated CpGs in the cerebellum, in predicting DNAm ages of other brain tissues, we could no longer observe a systematic overestimation for samples across all age

groups, instead only the individuals aged below 60 years old were overestimated, by contrast, the above 60 years old group was clearly underestimated (Figures 4.3e, 4.3i and 4.4). We conclude that this is due to the improper usage of the clock, as the $\text{CerebellumClock}_{\text{specific}}$ consists of many cerebellum-specific CpGs.

Why does the cerebellum have a much smaller number of age-associated CpGs? The observed age-associated methylation level change of CpGs sites in tissues with mixed cell types could arise through epigenetic drifts with mitotic divisions, cell type composition changes and intrinsic changes affected by cell inner metabolism. Above 80% of cells in the cerebellar grey matter are non-replicating neuronal cells [175]. As a result of this, retrospective birth dating of cells through ^{14}C bomb-pulse method indicates the average cell turnover rate in cerebellum is extremely low. In contrast, a much higher proportion of non-neuronal cells (mainly glial cells) in cerebral cortex makes it have a higher average cell turnover than the cerebellum [176]. *ELOVL2* hypermethylation has been demonstrated as a marker of cell divisions that occur throughout human ageing [177], the hypermethylation of a locus in *ELOVL2* which targeted by the probe of cg16867657, has been reported to show highest age correlation in whole blood [178, 179, 60]. Our results show that hypermethylation of cg16867657 is still significant (p=7.0e-08, effect size=0.0011) correlated with age in cerebellum though is much less significant than that in MTG (p=1.5e-12, effect size=0.0023), this is consistent with very low average mitotic rates in cerebellum. *FHL2* is another well-documented gene whose hypermethylation is strongly correlated with age [61, 180, 179]. Unlike *ELOVL2*, *FHL2* hypermethylation is not closely associated with cell replication [177]. Remarkably, the top three age-associated CpG sites in cerebellum are all mapped to *FHL2* gene and they exhibited similar age effect sizes in MTG (Supplementary Tables 1), confirming hypermethylation of *FHL2* gene is not mainly accompanied by cell divisions. Taken together, we speculate the smaller number of age-associated CpG sites found in cerebellum is largely attributed to its extremely low average cell replication rates.

It is easy to understand DNAm age comparisons between samples from the same tissues, i.e. we are confident that sample A is biologically younger than sample B when the DNAm age of sample A is much smaller than sample B and they are from the same tissue. However, we still lack sufficient evidence to compare the biological ages of samples from different tissues confidently. For example, as recently reported by Jonkman and colleagues, Horvath's multi-tissue clock predicts naive T cells to be up to 30 years younger than activated T cells from the same donor [181]. Can we conclude that naive T cells are biologically 30 years younger than activated T cells? Similarly, when predicted by our CerebellumClock$_{\text{common}}$, the non-cerebellar brain tissues are predicted to be at least 11 years older than the cerebellum (Figure 4.6a), however, we can not claim that those non-cerebellar brain tissues are biologically 11 years older than the cerebellum, as we could easily find one CpG or several CpGs combined that distinguishes the cerebellum from other brain tissue, then add it/them to the existing model and assigns it with a coefficient to counteract the 11 years gap. Then the new adjusted clock should not produce DNAm age predictions with systematic large differences between the cerebellum and other brain tissues. As proposed by Liu et al., the many non-age-related CpGs in Horvath's multi-tissue clock [40] may actually be reflecting and adjusting for tissue differences [174]. We have adequately demonstrated a single equation, BrainCortexClock, relying on only a subset of the 201 shared age-association CpGs provides unbiased DNAm age prediction for both cerebellum and cerebral cortex since given they have equal representation in the training dataset.

Another angle for ageing rate comparisons is to look at the Telomere Length (TL) shortening rates. Telomeres are protective DNA-protein complexes at the termini of chromosomes [182] and telomere attrition is considered an important hallmark of human ageing [32]. As comprehensively studied by Demanelis et al. [183], the average relative TL (RTL) varies across different tissue types, for instance, the average RTL is the lowest in whole blood and the longest in testis. Even though they found TL can shorten at different rates with ageing

between several tissue types, the majority of tissues do not show a significant difference in age-dependent shortening rates, and there is no evidence to claim that different tissue types age at rates proportional to their TL shortening rates.

We should acknowledge some limitations of this study. First, due to the scarcity of cerebellum samples, the majority of our collected cerebellum samples are from elderly individuals aged above 60 years old. It would be very valuable to test our hypothesis that the Horvath's multi-tissue clock would systematically overestimate the ages of cerebellum samples from young individuals aged below 30 years old. Second, our age EWASs on the cerebellum and MTG were also based on a very elderly population which has a relatively narrow age range, as demonstrated by Vershinina and colleagues [92], lots of age-associated CpGs do exhibit nonlinear methylation changes with age. Thus our age EWASs may have missed many CpGs that are strongly age-associated in the younger age group but be a much-attenuated association in the aged group. Future studies that include more young individuals should reveal a more complete picture of age-associated changes in the cerebellar methylome.

4.5 Conclusion

The large underestimations of age estimations for the cerebellum by widely used DNAm clocks are mainly due to inadequate cerebellum samples in their training datasets. We suggest the smaller number of age-associated CpG sites in cerebellum is largely attributed to its extremely low average cell replication rates. We have constructed a cerebellum-specific clock that can accurately predict cerebellum age and demonstrate conclusion from ageing rates comparison by DNA methylation clocks can be arbitrary by manipulating input CpG sites and the proportion of tissue types included in the training dataset. We believe our

findings can have wider implications for the use of ageing clocks.

Chapter5

Conclusion

In order to improve the precision of age prediction from methylation data, I collected a large number of DNA methylation microarray samples from public repositories. During the processing of those collected samples that are produced by different laboratories in different years, two new bioinformatic methods are proposed when realised the limitations of existing tools. The first one is a user-friendly sex classifier to accurately estimate sample sex from DNA methylation data. The second one is a novel two-step strategy to normalize DNA methylation microarray data avoiding sex bias. The two methods are useful in preprocessing methylation microarray samples, especially necessary when performing large cohort EWAS analysis. Lastly, in studying the epigenetic ageing rate differences between different tissues, the unique characteristics of age-associated methylome changes of the cerebellum are revealed, furthermore, I trained an accurate cerebellum age clock and suggest more evidence is required to support the claim that the cerebellum ages slower.

The sex classifier was built based on that females and males have different sex chromosome combinations, thus the overall methylation levels of the sex chromosomes between the two sexes are very different. By performing epigenome-wide sex association studies, 4047 CpGs on X chromosomes and 284 CpGs on Y chromosomes were found to be significantly differentially methylated between females and males. Then the sex classifier that consists of

two first principle components of the variance of methylation value of sex-associated CpGs on X chromosomes and Y chromosomes could clearly separate females and males. The sex classifier was demonstrated to perform well across a wide range of tissues or cell types despite it being originally built on methylation data from whole blood samples. It could be used to assign sex annotations for samples collected from public repositories that do not have such associations, and it also should be a common practice to compare the reported sex and the estimated sex from samples to identify questionable samples and then remove them from downstream analysis. Moreover, the sex classifier could identify samples with sex chromosome aneuploidy.

The two-step strategy provides an ideal solution to normalise female samples and male samples together without introducing technical sex bias. By this strategy, the first step is to normalize autosomal data separately by conventional normalization methods; then the second step is to infer the corrected values of sex chromosomes-linked CpGs by weighted averaging the normalised values of their nearest neighbours in the autosomes, this step is efficiently achieved by applying the interpolation algorithm. Furthermore, a useful parameter—the fraction of sex-explained variance, is proposed to be able to quantitatively measure a proportion of meaningful biological variance. It also demonstrated the beneficial effects of employing between-array normalisation methods to remove technical variance while retaining meaningful biological variance. The novel two-step strategy can be also employed by other quantile-based normalisation methods that are dedicated to dealing with other biological data such as RNA-seq data.

The claim that the cerebellum ages slowly was made upon a small subset of cerebellum samples and a single clock. To effectively verify this claim, a large set (N=752) of cerebellum samples from GEO was collected and their DNA methylation ages were estimated by six representative clocks. All cerebellum samples were severely underestimated compared to

their chronological ages by the six clocks, the underestimation effects were much more severe by clocks whose training datasets do not include brain cortex-related tissues, suggesting the observed huge underestimation effects mainly reflect the improper usage of the clocks. Comparative analysis of the epigenome-wide age association studies between the cerebellum and MTG from the same population reveals the cerebellum has a much smaller number of age-associated CpGs. And the age-associated CpGs between the two tissues are largely non-overlapped, indicating the unique age-related methylome changes of the cerebellum. A highly accurate age prediction model for the cerebellum is constructed demonstrating the strong and consistent ageing signals in the methylome of the cerebellum. Improper applying this cerebellum-specific model to other brain cortex tissues leads to systematical severe underestimations of their epigenetic ages compared to their chronological age. Clocks trained on only shared age-associated CpGs did show the cerebellum has smaller epigenetic ages, however, this is only valid on models constructed on a small proportion of CpGs (201), in which 94% (189) of them gain methylation with age, and cerebellum has lower median methylation levels for the majority of CpGs (85%, 171). To better perform ageing rate comparisons, future studies should look into other omic data as well, such as the transcriptome, proteome and metabolome.

Until now, there have been published many DNA methylation age clocks, nevertheless, only a very small proportion of, if not zero, CpGs are shared by any of the two clocks. For example, only 6 CpGs sites, i.e. cg09809672, cg04474832, cg22736354, cg06493994, cg19722847, cg05442902, exist in both Hannumn's clock and Horvath's multi-tissue clock. This seems to discourage these researchers who want to dissect the epigenetic clocks by diving into the individual clock-important CpGs. However, the small proportion of overlapping CpGs is not very unexpected considering the amount of CpGs whose methylation status changes with ageing are huge and they are widespread across all chromosomes. For example, Hannum et al. found 70,387 CpGs are age-associated among the 473,034 CpGs they examined. In gen-

eral, the observed age-associated methylation changes may derive from mitotic features, cell type composition changes, environmental influences and other disruptions that occur over time. Training of DNAm clocks often employs penalized linear regression, especially Elastic Net, to select a small number of CpGs among the huge amount of candidate age-associated CpGs. As many of the CpGs share similar changing patterns, any subtle difference in the training dataset or adjustments of the super-parameters of the training algorithms would result in very different outcome CpGs combinations even though the models have similar prediction performance. In this field, some researchers were very keen to discover biological meanings from the deviations between clock's estimated age and chronological age, but at present, the deviations from existing clocks mainly reflect technical variations, more future studies are needed to distil true, if exist, age acceleration effect from epigenetic data.

Bibliography

[1] Jean-Philippe Fortin et al. "Functional normalization of 450k methylation array data improves replication in large cancer studies". In: *Genome Biology* 15.11 (2014), p. 503. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0503-2.

[2] Ruth Pidsley et al. "A data-driven approach to preprocessing Illumina 450K methylation array data". In: *BMC genomics* 14.1 (2013), pp. 1–10.

[3] Conrad H Waddington. "The epigenotype". In: *Endeavour* 1 (1942), pp. 18–20.

[4] Eva Jablonka and Marion J Lamb. "The changing concept of epigenetics". In: *Annals of the New York Academy of Sciences* 981.1 (2002), pp. 82–96.

[5] Shelley L Berger et al. "An operational definition of epigenetics". In: *Genes & development* 23.7 (2009), pp. 781–783.

[6] Carrie Deans and Keith A Maggert. "What do you mean,"epigenetic"?" In: *Genetics* 199.4 (2015), pp. 887–896.

[7] Bradley E Bernstein, Alexander Meissner, and Eric S Lander. "The mammalian epigenome". In: *Cell* 128.4 (2007), pp. 669–681.

[8] Howard Cedar. "DNA methylation and gene activity." In: *Cell* 53.1 (1988), pp. 3–4.

[9] Rudolf W Hendriks et al. "The hypervariable DXS255 locus contains a LINE-1 repetitive element with a CpG island that is extensively methylated only on the active X chromosome". In: *Genomics* 14.3 (1992), pp. 598–603.

[10] Alysson R Muotri et al. "L1 retrotransposition in neurons is modulated by MeCP2". In: *Nature* 468.7322 (2010), pp. 443–446.

[11] T Mohandas, RS Sparkes, and LJ Shapiro. "Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation". In: *Science* 211.4480 (1981), pp. 393–396.

[12] En Li, Caroline Beard, and Rudolf Jaenisch. "Role for DNA methylation in genomic imprinting". In: *Nature* 366.6453 (1993), pp. 362–365.

