DNA Methylation: Methods and Analyses

Tyler J. Gorrie-Stone

A thesis submitted for the degree of Doctor of Philosophy in Biochemistry

School of Biological Sciences

University of Essex

September 2019

Abstract

Epigenome-wide Association Studies (EWAS) have been a popular method to investigate the genome over the past decade. From these experiments, more than 75,000 samples have been assayed using the high-throughput, cost-effective HumanMethylation450 microarray (450k) developed by Illumina. With the recent release of the HumanMethylationEPIC microarray, the size of data is expected to increase considerably so advances are needed in the methodologies used to analyse such data.

The first part of this thesis focuses on the development of tools that can be used for the analysis of DNA methylation microarray data. Firstly I develop a wide range of tools that can be used to quality control data. These tools focus specifically on data-driven aspects of quality control that are often overlooked and can cause problems during downstream analysis. Comparison of these tools to other popular methods demonstrate that the tools I created are effective in decreasing test statistic inflation while conserving the largest number of samples (Chapter 2). Secondly to accommodate the increase in the size of data, I developed the bigmelon R package which reduces the amount of memory required to perform the analysis typically required of EWAS (Chapter 3).

I then demonstrate how both the tools described in Chapters 2 and 3 can be used in EWAS settings. I perform an EWAS between DNA methylation and various blood-lipid traits and statin-use on a dataset comprising of 1,193 samples from the Understanding Society: UK Household Longitudinal study and replicate the findings of many previous EWAS (Chapter 4). Lastly, I demonstrate how the data from tens of thousands of microarrays can be utilised in preliminary analyses that focus on the wide-spread characterisation of the probes on the 450k microarray and how tissue-specific DNA methylation patterns may correlate with tissue-specific gene expression (Chapter 5).

Declarations

Chapter 2

Publicly available datasets were sourced from GEO under platform accession number GPL13534. I created and tested all of the tools that were added to the wateRmelon R package and I performed all of the analyses described in this chapter.

Chapter 3

Initial prototyping of the bigmelon R package was done by Professor Leonard Schalkwyk, Dr Ayden Saffari and Dr Karim Malki. Majority of the code and the resultant software was rewritten by myself for this thesis. I wrote every function, manual page and performed all of the benchmarking and statistical analysis to produce the resultant manuscript that is included as Chapter 3.

Chapter 4

Sample and data collection for the Understanding Society: UK Household Longitudinal Study was performed by Dr Melissa Smart, Dr Yanchun Bao and Professor Meena Kumari at the Institute of Socio-Economical Research at the University of Essex in collaboration with Professor Jon Mill, Dr Eilis Hannon and Dr Joe Burrage from the University of Exeter Medical School. I performed the data quality-control and preparation for the data submission for EGA. The statistical analysis describe in this chapter were also performed by me.

Chapter 5

I sourced all the data used from GEO and the Marmal-Aid database. Gene expression data was obtained from the InterMine Web Resource. All analysis described in this chapter was performed by me.

Acknowledgements

I wish to thank numerous people for their contributions towards this Thesis and my PhD in general. Firstly I would like to thank my PhD supervisor Professor Leonard Schalkwyk for the opportunity, the advice and wisdom you have afforded to me over the last four years. Additionally this work would not have been possible without the contributions of the people at ISER (Professor Meena Kumari, Dr Melissa Smart, Dr Yanchun Bao) and Exeter (Professor Jon Mill, Dr Joe Burrage, Dr Eilis Hannon) for providing data and invaluable academic insight.

Personally I would like to extend my gratitude to all members of the genomics group at Essex. Especially to Mohab, Patrick, Stuart, Myles, Dave and Louis the traitor. I would also like to thank the Knotty Boys: Tim 'Nameless' Less, Amonis Hawkwood III, Je'faac and Jim who provided countless entertainment most Thursday providing Sam was not being a flake.

I would also like to thank my family for their enduring support and bother.

Lastly I would like to thank Louise. Words cannot express how thankful I am that you have accompanied me through this experience.

List of Abbreviations

27k	HumanMethylatoin27 BeadChip microarray
450k	HumanMethylation450 BeadChip microarray
5hmc	5-hydroxymethylcytosine
5mc	5-methylcytosine
A	Adenine
BMI	Body mass index
bc	beadcount
C	Cytosine
CpG	Cytosine-Guanine Dinucleotide
CVD	Cardiovascular Disease
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
DMR	Differentially Methylated Region
DMP	Differentially Methylated Position
EWAS	Epigenome-Wide Association Study
EPIC	HumanMethylationEPIC BeadChip microarray
G	Guanine
GEO	Gene Expression Omnibus
GWAS	Genome-Wide Association Study
HDL-C	High Density Lipoprotein Cholesterol
IQR	Interquartile Range
LDL-C	Low Density Lipoprotein Cholesterol
MIAME	Minimum Information About Microarray Experiment
mRNA	mature Ribonucleic acid
mQTL	Methylation Quantitative Trait Loci
OOB	Out-of-band
RMSD	Root Mean Square of Difference
RNA	Ribonucleic acid