[13] Aharon Razin and Arthur D Riggs. "DNA methylation and gene function". In: *Science* 210.4470 (1980), pp. 604–610.

[14] Treat B Johnson and Robert D Coghill. "Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus1". In: *Journal of the American Chemical Society* 47.11 (1925), pp. 2838–2844.

[15] Ryan Lister et al. "Human DNA methylomes at base resolution show widespread epigenomic differences". In: *nature* 462.7271 (2009), pp. 315–322.

[16] Adrian P Bird. "CpG islands as gene markers in the vertebrate nucleus". In: *Trends in Genetics* 3 (1987), pp. 342–347.

[17] Serge Saxonov, Paul Berg, and Douglas L Brutlag. "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters". In: *Proceedings of the National Academy of Sciences* 103.5 (2006), pp. 1412–1417.

[18] Alexander Meissner et al. "Genome-scale DNA methylation maps of pluripotent and differentiated cells". In: *Nature* 454.7205 (2008), pp. 766–770.

[19] Rafael A Irizarry et al. "The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores". In: *Nature genetics* 41.2 (2009), pp. 178–186.

[20] Marina Bibikova et al. "High density DNA methylation array with single CpG site resolution". In: *Genomics* 98.4 (2011), pp. 288–295.

[21] Juan Sandoval et al. "Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome". In: *Epigenetics* 6.6 (2011), pp. 692–702.

[22] Masaki Okano et al. "DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development". In: *Cell* 99.3 (1999), pp. 247–257. ISSN: 0092-8674. DOI: https://doi.org/10.1016/S0092-8674(00)81656-6.

[23] Shaoping Xie et al. "Cloning, expression and chromosome locations of the human DNMT3 gene family". In: *Gene* 236.1 (1999), pp. 87–95.

[24] Andrea Hermann, Rachna Goyal, and Albert Jeltsch. "The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites". In: *Journal of Biological Chemistry* 279.46 (2004), pp. 48350–48359.

[25] Xiaoji Wu and Yi Zhang. "TET-mediated active DNA demethylation: Mechanism, function and beyond". In: *Nature Reviews Genetics* 18.9 (2017), pp. 517–534. ISSN: 14710064.

[26] Sonja Zeilinger et al. "Tobacco smoking leads to extensive genome-wide changes in DNA methylation". In: *PLoS ONE* 8.5 (2013), pp. 1–14.

[27] Keith D Robertson. "DNA methylation and human disease". In: *Nature Reviews Genetics* 6.8 (2005), pp. 597–610.

[28] Manabu Fuchikami et al. "DNA methylation profiles of the Brain-Derived Neurotrophic Factor (BDNF) gene as apotent diagnostic biomarker in major depression". In: *PLoS ONE* 6.8 (2011), pp. 1–7.

[29] Sergio Villicaña and Jordana T Bell. "Genetic impacts on DNA methylation: research findings and future perspectives". In: *Genome Biology* 22 (2021), p. 127. DOI: 10.1186/s13059-021-02347-6.

[30] Sebastian Moran, Carles Arribas, and Manel Esteller. "Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences". In: *Epigenomics* 8.3 (2016), pp. 389–399. ISSN: 1750192X. DOI: 10.2217/epi.15.114.

[31] *Ageing and health*. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health. Accessed: 2022-08-22.

[32] Carlos López-Otín et al. "The hallmarks of aging". In: *Cell* 153.6 (2013), pp. 1194–1217.

[33] Sangita Pal and Jessica K Tyler. "Epigenetics and aging". In: *Science advances* 2.7 (2016), e1600584.

[34] John C Newman et al. "Strategies and challenges in clinical trials targeting human aging". In: *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 71.11 (2016), pp. 1424–1434.

[35] Linda Partridge, Joris Deelen, and P Eline Slagboom. "Facing up to the global challenges of ageing". In: *Nature* 561.7721 (2018), pp. 45–56.

[36] Johannes Leth Nielsen, Daniela Bakula, and Morten Scheibye-Knudsen. "Clinical trials targeting aging". In: *Frontiers in Aging* 3 (2022).

[37] Petr Klemera and Stanislav Doubal. "A new approach to the concept and computation of biological age". In: *Mechanisms of ageing and development* 127.3 (2006), pp. 240–248.

[38] Dóra Révész et al. "Telomere length as a marker of cellular aging is associated with prevalence and progression of metabolic syndrome". In: *The Journal of Clinical Endocrinology & Metabolism* 99.12 (2014), pp. 4607–4615.

[39] Gregory Hannum et al. "Genome-wide methylation profiles reveal quantitative views of human aging rates". In: *Molecular cell* 49.2 (2013), pp. 359–367.

[40] Steve Horvath. "DNA methylation age of human tissues and cell types". In: *Genome biology* 14.10 (2013), pp. 1–20.

[41] Jason G Fleischer et al. "Predicting age from the transcriptome of human dermal fibroblasts". In: *Genome biology* 19.1 (2018), pp. 1–8.

[42] Maxim N Shokhirev and Adiv A Johnson. "Modeling the human aging transcriptome across tissues, health status, and sex". In: *Aging cell* 20.1 (2021), e13280.

[43] Oliver Robinson et al. "Determinants of accelerated metabolomic and epigenetic aging in a UK cohort". In: *Aging cell* 19.6 (2020), e13149.

[44] Johannes Hertel et al. "Measuring biological age via metabonomics: the metabolic age score". In: *Journal of proteome research* 15.2 (2016), pp. 400–410.

[45] Fedor Galkin et al. "Biohorology and biomarkers of aging: Current state-of-the-art, challenges and opportunities". In: *Ageing Research Reviews* 60 (2020), p. 101050. ISSN: 1568-1637. DOI: https://doi.org/10.1016/j.arr.2020.101050.

[46] Sofie Bekaert, Tim De Meyer, and Patrick Van Oostveldt. "Telomere attrition as ageing biomarker". In: *Anticancer research* 25.4 (2005), pp. 3011–3021.

[47] Ann M Valdes et al. "Obesity, cigarette smoking, and telomere length in women". In: *The lancet* 366.9486 (2005), pp. 662–664.

[48] Marjolein J Peters et al. "The transcriptional landscape of age in human peripheral blood". In: *Nature communications* 6.1 (2015), pp. 1–14.

[49] Sarah E Harris et al. "Age-related gene expression changes, and transcriptome wide association study of physical and cognitive aging traits, in the Lothian Birth Cohort 1936". In: *Aging (Albany NY)* 9.12 (2017), p. 2489.

[50] Polina Mamoshina et al. "Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification". In: *Frontiers in genetics* 9 (2018), p. 242.

[51] Juulia Jylhävä, Nancy L Pedersen, and Sara Hägg. "Biological age predictors". In: *EBioMedicine* 21 (2017), pp. 29–36.

[52] Robin Holliday. "The inheritance of epigenetic defects". In: *Science* 238.4824 (1987), pp. 163–170.

[53] BF Vanyushin, SG Tkacheva, and AN Belozersky. "Rare bases in animal DNA". In: *Nature* 225.5236 (1970), pp. 948–949.

[54] BF Vanyushin et al. "The 5-methylcytosine in DNA of rats". In: *Gerontology* 19.3 (1973), pp. 138–152.

[55] Vincent L Wilson et al. "Genomic 5-methyldeoxycytidine decreases with age." In: *Journal of Biological Chemistry* 262.21 (1987), pp. 9948–9951.

[56] Joseph Golbus, Thomas D Palella, and Bruce C Richardson. "Quantitative changes in T cell DNA methylation occur during differentiation and ageing". In: *European journal of immunology* 20.8 (1990), pp. 1869–1872.

[57] Reid S Alisch et al. "Age-associated DNA methylation in pediatric populations". In: *Genome research* 22.4 (2012), pp. 623–632.

[58] Hunter L Porter et al. "Many chronological aging clocks can be found throughout the epigenome: Implications for quantifying biological aging". In: *Aging Cell* 20.11 (2021), e13492.

[59] Meaghan J Jones, Sarah J Goodman, and Michael S Kobor. "DNA methylation and healthy human aging". In: *Aging cell* 14.6 (2015), pp. 924–932.

[60] Roderick C Slieker et al. "Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception". In: *Epigenetics & chromatin* 11.1 (2018), pp. 1–11.

[61] Paolo Garagnani et al. "Methylation of ELOVL 2 gene as a new epigenetic marker of age". In: *Aging cell* 11.6 (2012), pp. 1132–1134.

[62] Lindsay M Reynolds et al. "Age-related variations in the methylome associated with gene expression in human monocytes and T cells". In: *Nature communications* 5.1 (2014), pp. 1–8.

[63] Liina Tserel et al. "CpG sites associated with NRP1, NRXN2 and miR-29b-2 are hypomethylated in monocytes during ageing". In: *Immunity & Ageing* 11.1 (2014), pp. 1–5.

[64] Sven Bocklandt et al. "Epigenetic predictor of age". In: *PloS one* 6.6 (2011), e14821.

[65] Carola I. Weidner et al. "Aging of blood can be tracked by DNA methylation changes at just three CpG sites". In: *Genome Biology* 15.2 (2014). ISSN: 1474760X. DOI: 10.1186/gb-2014-15-2-r24.

[66] Steve Horvath et al. "Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies". In: *Aging* 10.7 (2018), pp. 1758–1775. ISSN: 19454589. DOI: 10.18632/aging.101508.

[67] Steve Horvath and Kenneth Raj. "DNA methylation-based biomarkers and the epigenetic clock theory of ageing". In: *Nature Reviews Genetics* 19.6 (2018), pp. 371–384.

[68] Sarah Voisin et al. "An epigenetic clock for human skeletal muscle". In: *Journal of cachexia, sarcopenia and muscle* 11.4 (2020), pp. 887–898.

[69] Lisa M McEwen et al. "The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells". In: *Proceedings of the National Academy of Sciences* 117.38 (2020), pp. 23329–23335.

[70] Gemma L Shireby et al. "Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex". In: *Brain* 143.12 (2020), pp. 3763–3775.

[71] Mariana Boroni et al. "Highly accurate skin-specific methylome analysis algorithm as a platform to screen and validate therapeutics for healthy aging". In: *Clinical epigenetics* 12.1 (2020), pp. 1–16.

[72] Steve Horvath et al. "Obesity accelerates epigenetic aging of human liver". In: *Proceedings of the National Academy of Sciences* 111.43 (2014), pp. 15538–15543.

[73] Steve Horvath and Andrew J Levine. "HIV-1 infection accelerates age according to the epigenetic clock". In: *The Journal of infectious diseases* 212.10 (2015), pp. 1563–1573.

[74] Steve Horvath et al. "Accelerated epigenetic aging in Down syndrome". In: *Aging cell* 14.3 (2015), pp. 491–495.

[75] Steve Horvath et al. "Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels". In: *Aging (Albany NY)* 8.7 (2016), p. 1485.

[76] Anna Maierhofer et al. "Accelerated epigenetic aging in Werner syndrome". In: *Aging (Albany NY)* 9.4 (2017), p. 1143.

[77] Riccardo E Marioni et al. "DNA methylation age of blood predicts all-cause mortality in later life". In: *Genome biology* 16.1 (2015), pp. 1–12.

[78] Qian Zhang et al. "Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing". In: *Genome medicine* 11.1 (2019), pp. 1–11.

[79] Valentin M Vetter et al. "Relationship Between 5 Epigenetic Clocks, Telomere Length, and Functional Capacity Assessed in Older Adults: Cross-Sectional and Longitudinal Analyses". In: *The Journals of Gerontology: Series A* 77.9 (2022), pp. 1724–1733.

[80] R Hochschild. "Validating biomarkers of aging—mathematical approaches and results of a 2462-person study". In: *Practical Handbook of Human Biologic Age Determination; CRC Press: Boca Raton, FL, USA* (1994), pp. 93–144.

[81] Morgan E Levine et al. "An epigenetic biomarker of aging for lifespan and healthspan". In: *Aging (Albany NY)* 10.4 (2018), p. 573.

[82] Ake T Lu et al. "DNA methylation GrimAge strongly predicts lifespan and healthspan". In: *Aging (Albany NY)* 11.2 (2019), p. 303.

[83] Xia Li et al. "Longitudinal trajectories, correlations and mortality associations of nine biological ages across 20-years follow-up". In: *Elife* 9 (2020), e51507.

[84] Cathal McCrory et al. "GrimAge outperforms other epigenetic clocks in the prediction of age-related clinical phenotypes and all-cause mortality". In: *The Journals of Gerontology: Series A* 76.5 (2021), pp. 741–749.

[85] Christopher G Bell et al. "DNA methylation aging clocks: challenges and recommendations." In: *Genome biology* 20.1 (2019), p. 249. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1824-y.

[86] Daniel J Simpson and Tamir Chandra. "Epigenetic age prediction". In: *Aging Cell* 20.9 (2021), e13452.

[87] Lisa M McEwen et al. "Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array". In: *Clinical epigenetics* 10.1 (2018), pp. 1–9.

[88] Karen Sugden et al. "Patterns of reliability: assessing the reproducibility and integrity of DNA methylation measurement". In: *Patterns* 1.2 (2020), p. 100014.

[89] Albert T Higgins-Chen et al. "A computational solution for bolstering reliability of epigenetic clocks: implications for clinical trials and longitudinal tracking". In: *Nature Aging* 2.7 (2022), pp. 644–661. ISSN: 2662-8465.

[90] Maitreyee Bose et al. "Evaluation of microarray-based DNA methylation measurement using technical replicates: the Atherosclerosis Risk In Communities (ARIC) Study". In: *BMC bioinformatics* 15.1 (2014), pp. 1–10.

[91] Mark W Logue et al. "The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples". In: *Epigenomics* 9.11 (2017), pp. 1363–1371.

[92] O Vershinina et al. "Disentangling age-dependent DNA methylation: deterministic, stochastic, and nonlinear". In: *Scientific reports* 11.1 (2021), pp. 1–12.

[93] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[94] Yan Xia et al. "Sex-differential DNA methylation and associated regulation networks in human brain implicated in the sex-biased risks of psychiatric disorders". In: *Molecular psychiatry* 26.3 (2021), pp. 835–848.

[95] Robin JG Hartman, Sarah E Huisman, and Hester M den Ruijter. "Sex differences in cardiovascular epigenetics—a systematic review". In: *Biology of sex Differences* 9.1 (2018), pp. 1–8.

[96] Olivia A Grant et al. "Characterising sex differences of autosomal DNA methylation in whole blood using the Illumina EPIC array". In: *Clinical epigenetics* 14.1 (2022), pp. 1–16.

[97] *Life expectancy of women vs life expectancy of men, 2021.* `https://ourworldindata.org/grapher/life-expectancy-of-women-vs-life-expectancy-of-women`. Accessed: 2012-10-21.

[98] Igor Yusipov et al. "Age-related DNA methylation changes are sex-specific: a comprehensive assessment". In: *Aging (Albany NY)* 12.23 (2020), p. 24057.

[99] Yucheng Wang et al. "DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy". In: *BMC Genomics* 22.1 (2021), p. 484. ISSN: 1471-2164.

[100] Vardhman K Rakyan et al. "Epigenome-wide association studies for common human diseases". In: *Nature Reviews Genetics* 12.8 (2011), pp. 529–541.

[101] Jingyu Liu et al. "A study of the influence of sex on genome wide methylation". In: *PLoS ONE* 5.4 (2010).

[102] Paul Yousefi et al. "Sex differences in DNA methylation assessed by 450K BeadChip in newborns". In: *BMC Genomics* 16.1 (2015), pp. 1–12.

[103] Lilah Toker, Min Feng, and Paul Pavlidis. "Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies". In: *F1000Research* 5 (2016), pp. 1–15.

[104] Jonathan A. Heiss and Allan C. Just. "Identifying mislabeled and contaminated DNA methylation microarray data: An extended quality control toolset with examples from GEO". In: *Clinical Epigenetics* 10.1 (2018), pp. 1–9.

[105] Nina S McCarthy et al. "Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns". In: *BMC Genomics* 15.1 (2014), p. 981. ISSN: 1471-2164.

[106] Martin J. Aryee et al. "Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays". In: *Bioinformatics* 30.10 (2014), pp. 1363–1369.

[107] Chol Hee Jung et al. "sEst: Accurate sex-estimation and abnormality detection in methylation microarray data". In: *International Journal of Molecular Sciences* 19.10 (2018).

[108] Eilis Hannon et al. "Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins". In: *PLoS genetics* 14.8 (2018), e1007544.

[109] Eilis Hannon et al. "Assessing the co-variability of DNA methylation across peripheral cells and tissues: implications for the interpretation of findings in epigenetic epidemiology". In: *bioRxiv* (2020).

[110] Lisa M Mcewen et al. "DNA methylation age estimator for pediatric buccal cells : The PedBE clock". In: *Proceedings of the National Academy of Sciences of the United States of America* (2019), pp. 1–7. DOI: 10.1073/pnas.1820843116.

[111] Shraddha Pai et al. "Differential methylation of enhancer at IGF2 is associated with abnormal dopamine synthesis in major psychosis". In: *Nature Communications* 10.1 (2019), p. 2046.

[112] Steve Horvath et al. "An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease". In: *Genome Biology* 17.1 (2016), pp. 0–22.

[113] Liguo Wang et al. "Alpha-1 antitrypsin deficiency liver disease, mutational homogeneity modulated by epigenetic heterogeneity with links to obesity". In: *Hepatology* 70.1 (2019), pp. 51–66.

[114] Samantha L Wilson et al. "Mining DNA methylation alterations towards a classification of placental pathologies". In: *Human molecular genetics* 27.1 (2018), pp. 135–146.

[115] Timothy G Jenkins et al. "Intra-sample heterogeneity of sperm DNA methylation". In: *MHR: Basic science of reproductive medicine* 21.4 (2015), pp. 313–319.

[116] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019.

[117] *The UK household longitudinal study*. URL: `https://www.understandingsociety.ac.uk/about/about-the-study`.

[118] E Riboli et al. "European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection". In: *Public Health Nutrition* 5.6b (2002), pp. 1113–1124.

[119] Taru Tukiainen et al. "Landscape of X chromosome inactivation across human tissues". In: *Nature* 550.7675 (2017), pp. 244–248.

[120] Daniel L McCartney et al. "Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip". In: *Genomics data* 9 (2016), pp. 22–24.

[121] Hervé Abdi and Lynne J Williams. "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[122] Xianglong Zhang et al. "Integrated functional genomic analyses of Klinefelter and Turner syndromes reveal global network effects of altered X chromosome dosage". In: *Proceedings of the National Academy of Sciences of the United States of America* 117.9 (2020), pp. 4864–4873.

[123] Ricky S S. Joshi et al. "DNA methylation profiling of uniparental disomy subjects provides a map of parental epigenetic bias in the human genome". In: *American Journal of Human Genetics* 99.3 (2016), pp. 555–566.

[124] Richard F. Walker et al. "Epigenetic age analysis of children who seem to evade aging". In: *Aging* 7.5 (2015), pp. 334–339.

[125] Joana Viana et al. "Epigenomic and transcriptomic signatures of a Klinefelter syndrome (47,XXY) karyotype in the brain". In: *Epigenetics* 9.4 (2014), pp. 587–599.

[126] Martin Cederlöf et al. "Klinefelter syndrome and risk of psychosis, autism and ADHD". In: *Journal of Psychiatric Research* 48.1 (2014), pp. 128–130. ISSN: 0022-3956.

[127] Anthony S. Zannas et al. "Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-κB-driven inflammation and cardiovascular risk". In: *Proceedings of the National Academy of Sciences of the United States of America* 166.23 (2019), pp. 11370–11379.

[128] Keely L Szilágyi et al. “Epigenetic contribution of the myosin light chain kinase gene to the risk for acute respiratory distress syndrome”. In: *Translational Research* 180 (2017), pp. 12–21.

[129] Monica Uddin et al. “Epigenetic meta-analysis across three civilian cohorts identifies NRG1 and HGS as blood-based biomarkers for post-traumatic stress disorder”. In: *Epigenomics* 10.12 (2018), pp. 1585–1601.

[130] Benjamin Lehne et al. “A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies”. In: *Genome Biology* 16.1 (2015), pp. 1–12.

[131] NT Ventham et al. “Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease”. In: *Nature communications* 7.1 (2016), pp. 1–14.

[132] Behrooz Torabi Moghadam et al. “Analyzing DNA methylation patterns in subjects diagnosed with schizophrenia using machine learning methods”. In: *Journal of Psychiatric Research* 114 (2019), pp. 41–47.

[133] L F Wockner et al. “Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients”. In: *Translational Psychiatry* 4.1 (2014), e339–e339.

[134] MARY F LYON. “Gene action in the X-chromosome of the mouse (Mus musculus L.)” In: *Nature* 190.4773 (1961), pp. 372–373.

[135] Andrew J. Sharp et al. “DNA methylation profiles of human active and inactive X chromosomes”. In: *Genome Research* 21.10 (2011), pp. 1592–1600.

[136] Andrew E. Jaffe and Rafael A. Irizarry. “Accounting for cellular heterogeneity is critical in epigenome-wide association studies”. In: *Genome Biology* 15.2 (2014), pp. 1–9.

[137] Eugene A. Houseman et al. “DNA methylation arrays as surrogate measures of cell mixture distribution”. In: *BMC Bioinformatics* 13.1 (2012).

[138] Graham J Burton and Abigail L Fowden. “The placenta: a multifaceted, transient organ”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1663 (2015), p. 20140066.

[139] Hongshan Guo et al. “The DNA methylation landscape of human early embryos”. In: *Nature* 511.7511 (2014), pp. 606–610.

[140] Yucheng Wang et al. “interpolatedXY: a two-step strategy to normalise DNA methylation microarray data avoiding sex bias”. In: *bioRxiv* (2021).

[141] Sarah Dedeurwaerder et al. "Evaluation of the Infinium Methylation 450K technology". In: *Epigenomics* 3.6 (2011), pp. 771–784. ISSN: 17501911. DOI: 10.2217/epi.11.105.

[142] Andrew E. Teschendorff et al. "A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data". In: *Bioinformatics* 29.2 (2013), pp. 189–196. ISSN: 13674803.

[143] Timothy J. Triche et al. "Low-level processing of Illumina Infinium DNA Methylation BeadArrays". In: *Nucleic Acids Research* 41.7 (2013), pp. 1–11. ISSN: 03051048. DOI: 10.1093/nar/gkt090.

[144] Allison M. Cotton et al. "Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation". In: *Human Molecular Genetics* 24.6 (2015), pp. 1528–1539. ISSN: 14602083. DOI: 10.1093/hmg/ddu564.

[145] Tyler J. Gorrie-Stone et al. "Bigmelon: Tools for analysing large DNA methylation datasets". In: *Bioinformatics* 35.6 (2019), pp. 981–986. ISSN: 14602059.

[146] Randi K Johnson et al. "Longitudinal DNA methylation differences precede type 1 diabetes". In: *Scientific Reports* 10.1 (2020), p. 3721. ISSN: 2045-2322. DOI: 10.1038/s41598-020-60758-0.

[147] Jean-Philippe Fortin, Jr Triche Timothy J., and Kasper D Hansen. "Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi". In: *Bioinformatics* 33.4 (2016), pp. 558–560. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw691.

[148] Jenny Van Dongen et al. "Genetic and environmental influences interact with age and sex in shaping the human methylome". In: *Nature communications* 7.1 (2016), pp. 1–13.

[149] Yuan Tian et al. "ChAMP: Updated methylation analysis pipeline for Illumina BeadChips". In: *Bioinformatics* 33.24 (2017), pp. 3982–3984. ISSN: 14602059.

[150] Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. "SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips." In: *Genome biology* 13.6 (2012), pp. 1–12. ISSN: 14656914. DOI: 10.1186/gb-2012-13-6-r44.

[151] Sarah Dedeurwaerder et al. "A comprehensive overview of Infinium Human Methylation450 data processing". In: *Briefings in Bioinformatics* 15.6 (2013), pp. 929–941. ISSN: 14774054. DOI: 10.1093/bib/bbt054.

[152] Yucheng Wang et al. "Recalibrating the cerebellum DNA methylation clock: implications for ageing rates comparison". In: *bioRxiv* (2022), pp. 2022–05.

[153] Steve Horvath et al. "The cerebellum ages slowly according to the epigenetic clock". In: *Aging (Albany NY)* 7.5 (2015), p. 294.

[154] Mary E Sehl et al. "DNA methylation age is elevated in breast tissue of healthy women". In: *Breast cancer research and treatment* 164.1 (2017), pp. 209–219.

[155] Steve Horvath and Kenneth Raj. "DNA methylation-based biomarkers and the epigenetic clock theory of ageing". In: *Nature Reviews Genetics* 19.6 (2018), pp. 371–384.

[156] Christopher G Bell et al. "DNA methylation aging clocks: challenges and recommendations". In: *Genome biology* 20.1 (2019), pp. 1–24.

[157] Louis Y El Khoury et al. "Systematic underestimation of the epigenetic clock and age acceleration in older subjects". In: *Genome biology* 20.1 (2019), pp. 1–10.

[158] Danielle L Brokaw et al. "Cell death and survival pathways in Alzheimer's disease: an integrative hypothesis testing approach utilizing-omic data sets". In: *Neurobiology of Aging* 95 (2020), pp. 15–25.

[159] Katie Lunnon et al. "Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease". In: *Nature neuroscience* 17.9 (2014), pp. 1164–1170.

[160] Adam R Smith et al. "Parallel profiling of DNA methylation and hydroxymethylation highlights neuropathology-associated epigenetic variation in Alzheimer's disease". In: *Clinical epigenetics* 11.1 (2019), pp. 1–13.

[161] Stephen A Semick et al. "Integrated DNA methylation and gene expression profiling across multiple brain regions implicate novel genes in Alzheimer's disease". In: *Acta neuropathologica* 137.4 (2019), pp. 557–569.

[162] Ruth Pidsley et al. "Methylomic profiling of human brain tissue supports a neurodevelopmental origin for schizophrenia". In: *Genome biology* 15.10 (2014), pp. 1–11.

[163] Andrew E Jaffe et al. "Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex". In: *Nature neuroscience* 19.1 (2016), pp. 40–47.

[164] Rebecca G Smith et al. "Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology". In: *Alzheimer's & Dementia* 14.12 (2018), pp. 1580–1588.

[165] Tyler J Gorrie-Stone et al. "Bigmelon: tools for analysing large DNA methylation datasets". In: *Bioinformatics* 35.6 (2019), pp. 981–986.

[166] Yucheng Wang et al. "DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy". In: *BMC genomics* 22.1 (2021), pp. 1–11.

[167] Yucheng Wang. *dnaMethyAge: a user friendly R package to predict epigenetic age and calculate age acceleration from DNA methylation data*. 2021. URL: `https://github.com/yiluyucheng/dnaMethyAge`.

[168] Yi-an Chen et al. "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray". In: *Epigenetics* 8.2 (2013), pp. 203–209.

[169] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

[170] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1 (2010), p. 1.

[171] Belinda Phipson, Jovana Maksimovic, and Alicia Oshlack. "missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform". In: *Bioinformatics* 32.2 (2016), pp. 286–288.

[172] Daniel M Wolpert, R Chris Miall, and Mitsuo Kawato. "Internal models in the cerebellum". In: *Trends in cognitive sciences* 2.9 (1998), pp. 338–347.

[173] Gary L Wenk et al. "Neuropathologic changes in Alzheimer's disease". In: *Journal of Clinical Psychiatry* 64 (2003), pp. 7–10.

[174] Zuyun Liu et al. "Underlying features of epigenetic aging clocks in vivo and in vitro". In: *Aging Cell* 19.10 (2020), e13229.

[175] Frederico AC Azevedo et al. "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain". In: *Journal of Comparative Neurology* 513.5 (2009), pp. 532–541.

[176] Kirsty L Spalding et al. "Retrospective birth dating of cells in humans". In: *Cell* 122.1 (2005), pp. 133–143.

[177] Maria Giulia Bacalini et al. "Systemic age-associated DNA hypermethylation of ELOVL2 gene: in vivo and in vitro evidences of a cell replication process". In: *Journals of*

Gerontology Series A: Biomedical Sciences and Medical Sciences 72.8 (2017), pp. 1015–1023.

[178] Åsa Johansson, Stefan Enroth, and Ulf Gyllensten. "Continuous aging of the human DNA methylome throughout the human lifespan". In: *PloS one* 8.6 (2013), e67378.

[179] Renata Zbieć-Piekarska et al. "Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science". In: *Forensic Science International: Genetics* 14 (2015), pp. 161–167.

[180] Ines Florath et al. "Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites". In: *Human molecular genetics* 23.5 (2014), pp. 1186–1201.

[181] Thomas H Jonkman et al. "Functional genomics analysis identifies T and NK cell activation as a driver of epigenetic clock progression". In: *Genome Biology* 23.1 (2022), p. 24. ISSN: 1474-760X.

[182] Elizabeth H Blackburn, Elissa S Epel, and Jue Lin. "Human telomere biology: a contributory and interactive factor in aging, disease risks, and protection". In: *Science* 350.6265 (2015), pp. 1193–1198.

[183] Kathryn Demanelis et al. "Determinants of telomere length across human tissues". In: *Science* 369.6509 (2020), eaaz6876.

AppendixA

Appendix Codes

1. The original codes for adjustedDasen:

```
#'
#' adjustedDasen
#'
#' @description adjustedDasen utilizes dasen normliasation to
   normalise autosomal
#' CpGs, and infers the sex chromosome linked CpGs by linear
   interpolation on
#' corrected autosomal CpGs.
#'
#' @param mns matrix of methylated signal intensities, samples in
   column and
#' probes in row.
#' @param uns matrix of unmethylated signal intensities, samples in
   column and
#' probes in row.
#' @param onetwo character vector or factor of length nrow(mns)
   indicating assay
#'  type 'I' or 'II'.
```

```
14 #' @param chr character vector stores the mapped chromosomes for
       all probes, e.g.
15 #' chr <- c('1', 'X', '21', ..., 'Y').
16 #' @param offset_fit logical (default is TRUE). To use dasen, set
       it TRUE; to use
17 #' nasen, set it FALSE.
18 #' @param cores an integer(e.g. 8) defines the number of cores to
       parallel processing.
19 #' Default value is 1, set to -1 to use all available cores.
20 #' @param ret2 logical (default is FALSE), if TRUE, returns a list
       of intensities
21 #' and betas instead of a naked matrix of betas.
22 #' @param fudge default 100, a value added to total intensity to
       prevent denominators
23 #' close to zero when calculating betas, e.g. betas <- mns / (mns +
       uns + fudge).
24 #' @param ... additional argument roco for dfsfit giving Sentrix
       rows and
25 #' columns.  This allows a background gradient model to be fit.
       This is split
26 #' from data column names by default.  roco=NULL disables model
       fitting (and
27 #' speeds up processing), otherwise roco can be supplied as a
       character vector
28 #' of strings like 'R01C01' (only 3rd and 6th characters used).
29 #'
30 #' @return a matrix of normalised beta values.
31 #' @export
32 #' adjustedDasen
```

```
33 #'
34 #' @references
35 #' A data-driven approach to preprocessing Illumina 450K
      methylation array data,
36 #' Pidsley et al, BMC Genomics. \cr
37 #' interpolatedXY: a two-step strategy to normalise DNA methylation
38 #' microarray data avoiding sex bias, Wang et al., 2021.
39 #' @examples
40 #' data(melon)
41 #' normalised_betas <- adjustedDasen(mns = methylated(melon), uns =
      unmethylated(melon), onetwo = fData(melon)[,fot(melon)], chr =
      fData(melon)$CHR, cores=1)
42 #' ## if input is an object of methylumiset or methylset
43 #' normalised_betas <- adjustedDasen(melon)
44 #'
45 adjustedDasen <- function(mns, uns, onetwo, chr, offset_fit=TRUE,
      cores=1, ret2=FALSE, fudge=100,...){
46     stopifnot(nrow(mns) == length(chr))
47     stopifnot(nrow(uns) == length(chr))
48     stopifnot(nrow(mns) == length(onetwo))
49     stopifnot(nrow(uns) == length(onetwo))
50     stopifnot(length(chr) == length(onetwo))
51
52     if(!is.logical(chr)){
53         is_sex <- grepl('(X|chrX|Y|chrY|23|24)', as.character(chr))
54     } else {
55         is_sex <- chr
56     }
57
```

```
58      if (cores < 1) {
59          cores <- detectCores()
60      }
61
62      if(Sys.info()["sysname"] != "Linux"){
63          cores <- 1
64      }
65
66      ## to use 'nasen', set offset_fit=FALSE
67      if(offset_fit){
68          mns <- p_dfsfit(mns,  onetwo, cores=cores)
69          uns <- p_dfsfit(uns,  onetwo, roco=NULL, cores=cores)
70      }
71
72      mns[onetwo == 'I' , ] <- uSexQNengine(A = mns[onetwo == 'I' ,
            ], is_sex = is_sex[onetwo == 'I' ], cores = cores)
73      mns[onetwo == 'II', ] <- uSexQNengine(A = mns[onetwo == 'II',
            ], is_sex = is_sex[onetwo == 'II'], cores = cores)
74      uns[onetwo == 'I' , ] <- uSexQNengine(A = uns[onetwo == 'I' ,
            ], is_sex = is_sex[onetwo == 'I' ], cores = cores)
75      uns[onetwo == 'II', ] <- uSexQNengine(A = uns[onetwo == 'II',
            ], is_sex = is_sex[onetwo == 'II'], cores = cores)
76
77      betas = (mns) / (mns+uns+fudge)
78      if(ret2){
79          return(list(betas = betas, methylated = mns, unmethylated =
                uns))
80      } else {
81          return(betas)
```

```
82      }
83  }
84
85
86  sort_order <- function(d, tie=TRUE){
87      ## obtain the sorted values and their index
88      Si <- sort(d, method = "quick", index.return = TRUE) # NA will
            be ignored or removed.
89      if (tie){
90          Si$ix <- NA
91      }
92      # deal with NA in input d
93      nobsj <- length(Si$x)
94      n_1 <- length(d)
95      isna <- is.na(d)
96      if (sum(isna) > 0) {
97          i <- (0:(n_1 - 1))/(n_1 - 1)
98          Si$x <- approx((0:(nobsj - 1))/(nobsj - 1), Si$x, i, ties =
              list("ordered", mean))$y  # Si$x will not contain NAs
              any more.
99          if (!tie) {
100             O_i <- rep(NA, n_1)
101             O_i[!isna] <- ((1:n_1)[!isna])[Si$ix]
102             Si$ix <- O_i
103         }
104     }
105     return(Si)
106 }
107
```

```
108
109 tie_norm <- function(d, is_sex, rank2mean){
110     ## normalise d differently on autosomes and XY
111     d_sex <- d[is_sex]
112     d_autosome <- d[!is_sex]
113     r_autosome <- rank(d_autosome) # NA will be counted and placed
            at the end.
114     # Get the ranks of sexual cpgs based on ranks of  autosomal
            cpgs;
115     # rule=2 means the value at the closest data extreme is used
            when new x is greater than max(x)
116     r_sex <- approx(d_autosome, r_autosome, d_sex, ties = mean,
            rule=2)$y
117
118     # Produce the final values of non-NA  autosomal cpgs based on
            their ranks
119     notna <- !is.na(d_autosome)
120     nobsj <- sum(notna)
121     d[!is_sex][notna] <- rank2mean((r_autosome[notna] - 1)/(nobsj -
            1))
122     # Produce the final values of non-NA sexual cpgs based on their
            ranks
123     notna_sex <- !is.na(d_sex)
124     d[is_sex][notna_sex] <- rank2mean((r_sex[notna_sex] - 1)/(nobsj
            - 1))
125     return(d)
126 }
127
128
```

```
129 uSexQNengine <- function(A, is_sex, cores=1) {
130     ## A: a dataframe or matrix;
131     ## chr: a vector, like c('1', '2', 'X', 'Y')
132     stopifnot(nrow(A) == length(is_sex))
133     A <- data.frame(A, check.names=FALSE)
134     A_autosome <- A[!is_sex, ]
135     n_1 <- nrow(A_autosome)
136     sort_Aa <- mclapply(A_autosome, sort_order, mc.cores=cores,
            tie=TRUE)
137     S_autosome <- sapply(sort_Aa, function(x) x$x)
138     m_autosome <- rowMeans(S_autosome)
139
140     # Get a function which gives relationships between orders and
            mean values.
141     i <- (0:(n_1 - 1))/(n_1 - 1)
142     rank2mean <- approxfun(i, m_autosome, ties = list("ordered",
            mean))
143
144     #rm(S_autosome, A_autosome, sort_Aa)
145     # For each sample, find its normalised values
146     A <- mclapply(A, tie_norm, is_sex=is_sex, mc.cores=cores,
            rank2mean=rank2mean)
147     A <- sapply(A, function(x) x)
148
149     return(A)
150 }
151
152
153 p_dfsfit <- function (mn, onetwo, cores=1,
```

```
    roco=substring(colnames(mn), regexpr("R0[1-9]C0[1-9]",
    colnames(mn))), ...){
154     mn <- data.frame(mn, check.names=FALSE)
155     mdf <- mclapply(mn, dfs2, onetwo, mc.cores=cores)
156     mdf <- sapply(mdf, function(x) x)
157     if (!is.null(roco)) {
158         scol <- as.numeric(substr(roco, 6, 6))
159         srow <- as.numeric(substr(roco, 3, 3))
160         fit <- try(lm(mdf ~ srow + scol), silent = TRUE)
161         if (!inherits(fit, "try-error")) {
162             mdf <- fit$fitted.values
163         }
164         else {
165             message("Sentrix position model failed, skipping")
166         }
167     }
168     otcor <- matrix(rep(mdf, sum(onetwo == "I")), byrow = T, nrow =
        sum(onetwo == "I"))
169     mn[onetwo == "I", ] <- mn[onetwo == "I", ] - otcor
170     mn
171 }
```

2. The original codes for adjustedFunnorm:

```
1 ### ORIGINAL AUTHOR: Jean-Philippe Fortin, Sept 24 2013 (Functional
    normalization of 450k methylation array data improves
    replication in large cancer studies, Genome Biology, 2014)
2 ### Adopted by Yucheng Wang
3
```

```
4  ########################################################
5  ## Functional normalization of the 450k array
6  ## Jean-Philippe Fortin
7  ## Sept 24 2013
8  ########################################################
9
10 ##
11
12
13 #' adjustedFunnorm
14 #'
15 #' @description adjustedFunnorm utilizes functional normliasation
      to normalise autosomal
16 #' CpGs, and infers the sex chromosome linked CpGs by linear
      interpolation on
17 #' corrected autosomal CpGs.
18 #'
19 #' @param rgSet An object of class "RGChannelSet".
20 #' @param nPCs Number of principal components from the control
      probes PCA.
21 #' @param sex An optional numeric vector containing the sex of the
      samples.
22 #' @param bgCorr Should the NOOB background correction be done,
      prior to
23 #' functional normalization (see "preprocessNoob")
24 #' @param dyeCorr Should dye normalization be done as part of the
      NOOB
25 #' background correction (see "preprocessNoob")?
26 #' @param keepCN Should copy number estimates be kept around?
```

```
   Setting to 'FALSE'
#'  will decrease the size of the output object significantly.
#' @param ratioConvert Should we run "ratioConvert", ie. should the
   output be a
#' "GenomicRatioSet" or should it be kept as a "GenomicMethylSet";
   the latter
#' is for experts.
#' @param verbose Should the function be verbose?
#'
#' @return an object of class "GenomicRatioSet", unless
   "ratioConvert=FALSE" in
#' which case an object of class "GenomicMethylSet".
#' @export
#' adjustedFunnorm
#' @references
#' Functional normalization of 450k methylation array data improves
   replication
#' in large cancer studies, Fortin et al., 2014, Genome biology. \cr
#' interpolatedXY: a two-step strategy to normalise DNA methylation
#' microarray data avoiding sex bias, Wang et al., 2021.
#'
#' @examples
#' \dontrun{
#' GRset <- adjustedFunnorm(RGSet)
#' }
#'
adjustedFunnorm <- function(rgSet, nPCs=2, sex = NULL, bgCorr =
   TRUE, dyeCorr = TRUE, keepCN = TRUE, ratioConvert = TRUE,
   verbose = TRUE) {
```

```
49
50      .isMatrixBackedOrStop(rgSet, "adjustedFunnorm")
51
52      .isRGOrStop(rgSet)
53      rgSet <- updateObject(rgSet) ## FIXM: might not KDH:
            technically, this should not be needed, but might be nice
54
55      # Background correction and dye bias normalization:
56      if (bgCorr){
57          if(verbose && dyeCorr) {
58              message("[adjustedFunnorm] Background and dye bias
                    correction with noob")
59          } else {
60              message("[adjustedFunnorm] Background correction with
                    noob")
61          }
62          gmSet <- preprocessNoob(rgSet, dyeCorr = dyeCorr)
63          if(verbose) message("[adjustedFunnorm] Mapping to genome")
64          gmSet <- mapToGenome(gmSet)
65      } else {
66          if(verbose) message("[adjustedFunnorm] Mapping to genome")
67          gmSet <- mapToGenome(rgSet)
68      }
69
70      subverbose <- max(as.integer(verbose) - 1L, 0)
71
72      if(verbose) message("[adjustedFunnorm] Quantile extraction")
73      extractedData <- .extractFromRGSet450k(rgSet)
74      rm(rgSet)
```

```
75
76      if (is.null(sex)) {
77          gmSet <- addSex(gmSet, getSex(gmSet, cutoff = -3))
78          sex <- rep(1L, length(gmSet$predictedSex))
79          sex[gmSet$predictedSex == "F"] <- 2L
80      }
81      if(verbose) message("[adjustedFunnorm] Normalization")
82      if(keepCN) {
83          CN <- getCN(gmSet)
84      }
85      gmSet <- .adjusted_normalizeFunnorm450k(object = gmSet,
           extractedData = extractedData,
86                                                  sex = sex, nPCs = nPCs,
                                                       verbose = subverbose)
87      preprocessMethod <- c(preprocessMethod(gmSet),
88                            mu.norm = sprintf("Funnorm, nPCs=%s",
                                 nPCs))
89      if(ratioConvert) {
90          grSet <- ratioConvert(gmSet, type = "Illumina", keepCN =
               keepCN)
91          if(keepCN) {
92              assay(grSet, "CN") <- CN
93          }
94          grSet@preprocessMethod <- preprocessMethod
95          return(grSet)
96      } else {
97          gmSet@preprocessMethod <- preprocessMethod
98          return(gmSet)
99      }
```

```
}


.getFunnormIndices <- function(object) {
    ## WYC
    .isGenomicOrStop(object)
    probeType <- getProbeType(object, withColor = TRUE)
    # autosomal <- (seqnames(object) %in% paste0("chr", 1:22))
    indices <- list(IGrn = (probeType == "IGrn"),
                    IRed = (probeType == "IRed"),
                    II = (probeType == "II" ),
                    XY = as.vector(seqnames(object)) %in% c("chrX",
                       "chrY"))
    indices
}


.adjusted_normalizeFunnorm450k <- function(object, extractedData,
   nPCs, sex, verbose = TRUE) {
    #normalizeQuantiles <- function(matrix, indices, sex = NULL) {
    #    matrix <- matrix[indices,,drop=FALSE]
    #    ## uses probs, model.matrix, nPCS, through scoping)
    #    oldQuantiles <- t(colQuantiles(matrix, probs = probs))
    #    if(is.null(sex)) {
    #        newQuantiles <- .returnFit(controlMatrix =
       model.matrix, quantiles = oldQuantiles, nPCs = nPCs)
    #    } else {
    #        newQuantiles <- .returnFitBySex(controlMatrix =
       model.matrix, quantiles = oldQuantiles, nPCs = nPCs, sex =
```

```
    sex)
#     }
#     .normalizeMatrix(matrix, newQuantiles)
#}

interpolatedXY <- function(ra_signal, na_signal, ru_signal){
    # construct a function which reflects relationships between
      orders and final norm values.
    n_1 <- length(ra_signal)
    rank2mean <- approxfun((0:(n_1 - 1))/(n_1 - 1),
      sort(na_signal, method = "quick"), ties =
      list("ordered", mean))

    rank_autosome <- rank(ra_signal) # NA will be counted and
      placed at the end.
    # Get the ranks of sexual cpgs based on ranks of  autosomal
      cpgs;
    # rule=2 means the value at the closest data extreme is
      used when new x is greater than max(x)
    rank_sex <- approx(ra_signal, rank_autosome, ru_signal,
      ties = mean, rule=2)$y

    # Produce the final values of non-NA sexual cpgs based on
      their ranks
    notNA <- !is.na(ru_signal)
    ru_signal[notNA] <- rank2mean((rank_sex[notNA] - 1)/(n_1 -
      1))
    ru_signal
}
```

```
144
145     unbiased_normalizeQuantiles <- function(mat, indices, sex_probe
            = NULL) {
146         mat_auto <- mat[(indices & !sex_probe),,drop=FALSE]
147         mat_sex <- mat[(indices & sex_probe),,drop=FALSE]
148         ## uses probs, model.matrix, nPCS, through scoping)
149         oldQuantiles <- t(colQuantiles(mat_auto, probs = probs))
150         newQuantiles <- .returnFit(controlMatrix = model.matrix,
                quantiles = oldQuantiles, nPCs = nPCs)
151         n_matrix <- .normalizeMatrix(mat_auto, newQuantiles)
152         for(j in 1:ncol(n_matrix)){
153             mat_sex[, j] <- interpolatedXY(mat_auto[, j],
                    n_matrix[, j], mat_sex[, j])
154         }
155         mat[(indices & sex_probe), ] <- mat_sex
156         mat[(indices & !sex_probe), ] <- n_matrix
157         mat
158     }
159
160     indicesList <- .getFunnormIndices(object)
161     model.matrix <- .buildControlMatrix450k(extractedData)
162     probs <- seq(from = 0, to = 1, length.out = 500)
163     Meth <- getMeth(object)
164     Unmeth <- getUnmeth(object)
165     if (nPCs > 0){
166         for (type in c("IGrn", "IRed", "II")) {
167             indices <- indicesList[[type]]
168             if(length(indices) > 0) {
169                 if(verbose) message(sprintf("[InterpolatedXY
```

```
                    adjustedFunnorm] Normalization of the %s
                    probes", type))
170                 Unmeth <- unbiased_normalizeQuantiles(Unmeth,
                    indices=indices, sex_probe=indicesList$XY)
171                 Meth <- unbiased_normalizeQuantiles(Meth,
                    indices=indices, sex_probe=indicesList$XY)
172             }
173         }
174     }
175     assay(object, "Meth") <- Meth
176     assay(object, "Unmeth") <- Unmeth
177     return(object)
178 }
179
180
181 ### To extract quantiles and control probes from rgSet
182 .extractFromRGSet450k <- function(rgSet) {
183     rgSet <- updateObject(rgSet)
184     controlType <- c("BISULFITE CONVERSION I",
185                      "BISULFITE CONVERSION II",
186                      "EXTENSION",
187                      "HYBRIDIZATION",
188                      "NEGATIVE",
189                      "NON-POLYMORPHIC",
190                      "NORM_A",
191                      "NORM_C",
192                      "NORM_G",
193                      "NORM_T",
194                      "SPECIFICITY I",
```

```
                    "SPECIFICITY II",
                    "TARGET REMOVAL",
                    "STAINING")

    array <- annotation(rgSet)[["array"]]
    ## controlAddr <- getControlAddress(rgSet, controlType =
       controlType, asList = TRUE)
    ctrls <- getProbeInfo(rgSet, type = "Control")
    ctrls <- data.frame(ctrls)
    if(!all(controlType %in% ctrls$Type))
        stop("The 'rgSet' does not contain all necessary control
           probes")

    ctrlsList <- split(ctrls, ctrls$Type)[controlType]
    redControls <- getRed(rgSet)[ctrls$Address,,drop=FALSE]
    redControls <- lapply(ctrlsList, function(ctl)
       redControls[ctl$Address,,drop=FALSE])
    greenControls <- getGreen(rgSet)[ctrls$Address,,drop=FALSE]
    greenControls <- lapply(ctrlsList, function(ctl)
       greenControls[ctl$Address,,drop=FALSE])

    ## Extraction of the undefined negative control probes
    oobRaw <- getOOB(rgSet)
    probs <- c(0.01, 0.50, 0.99)
    greenOOB <- t(colQuantiles(oobRaw$Grn, na.rm = TRUE, probs =
       probs))
    redOOB   <- t(colQuantiles(oobRaw$Red, na.rm=TRUE,  probs =
       probs))
    oob      <- list(greenOOB = greenOOB, redOOB = redOOB)
```

```
218
219     return(list(
220         greenControls = greenControls,
221         redControls = redControls,
222         oob = oob, ctrlsList = ctrlsList,
223         array = array))
224 }
225
226
227 ## Extraction of the Control matrix
228 .buildControlMatrix450k <- function(extractedData) {
229     getCtrlsAddr <- function(exType, index) {
230         ctrls <- ctrlsList[[index]]
231         addr <- ctrls$Address
232         names(addr) <- ctrls$ExtendedType
233         na.omit(addr[exType])
234     }
235
236     array <- extractedData$array
237     greenControls <- extractedData$greenControls
238     redControls <- extractedData$redControls
239     controlNames <- names(greenControls)
240     ctrlsList <- extractedData$ctrlsList
241
242     ## Bisulfite conversion extraction for probe type II:
243     index <- match("BISULFITE CONVERSION II", controlNames)
244     redControls.current <- redControls[[ index ]]
245     bisulfite2 <- colMeans2(redControls.current, na.rm = TRUE)
246
```

```
247     ## Bisulfite conversion extraction for probe type I:
248     index <- match("BISULFITE CONVERSION I", controlNames)
249     if (array=="IlluminaHumanMethylation450k"){
250         addr <- getCtrlsAddr(exType = sprintf("BS Conversion
            I%sC%s", c(" ", "-", "-"), 1:3), index = index)
251     } else {
252         addr <- getCtrlsAddr(exType = sprintf("BS Conversion
            I%sC%s", c("-", "-"), 1:2), index = index)
253     }
254     greenControls.current <- greenControls[[ index
        ]][addr,,drop=FALSE]
255     if (array=="IlluminaHumanMethylation450k"){
256         addr <- getCtrlsAddr(exType = sprintf("BS Conversion
            I-C%s", 4:6), index = index)
257     } else {
258         addr <- getCtrlsAddr(exType = sprintf("BS Conversion
            I-C%s", 3:5), index = index)
259     }
260     redControls.current <- redControls[[ index ]][addr,, drop=FALSE]
261     if (nrow(redControls.current)==nrow(greenControls.current)){
262         bisulfite1 <- colMeans2(redControls.current +
            greenControls.current, na.rm = TRUE)
263     } else {
264         bisulfite1 <- colMeans2(redControls.current, na.rm=TRUE) +
            colMeans2(greenControls.current, na.rm = TRUE)
265     }
266
267
268     ## Staining
```

```
269     index <- match("STAINING", controlNames)
270     addr <- getCtrlsAddr(exType = "Biotin (High)", index = index)
271     stain.green <- t(greenControls[[ index ]][addr,,drop=FALSE])
272     addr <- getCtrlsAddr(exType = "DNP (High)", index = index)
273     stain.red <- t(redControls[[ index ]][addr,, drop=FALSE ])
274
275     ## Extension
276     index <-    match("EXTENSION", controlNames)
277     addr <- getCtrlsAddr(exType = sprintf("Extension (%s)", c("A",
            "T")), index = index)
278     extension.red <- t(redControls[[index]][addr,,drop=FALSE])
279     colnames(extension.red) <- paste0("extRed",
            1:ncol(extension.red))
280     addr <- getCtrlsAddr(exType = sprintf("Extension (%s)", c("C",
            "G")), index = index)
281     extension.green <- t(greenControls[[index]][addr,,drop=FALSE])
282     colnames(extension.green) <- paste0("extGrn",
            1:ncol(extension.green))
283
284     ## Hybridization should be monitored only in the green channel
285     index <- match("HYBRIDIZATION", controlNames)
286     hybe <- t(greenControls[[index]])
287     colnames(hybe) <- paste0("hybe", 1:ncol(hybe))
288
289     ## Target removal should be low compared to hybridization probes
290     index <- match("TARGET REMOVAL", controlNames)
291     targetrem <- t(greenControls[[index]])
292     colnames(targetrem) <- paste0("targetrem", 1:ncol(targetrem))
293
```

```
294     ## Non-polymorphic probes
295     index <- match("NON-POLYMORPHIC", controlNames)
296     addr <- getCtrlsAddr(exType = sprintf("NP (%s)", c("A", "T")),
           index = index)
297     nonpoly.red <- t(redControls[[index]][addr, ,drop=FALSE])
298     colnames(nonpoly.red) <- paste0("nonpolyRed",
           1:ncol(nonpoly.red))
299     addr <- getCtrlsAddr(exType = sprintf("NP (%s)", c("C", "G")),
           index = index)
300     nonpoly.green <- t(greenControls[[index]][addr, ,drop=FALSE])
301     colnames(nonpoly.green) <- paste0("nonpolyGrn",
           1:ncol(nonpoly.green))
302
303     ## Specificity II
304     index <- match("SPECIFICITY II", controlNames)
305     greenControls.current <- greenControls[[index]]
306     redControls.current <- redControls[[index]]
307     spec2.green <- t(greenControls.current)
308     colnames(spec2.green) <- paste0("spec2Grn", 1:ncol(spec2.green))
309     spec2.red <- t(redControls.current)
310     colnames(spec2.red) <- paste0("spec2Red", 1:ncol(spec2.red))
311     spec2.ratio <- colMeans2(greenControls.current, na.rm = TRUE) /
312         colMeans2(redControls.current, na.rm = TRUE)
313
314     ## Specificity I
315     index <- match("SPECIFICITY I", controlNames)
316     addr <- getCtrlsAddr(exType = sprintf("GT Mismatch %s (PM)",
           1:3), index = index)
317     greenControls.current <-
```

```
      greenControls[[index]][addr,,drop=FALSE]
318   redControls.current <- redControls[[index]][addr,,drop=FALSE]
319   spec1.green <- t(greenControls.current)
320   colnames(spec1.green) <- paste0("spec1Grn", 1:ncol(spec1.green))
321   spec1.ratio1 <- colMeans2(redControls.current, na.rm = TRUE) /
322       colMeans2(greenControls.current, na.rm = TRUE)
323
324   index <- match("SPECIFICITY I", controlNames) # Added that line
325   addr <- getCtrlsAddr(exType = sprintf("GT Mismatch %s (PM)",
      4:6), index = index)
326   greenControls.current <-
      greenControls[[index]][addr,,drop=FALSE]
327   redControls.current <- redControls[[index]][addr,,drop=FALSE]
328   spec1.red <- t(redControls.current)
329   colnames(spec1.red) <- paste0("spec1Red", 1:ncol(spec1.red))
330   spec1.ratio2 <- colMeans2(greenControls.current, na.rm = TRUE) /
331       colMeans2(redControls.current, na.rm = TRUE)
332   spec1.ratio <- (spec1.ratio1 + spec1.ratio2) / 2
333
334   ## Normalization probes:
335   index <- match(c("NORM_A"), controlNames)
336   normA <- colMeans2(redControls[[index]], na.rm = TRUE)
337   index <- match(c("NORM_T"), controlNames)
338   normT <- colMeans2(redControls[[index]], na.rm = TRUE)
339   index <- match(c("NORM_C"), controlNames)
340   normC <- colMeans2(greenControls[[index]], na.rm = TRUE)
341   index <- match(c("NORM_G"), controlNames)
342   normG <- colMeans2(greenControls[[index]], na.rm = TRUE)
343
```

```
344     dyebias <- (normC + normG)/(normA + normT)
345
346     oobG <- extractedData$oob$greenOOB
347     oobR <- extractedData$oob$redOOB
348     oob.ratio <- oobG[2,]/oobR[2,]
349     oobG <- t(oobG)
350     colnames(oobG) <- paste0("oob", c(1,50,99))
351
352     model.matrix <- cbind(
353         bisulfite1, bisulfite2, extension.green, extension.red,
              hybe,
354         stain.green, stain.red, nonpoly.green, nonpoly.red,
355         targetrem, spec1.green, spec1.red, spec2.green, spec2.red,
              spec1.ratio1,
356         spec1.ratio, spec2.ratio, spec1.ratio2, normA, normC,
              normT, normG, dyebias,
357         oobG, oob.ratio)
358
359
360     ## Imputation
361     for (colindex in 1:ncol(model.matrix)) {
362         if(any(is.na(model.matrix[,colindex]))) {
363             column <- model.matrix[,colindex]
364             column[is.na(column)]    <- mean(column, na.rm = TRUE)
365             model.matrix[ , colindex] <- column
366         }
367     }
368
369     ## Scaling
```

```
370     model.matrix <- scale(model.matrix)
371
372     ## Fixing outliers
373     model.matrix[model.matrix > 3] <- 3
374     model.matrix[model.matrix < (-3)] <- -3
375
376     ## Rescaling
377     model.matrix <- scale(model.matrix)
378
379     return(model.matrix)
380 }
381
382
383 ### Return the normalized quantile functions
384 .returnFit <- function(controlMatrix, quantiles, nPCs) {
385     stopifnot(is.matrix(quantiles))
386     stopifnot(is.matrix(controlMatrix))
387     stopifnot(ncol(quantiles) == nrow(controlMatrix))
388     ## Fixing potential problems with extreme quantiles
389     quantiles[1,] <- 0
390     quantiles[nrow(quantiles),] <- quantiles[nrow(quantiles) - 1,]
          + 1000
391     meanFunction <- rowMeans2(quantiles)
392     res <- quantiles - meanFunction
393     controlPCs <- prcomp(controlMatrix)$x[,1:nPCs,drop=FALSE]
394     design <- model.matrix(~controlPCs)
395     fits <- lm.fit(x = design, y = t(res))
396     newQuantiles <- meanFunction + t(fits$residuals)
397     newQuantiles <- .regularizeQuantiles(newQuantiles)
```

```
398     return(newQuantiles)
399 }
400 
401 .returnFitBySex <- function(controlMatrix, quantiles, nPCs, sex) {
402     stopifnot(is.matrix(quantiles))
403     stopifnot(is.matrix(controlMatrix))
404     stopifnot(ncol(quantiles) == nrow(controlMatrix))
405     sex    <- as.character(sex)
406     levels <- unique(sex)
407     nSexes <- length(levels)
408     if (nSexes == 2) {
409         sex1 <- sum(sex == levels[1])
410         sex2 <- sum(sex == levels[2])
411 
412     } else {
413         sex1 <- sum(sex == levels[1])
414         sex2 <- 0
415     }
416 
417     ## When normalization should not be performed by sex separately:
418     if ((sex1 <= 10) | (sex2 <= 10)) {
419         newQuantiles <- .returnFit(controlMatrix = controlMatrix,
420                                    quantiles = quantiles,
421                                    nPCs = nPCs)
422     } else {
423         quantiles1 <- quantiles[, sex == levels[1]]
424         controlMatrix1 <- controlMatrix[sex == levels[1], ]
425 
426         newQuantiles1 <- .returnFit(controlMatrix = controlMatrix1,
```

```
                                    quantiles = quantiles1,
                                    nPCs = nPCs)

        quantiles2 <- quantiles[, sex == levels[2]]
        controlMatrix2 <- controlMatrix[sex == levels[2], ]

        newQuantiles2 <- .returnFit(controlMatrix = controlMatrix2,
                                    quantiles = quantiles2,
                                    nPCs = nPCs)

        newQuantiles <- quantiles
        newQuantiles[, sex == levels[1]] <- newQuantiles1
        newQuantiles[, sex == levels[2]] <- newQuantiles2
    }

    return(newQuantiles)
}


### Normalize a matrix of intensities
.normalizeMatrix <- function(intMatrix, newQuantiles) {
    ## normMatrix <- matrix(NA, nrow(intMatrix), ncol(intMatrix))
    n <- nrow(newQuantiles)
    normMatrix <- sapply(1:ncol(intMatrix), function(i) {
        crtColumn <- intMatrix[ , i]
        crtColumn.reduced <- crtColumn[!is.na(crtColumn)]
        ## Generation of the corrected intensities:
        target <- sapply(1:(n-1), function(j) {
            start <- newQuantiles[j,i]
```

```
456             end <- newQuantiles[j+1,i]
457             if (!isTRUE(all.equal(start,end))){
458                 sequence <- seq(start, end,( end-start)/n)[-(n+1)]
459             } else {
460                 sequence <- rep(start, n)
461             }
462             return(sequence)
463         })
464         target <- as.vector(target)
465         result <- preprocessCore::normalize.quantiles.use.target(
              matrix(crtColumn.reduced), target)
466         return(result)
467     })
468     return(normMatrix)
469 }
470
471 # To ensure a monotonically increasing and non-negative quantile
    function
472 # Necessary for pathological cases
473 .regularizeQuantiles <- function(x){
474     x[x<0] <- 0
475     colCummaxs(x)
476 }
477
478
479 ## WYC
480 .isMatrixBackedOrStop <- function(object, FUN) {
481     if (!.isMatrixBacked(object)) {
482         stop("'", FUN, "()' only supports matrix-backed minfi
```

```
            objects.",
483             call. = FALSE)
484     }
485 }
486
487
488 .isMatrixBacked <- function(object) {
489     stopifnot(is(object, "SummarizedExperiment"))
490     all(vapply(assays(object), is.matrix, logical(1L)))
491 }
492
493
494 .isRGOrStop <- function(object) {
495     if (!is(object, "RGChannelSet")) {
496         stop("object is of class '", class(object), "', but needs
            to be of ",
497             "class 'RGChannelSet' or 'RGChannelSetExtended'")
498     }
499 }
500
501
502 .isGenomicOrStop <- function(object) {
503     if (!is(object, "GenomicMethylSet") && !is(object,
        "GenomicRatioSet")) {
504         stop("object is of class '", class(object), "', but needs
            to be of ",
505             "class 'GenomicMethylSet' or 'GenomicRatioSet'")
506     }
507 }
```

AppendixB

Appendix Tables

Table B.1: Enriched GO terms of age-associated CpGs from the CBL

ONTOLOGY	TERM	N	DE	P.DE	FDR
MF	DNA-binding transcription factor activity, RNA polymerase II-specific	1537	66.17	4.41E-08	0.000770003
MF	DNA-binding transcription repressor activity, RNA polymerase II-specific	224	19	3.78E-06	0.022599894
MF	DNA-binding transcription activator activity, RNA polymerase II-specific	390	26.5	3.88E-06	0.022599894
MF	RNA polymerase II proximal promoter sequence-specific DNA binding	379	25.5	9.89E-06	0.043204613

Table B.2: Enriched GO terms of age-associated CpGs from the MTG

ONTOLOGY	TERM	N	DE	P.DE	FDR
MF	DNA-binding transcription factor activity, RNA polymerase II-specific	1537	311.83	2.50E-30	4.37E-26
MF	sequence-specific DNA binding	297	85	2.75E-17	2.40E-13
MF	DNA-binding transcription activator activity, RNA polymerase II-specific	390	104.5	8.54E-16	4.98E-12
BP	positive regulation of transcription by RNA polymerase II	975	197	2.11E-15	9.20E-12
BP	homophilic cell adhesion via plasma membrane adhesion molecules	129	44.27	1.00E-12	3.51E-09
CC	nucleosome	97	24.5	3.65E-10	1.06E-06
MF	DNA-binding transcription factor activity	480	103.83	4.55E-10	1.14E-06
BP	negative regulation of transcription by RNA polymerase II	707	135.5	9.92E-09	2.17E-05
BP	proximal/distal pattern formation	24	14	2.64E-07	0.0005
BP	neuron differentiation	87	27	1.29E-06	0.00
CC	transcription factor complex	145	38.5	2.01E-06	0.003
MF	transcription factor binding	264	60.5	2.30E-06	0.003
BP	dopaminergic neuron differentiation	19	11	4.21E-06	0.01
BP	embryonic skeletal system morphogenesis	34	15	5.38E-06	0.01
MF	protein heterodimerization activity	497	83.5	6.53E-06	0.01
BP	nucleosome assembly	101	20.5	7.22E-06	0.01
MF	RNA polymerase II regulatory region sequence-specific DNA binding	162	40	7.71E-06	0.01
BP	anterior/posterior pattern specification	62	21	1.06E-05	0.01
BP	female pregnancy	86	22	1.23E-05	0.01
BP	noradrenergic neuron differentiation	5	5	1.35E-05	0.01
CC	nuclear nucleosome	38	10.5	1.56E-05	0.01
MF	DNA-binding transcription repressor activity, RNA polymerase II-specific	224	50	1.84E-05	0.01
BP	regulation of signaling receptor activity	439	59	2.41E-05	0.02
MF	RNA polymerase II proximal promoter sequence-specific DNA binding	379	74.5	2.76E-05	0.02
MF	DNA binding	1329	178.67	3.84E-05	0.03
BP	DNA replication-dependent nucleosome assembly	32	9.5	4.35E-05	0.03
BP	telomere organization	27	8.5	5.85E-05	0.04
BP	G protein-coupled receptor signaling pathway	868	101.5	6.68E-05	0.04
BP	regulation of transcription, DNA-templated	420	72	7.27E-05	0.04
BP	embryonic digit morphogenesis	56	19	8.02E-05	0.05

Table B.3: Coefficients of probes used in the clock of $\text{CerebellumClock}_{\text{specific}}$

Probe	Coefs	Probe	Coefs	Probe	Coefs
Intercept	4.615397483	cg18598861	-0.025145678	cg15477738	-0.030625033
cg24079702	0.023080342	cg13933080	0.004864906	cg00540067	-0.055091553
cg06639320	0.168340984	cg21071793	0.010390503	cg27181013	0.010391212
cg14919554	0.135228392	cg10533159	-0.262225284	cg02058002	-0.026179963
cg07131451	0.027926811	ch.12.1023240F	-0.048993901	cg10225197	0.025456288
cg10621377	0.116252714	cg18984151	0.025311564	cg16748413	0.006224362
cg19958021	0.045621199	cg25540854	0.064849453	cg03968755	0.029299748
cg11520866	-0.033161158	cg22851200	0.142842697	cg25159610	-0.012999504
cg17758721	0.068516975	cg24099956	0.013842244	cg02624770	-0.004602038
cg25801292	0.056992305	cg23464360	0.032982053	cg12093180	0.022573024
cg23040782	0.078381562	cg23052669	-0.0509996	cg05006304	-0.007417847
cg19451698	0.010059917	cg07570470	0.047478856	cg09457766	-0.062591081
cg23981354	0.118808334	cg18473521	0.06885145	cg02114954	-0.031664712
cg02721182	0.04343659	cg26092675	0.042695351	cg13289553	0.015110518
cg15386103	-0.133766339	cg17885226	0.026552041	cg01504656	0.00964106
cg14085673	0.054167078	cg13850871	-0.022887637	cg19691659	-0.025469587
cg21182694	0.142794467	cg23995914	0.029850588	cg07983394	-0.042923899
cg06648759	0.142948773	cg09824900	-0.041203042	cg05666820	0.019658612
cg27529628	0.080521326	cg12141030	0.016392159	cg07620889	0.059441433
cg03742763	0.165274639	cg14020846	0.010521608	cg26529516	-0.000406048
cg14848772	0.041272709	cg02784202	0.050291121	cg09126541	0.003380488
cg04234190	0.029699297	cg09132058	0.011297103	cg14161159	-0.023423127
cg04271792	0.050007943	cg25401874	0.066767009	cg08842907	0.016287509
cg21493505	0.049264117	cg14206898	0.106371241	cg19673233	0.057396527
cg25230305	-0.033123222	cg16029256	-0.024679029	cg06479142	-0.034768728
cg22510037	-0.027707104	cg24136700	-0.004587933	cg10457539	0.039726801
cg04880546	0.058030791	cg01774335	-0.056447729	cg04861640	0.015030797
cg06734271	0.009960723	cg15962547	-0.025071243	cg08702413	0.022465449
cg26392005	0.006969648	cg22830707	0.114392318	cg00823526	0.004314051
cg14042099	-0.104432704	cg10715640	-0.066410483	cg25393429	0.164980051
cg15571405	-0.078733025	cg13632655	0.105151493	cg16312552	0.136943138
cg06144905	0.070221183	cg17486101	0.011148657	cg00088042	0.057445818
cg21572722	0.000503664	cg10097215	-0.002517205	cg15678861	-0.0505892
cg13575161	-0.014866112	cg15154411	-0.013758118	cg16849201	-0.086156493
cg16867657	0.023892664	cg04099767	-0.160057272	cg23201938	-0.021559198
cg00734683	-0.006087136	cg22118147	0.09170626	cg24691835	0.091818491
cg01968178	0.009844345	cg05020257	-0.021559953	cg04502490	0.02114072
cg23941599	0.073364656	cg03314644	0.040439428	cg21113478	-0.00982121
cg20701901	-0.038097599	cg01555253	-0.07817859	cg03320170	-0.077616957
cg19996355	0.000554958	cg12023246	-0.000320073	cg13644645	-0.074853868

cg01259029	0.031875129
cg24883601	0.037688603
cg23376861	0.025921371
cg17535314	0.046054948
cg19399220	0.045685254
cg08342886	0.061119263
cg13873655	-0.027035933
cg07146912	0.10837175
cg13327545	0.092207636
cg01607258	-0.023622724
cg18240400	0.001084152
cg08855249	-0.038756613
cg11999288	-0.085046807
cg16738971	0.01624021
cg15686615	0.02585977
cg23506842	0.048777747
cg17117277	0.101211633
cg16489193	0.026075575
cg17328472	0.055219497
cg14611683	0.071013312
ch.9.126316596R	0.032335964
cg13933043	-0.026806121
cg05708550	0.018393835
cg00533390	0.04117117
cg11071401	0.05313246
cg16367511	0.032969583
cg02406092	-0.148092249
cg16488580	-0.010221948
cg09773897	-0.007301405
cg26647200	-0.024560635
cg07630078	-0.087164948
cg22663995	-0.066305959
cg05944661	-0.012422877
cg17315964	-0.054426894
cg08903089	0.046354954
cg16408394	-0.060224481
cg03702413	-0.116569693
cg05213896	0.064054796
cg16204618	-0.037639265
cg08729686	-0.027797572
cg07158939	-0.101935316
cg22897522	-0.073963718
cg10949007	-0.098470494
cg11467638	0.05671211
cg01180628	-0.024948059
cg27509306	0.021796033
cg27513684	-0.120064989
cg12203250	-0.083282445
cg16475423	-0.10494099
cg01912119	0.020092496
cg18348836	-2.97E-05
cg08606497	-0.016002454
cg00087244	0.008841517
cg18515031	-0.041644991
cg17082959	0.084513841
cg26490949	0.012850493
cg16591502	0.075984488
cg02520768	-0.132671166
cg13165009	0.028965348
cg26472684	-0.061851766
ch.8.1011566R	-0.021161426
cg23530707	-0.005104421
cg08621277	0.081705154
cg05460965	-0.04016868
cg23352942	-0.019932186
cg17951244	0.026243212
cg04837533	0.012797114
cg26672098	-0.045146994
cg23251798	-0.044956663
cg00863378	-0.041966231
cg06851240	-0.08761131
cg00563348	-0.023992116
cg22358580	0.035686192
cg19343530	0.051937272
cg09974780	-0.003726144
cg10729854	-0.038369649
cg19225068	0.05909191
cg00252781	-0.014766346
cg15393490	-0.028438948
cg00576075	0.008832052
cg16135310	-0.077762154
cg07631435	-0.01109643
cg05256179	-0.076325508
cg26790247	-0.010472978
cg07617759	0.053064949
cg03838714	0.069983421
cg11935831	0.023354905
cg19489509	-0.076444281
cg00881372	-0.035160499
cg04157658	-0.026945588
ch.2.20642108R	-0.003824863
cg18868483	0.008508021
cg15919105	-0.023976384
cg11930955	-0.038526599
cg04339860	0.02203238
cg15852490	-0.056741955
cg14989316	0.015218331
cg04362096	-0.117178904
cg05555455	0.005949724
cg04511702	0.014170549
cg22130673	-0.026204679
cg27108925	-0.011217146
cg00580230	-0.004076676
cg04044664	0.109479384
cg24888049	0.081266381
cg16052972	-0.075718869
cg15642666	-0.002617533
cg02251315	0.013387769
cg13814485	0.00959014
cg15696627	0.027896023
cg09576978	-0.061385091
cg01436550	-0.033901622
ch.18.189111R	-0.01196404
cg10775173	0.028147659
cg05331472	-0.056819163
cg22215392	-0.082743141
cg24035598	0.027555373
cg21249595	0.027036319
cg20085953	-0.011226024
cg00049440	-0.054775271
cg24332710	0.000930959
cg06121469	0.019779493
cg15243034	0.039437513
cg04369903	-0.017892085
cg22234080	-0.063002826
cg24481868	0.005466792
cg27560229	-0.025820071
cg20482280	0.030623018
cg17521134	-4.82E-05
cg16018474	0.034059135
cg21597754	-0.063456306
cg10628699	0.105896858
cg09286183	0.012994038
cg19955284	-0.089526012
cg17412102	0.045543294

cg25410668	0.141871138	cg01457778	0.044657464	cg15144453	-0.018589979
cg07044414	-0.001617866	cg08157575	0.004021321	cg21209510	-0.050135807
cg14217495	0.018245777	cg26002103	0.00308299	cg06504162	-0.030349137
cg20098887	-0.00783349	cg13343159	-0.054473721	cg08095452	0.172098194
cg10975586	0.004526433	cg23239612	0.004020824	cg11416276	-0.004174949
cg20523947	0.045187257	ch.7.2635062R	-0.074672194	cg19500863	-0.016578963
cg11528594	0.019479881	cg27111444	0.082955575	cg05878098	-0.005806675

Table B.4: Coefficients of probes used in the clock of $\text{CerebellumClock}_{\text{common}}$

Probe	Coefs	Probe	Coefs	Probe	Coefs
Intercept	3.602442536	cg14611683	0.097166643	cg20426994	0.011610174
cg06639320	0.470160376	cg13933043	-0.181496052	cg22127570	-0.19351547
cg14919554	0.401790505	cg05708550	0.057027612	cg00252781	-0.064329894
cg07131451	0.061348162	cg11071401	0.118806624	cg05970314	-0.204630796
cg25801292	0.221100444	cg16367511	0.071844042	cg02746869	-0.001023435
cg02721182	0.178851061	cg06711656	0.179393753	cg23174607	0.046721036
cg03020208	0.021348759	cg02385167	-0.000690975	cg15393490	-1.175502996
cg14085673	0.295429673	cg10569694	-0.173044977	cg00576075	0.04450524
cg21182694	0.218942228	cg04090392	0.064016789	cg22331349	0.004182091
cg06648759	0.162991894	cg05213896	0.032985847	cg15648389	0.151756651
cg27529628	0.168589903	cg05024939	-0.034923429	cg25047092	-0.018347218
cg18514820	0.255797625	cg26542283	0.197904418	cg19230755	0.027700808
cg17243289	-0.093274329	cg10658666	0.023591668	cg03925294	0.0892044
cg01196788	0.270757891	cg18984151	0.055717203	cg21522254	0.050420261
cg06261926	0.019268481	cg12100751	0.047006428	cg02624770	0.093429522
cg27541691	0.098574016	cg07570470	0.066865611	cg02375320	-0.020047817
cg01122755	0.012124271	cg18473521	0.024059245	cg01504656	0.091119709
cg21493505	0.151383321	cg22736354	-0.120524934	cg05666820	0.085796742
cg22510037	-0.211290927	cg26092675	0.153170509	cg03660500	0.30927725
cg04880546	0.068327782	cg12141030	-0.004742673	cg26529516	-0.15467987
cg06144905	0.091746187	cg14020846	0.122570462	cg23404330	0.066777368
cg21572722	0.203084005	cg03143886	-0.160982731	cg06479142	-0.239799252
cg13575161	-0.11785417	cg24853724	-0.013974728	cg08702413	0.265128011
cg16867657	0.040961471	cg06704773	0.12885834	cg17140307	0.016594567
cg22310062	-0.010053946	cg13202816	-0.03480696	cg23813012	0.090296977
cg19996355	-0.047739805	cg10864952	-0.220964466	cg09935994	0.264109539
cg03664992	-0.162894449	cg20974724	0.302870286	cg27061971	0.053355691
cg01259029	0.22156056	cg22830707	0.104617502	cg04044664	0.10202789
cg10625705	-0.032553158	cg17486101	0.252054367	cg18184411	-0.021772992
cg08342886	0.090288858	cg19929355	-0.097772641	cg23893898	0.0320832
cg06580318	0.015331169	cg03314644	0.162135306	ch.18.189111R	-0.092986133
cg06022942	0.053663208	cg16063312	0.041509305	cg24035598	0.538564173
cg13327545	0.080862072	cg08606497	-0.238644702	cg18943383	0.080650976
cg05009601	-0.000864543	cg05460965	-0.308479657	cg24332710	0.173033933
cg05218976	0.11743311	cg23352942	-0.127213058	cg07935568	-0.099109874
cg16738971	0.200472009	cg26956371	-0.156507046	cg15243034	0.165153932
cg24222995	-0.083280226	cg17365504	0.04245546	ch.1.173201044F	0.063052098
cg01534416	0.036228589	cg26256521	0.064382481	cg17729667	-0.211893843
cg03013329	0.02957672	ch.2.1904845F	-0.097059293	cg24481868	-0.022548251
cg17117277	0.162948458	cg22358580	0.15973967	cg20224218	-0.43543136

cg11614451	0.045614745	cg21878188	0.275490752	cg27560229	-0.138749013

Table B.5: Coefficients of probes used in the clock of $\text{CortexClock}_{\text{common}}$

Probe	Coefs	Probe	Coefs	Probe	Coefs
Intercept	2.899069802	cg26092675	0.098046763	cg16738971	0.024272193
cg13327545	0.124032967	cg15565897	0.004150687	cg18556005	0.225584474
cg13806070	-0.01345442	cg07570470	0.061866246	cg02605173	-0.035888224
cg21572722	-0.03523408	cg19996355	-0.00033944	cg06479142	-0.094544932
cg20591728	-0.002566897	cg19929355	0.045599085	cg02624770	-0.014241251
cg23995914	0.313358375	cg24222995	0.036144106	cg14085673	0.101661778
cg23813012	0.04414117	cg04427498	-0.12811543	cg06711656	0.335178631
cg18514820	0.261478701	cg01196788	0.191756183	cg23376861	-0.048206495
cg18473521	0.463576259	cg07584066	0.044553959	cg02812207	0.009833827
cg05213896	0.045450757	cg03925294	-0.196961171	cg22127570	-0.092560902
cg04792813	8.71E-05	cg18008766	-0.134848674	cg03660500	-0.081092738
cg20692569	0.0086291	cg27061971	-0.048110592	cg01504656	0.018939548
cg09784307	0.017193828	cg20234855	-0.1382834	cg07935568	-0.091220933
cg16867657	0.595893704	cg24035598	0.021908904	cg04710764	-0.215115358
cg24079702	0.042527025	cg07589899	-0.021938945	cg05555455	0.003279565
cg13814485	-0.06638669	cg03020208	0.217607184	cg13202816	-0.134134708
cg12100751	0.019463134	cg06335143	-0.046307374	cg09935994	0.064710477
cg22331349	0.223587802	cg26060489	-0.245588032	cg23352942	-0.096288031
cg07131451	0.180130903	cg04090392	0.111908067	cg25801292	0.201993478
cg16969368	-0.036313278	cg07924892	0.139242162	cg01534416	0.156939942
cg17486101	-0.055849862	cg13543854	0.089015203	cg10235817	-0.056110319
cg15341124	0.31317102	cg25453381	0.153321159	cg26542283	0.161976479
cg17117277	0.079855792	cg00088042	0.226231773	cg24332710	0.054625671
cg17885226	0.125460539	cg02721182	0.045234928	cg16367511	0.110015172
cg02375320	0.111181076	cg21462428	0.171436788	cg17140307	0.098439357
cg15393490	-1.053282745	cg01812045	0.22145375	cg06704773	0.293900522
cg02746869	0.051878906	cg06144905	0.198688018	cg08708711	0.001015099
cg05218976	0.103717114	cg23956238	0.078719163	cg23040782	-0.012175818
cg12141030	-0.251854938	cg24481868	0.023498387	cg26256521	0.094864842
cg21725716	-0.004533933	cg19945840	-0.161008075	cg26490949	0.106826936
cg22454769	0.053018834	cg25047092	0.074038461	cg02385167	-0.071522603
cg04880546	0.496107589	cg13933043	-0.216796839	cg18184411	0.17231342
cg14020846	0.084794981	cg10864952	-0.067104287	cg15648389	0.082224763
cg01429039	0.068758382	cg21868699	-0.022019974	cg22510037	-0.322145178
cg14064148	-0.025077062	cg06648759	0.090731691	cg26529516	-0.092744956
cg26830108	0.041844122	cg10625705	0.059857785	cg25505610	0.009938434
cg18984151	0.063946327	cg10658666	-0.025578742	cg05970314	-0.038574201
cg15731815	0.047360731	cg19230755	0.027242871	cg14311320	-0.00540358
cg12462224	0.43112335	cg08460435	0.006654508	cg04044664	0.086927454
cg17243289	0.02867472	cg20974724	0.072607842	cg10457539	-0.055613751

cg13933080	0.051069457	cg11614451	0.31008648	cg09816471	-0.022277599
cg03607117	0.032715105	cg05917988	0.036496614	cg06625811	-0.148075207
cg05708550	0.027868238	cg18566594	-0.000505277	cg23404330	0.072987599
cg20426994	-0.042219604	cg21182694	0.130596925	cg05460965	-0.073716982
cg00474746	-0.27550611	cg08342886	0.212944609	cg17365504	-0.153255584
cg14611683	0.071232601	cg23174607	0.339575721	cg14405924	-0.057680043
cg27529628	0.241767112	cg00394316	0.11240713	cg19451698	-0.175360647
cg08798295	0.094618303	cg16063312	0.095437127	cg22830707	-0.155444458
cg16549027	0.030715596	cg20224218	-0.8064096	cg08606497	-0.467921124
cg03013329	-0.05128788	cg03391642	-0.149686166	ch.2.1904845F	-0.176878448
cg04374006	0.014291001	cg14919554	0.060210346	cg05009601	0.037444723

Table B.6: Coefficients of probes used in the clock of BrainCortexClock

Probe	Coefs	Probe	Coefs	Probe	Coefs
Intercept	3.688794786	cg00394316	0.02445234	cg26490949	0.140039695
cg24079702	0.105352434	cg14611683	0.114884063	cg05460965	-0.205130123
cg06639320	0.127940455	cg13933043	-0.159993091	cg14405924	-0.066648197
cg14919554	0.129796134	cg05708550	0.124039128	cg23352942	-0.165305973
cg07131451	0.181749371	cg15731815	0.015700942	cg26956371	-0.262711559
cg25801292	0.145318971	cg00593900	0.053302144	cg07584066	0.07404641
cg23040782	-0.024494236	cg11071401	0.062153637	cg26256521	0.072170757
cg19451698	-0.076758861	cg16367511	0.062022617	ch.2.1904845F	-0.178646479
cg02721182	0.060883435	cg06711656	0.283152471	cg22358580	0.008055857
cg26060489	-0.084459055	cg02385167	-0.029117218	cg21878188	0.15763488
cg09784307	0.0603932	cg19945840	-0.174562875	cg22127570	-0.22342384
cg03020208	0.087979542	cg10569694	-0.099333834	cg05970314	-0.222611225
cg14085673	0.188088238	cg16969368	-0.096217267	cg04710764	-0.019580107
cg21182694	0.214867583	cg06625811	0.08710309	cg15393490	-0.986002212
cg06648759	0.02213437	cg04090392	0.037474715	cg22331349	0.129963827
cg27529628	0.174942081	cg05213896	0.053156048	cg21725716	-0.053997311
cg18514820	0.337489177	cg05024939	-0.175575206	cg15648389	0.161161011
cg17243289	-0.022964103	cg01812045	0.158654567	cg25047092	-0.001320272
cg01196788	0.146308137	cg07924892	0.130291605	cg19230755	0.111021837
cg06261926	0.073916836	cg04374006	0.101672449	cg26002103	0.064567214
cg00474746	-0.010535704	cg18549036	-0.056766404	cg21522254	-0.080959644
cg27541691	0.091901071	cg26542283	0.219885805	cg08460435	0.074420264
cg01122755	0.020558435	cg10658666	-0.077226715	cg02624770	0.089294219
cg21493505	0.231591102	cg04427498	-0.010669874	cg02375320	-0.031895224
cg22510037	-0.305586147	cg13543854	-0.147806704	cg01504656	0.062821012
cg04880546	0.058008174	cg18984151	0.092915504	cg05666820	0.074404984
cg14311320	0.010401243	cg20591472	-0.036959876	cg26529516	-0.129175556
cg06144905	0.087120575	cg12100751	0.133274311	cg23404330	0.009183995
cg13575161	-0.106570556	cg07570470	0.06181963	cg06479142	-0.165540804
cg16867657	0.018796186	cg18473521	0.005458251	cg10457539	-0.033258358
cg22310062	-0.029837424	cg22736354	-0.061951302	cg08702413	0.185595624
cg19996355	-0.14607557	cg26092675	0.142913282	cg17140307	0.10454815
cg02812207	-0.044986453	cg17885226	0.057049496	cg02662658	0.07931523
cg03664992	-0.005095546	cg05917988	-0.040402826	cg00088042	0.05380794
cg01259029	0.267677518	cg23995914	0.020268348	cg21462428	0.386549244
cg23376861	0.063474975	cg25453381	0.096282582	cg21868699	-0.070282407
cg12462224	0.272313917	cg12141030	-0.073090238	cg06335143	-0.02324295
cg10625705	-6.01E-05	cg14020846	0.044453017	cg09935994	0.187681998
cg08342886	0.113049161	cg13806070	0.084309734	cg04044664	0.036733378
cg06580318	0.091325635	cg03143886	-0.020970338	cg23893898	0.057767953

cg15341124	0.184941278	cg24853724	-0.024324929	ch.18.189111R	-0.18012466
cg06022942	0.021317129	cg06704773	0.301418405	cg24035598	0.08226459
cg13327545	0.132689907	cg13202816	-0.085364763	cg18943383	0.095811692
cg05009601	-0.099685991	cg01429039	0.086366709	cg24332710	0.146253278
cg05218976	0.054778806	cg10864952	-0.217354131	cg07935568	-0.202425565
cg16738971	0.005559577	cg20974724	0.245789111	cg15243034	0.079587812
cg24222995	0.049563078	cg17486101	0.005732704	ch.1.173201044F	0.083795909
cg01534416	0.062130936	cg18556005	0.034654625	cg17729667	-0.223709359
cg03013329	0.085327187	cg19929355	-0.002784753	cg24481868	-0.006090312
cg25505610	-0.006142712	cg03314644	0.014435737	cg20224218	-0.445360996
cg17117277	0.079798051	cg16063312	0.042388076	cg16549027	-0.009629751
cg26830108	0.128985952	cg08139499	-0.006171374	cg27560229	-0.036199983
cg11614451	0.161783741	cg08606497	-0.294246678	cg13785883	-0.040965789