RRBS	Reduced-Representation Bisulfite Sequencing
SNP	Single Nucleotide Polymorphism
SDD	Standard Deviation of Difference
T	Thymine
TC	Total Cholesterol
TCGA	The Cancer Genome Atlas
TG	Tryglyceride
TSS	Transcription Start Site
UKHLS	Understanding Society: UK Household Longitudinal study
UTR	Untranslated Region
WC	Waist Circumference
WGBS	Whole Genome Bisulfite Sequencing
ZF	Zinc Finger domain
ZF-D3A	Zinc Finger domain - DMNT3A fusion protein

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Epigenome-wide Association Studies (EWAS) between DNA methylation and disease or environmental exposures have become increasingly popular in biomedical research (Rakyan *et al.*, 2011). Relying on high-throughput and cost-effective microarray technology EWAS can examine thousands of samples reliably across hundreds of thousands of loci spread throughout the genome (Bibikova *et al.*, 2011). Findings from these studies can be used to develop our understanding concerning how a disease or trait can manifest within the human body.

Due to popularity and relative ease, EWAS within the field of epidemiology are being performed at a rapid pace (Li *et al.*, 2019). With each new study hoping to leverage a better understanding of the cause of disease from the wealth of information that is produced by these microarrays. In accompaniment to this data is the steady development of methodologies that can be used to analyse the data, test genome-wide associations and interpret these findings. However, there is a need for some caution with respects to EWAS as the exact methods used for analysis can often go unreported. This can lead to difficulty when trying to reproduce or compare results between studies.

In this thesis, I set out to address some of the concerns relating to EWAS in general and suggest some methodological improvements where appropriate. Methods used for both statistical testing and data normalisation are generally well described and reported in most studies. However, methods that are used

during preliminary stages of analysis such as quality control are distinctly lacking in both design and how they are reported.

Another feature of EWAS that is regularly overlooked are limitations associated with the popular analysis platform, the R programming language (R Core Team, 2017), that is used to carry out most of these investigations. The R language is a useful tool for biologists and provides enough flexibility to perform almost any type of analysis. Additionally, the scientific community encourages that developers of the R language contribute their efforts to public repositories such as Bioconductor (Gentleman *et al.*, 2004) to encourage the development of reproducible research. One drawback of the R language is that it makes extensive use of computers memory to store data and perform analyses, this becomes the main bottleneck for any analysis that attempts to process large datasets. It is possible to avoid requiring large amounts of memory by selectively reading in small amounts of data, however this process requires prior, specific, knowledge of the data being analysed which is rarely the case when handling biological data. This bottleneck may disappear as access to high-performance computing clusters become more accessible within academic settings or as computers become cheaper and more powerful. However, as data sizes are anticipated to increase rapidly, it is likely that new algorithms to analyse data are required to be able to cope with the increasing burden of data size efficiently. The memory overheads currently associated with R, limit the ability to perform large scale analyses. Notable attempts such as those by Horvath (2013) and Lowe & Rakyan (2013) have been made but either focus on a small portion of the data, or the resources that were developed to facilitate such large scale analyses were unable to handle the computational requirements. Therefore, I would like to demonstrate what sort of analyses are possible when such limitations are alleviated.

This chapter introduces the concept of epigenetics and how the complex of epigenetic factors and the genetic sequence results in a change in gene expression - thus giving rise to the hundreds of cell-types that exist. I will also describe how the field of epigenetics was established over time with a particular interest in the relationship between DNA methylation and blood-lipid phenotypes and cardiovascular disease.

1.1 Epigenetics

Out of the 38 trillion cells estimated to be in the human body (Sender *et al.*, 2016), the majority of these will contain the same complement of DNA. Despite being genetically identical, these cells vary wildly in size, shape and function. This variation led scientists to wonder how it was possible for so many phenotypes to arise from a single genotype.

The term 'epigenetics' was first coined by Waddington (1942) to describe mitotically heritable cellular events, that ultimately lead to a change in gene expression, which are not caused by changes in the genetic sequence. This definition of 'epigenetics' describes the process of epigenesis which was the prevailing theory whereby all cells differentiate from a single undifferentiated cell. This is classically described, diagrammatically, by the journey of a marble traversing a down a canalised landscape (Figure 1.1). Where once the marble has chosen a course it has to continue down the chosen path until its ultimate fate.

Since then the definition of 'epigenetics' has changed numerous times (Haig, 2004) with the current and most popular definition currently describing the collection of heritable changes in gene function that are not explained by a change in the genetic sequence. This definition can be seen as a derivation from the words 'epi' (upon) and 'genetics' (genetic sequence) to reflect the idea that there are additional layers of information that exists on top of the DNA sequence that is thought to influence gene expression directly. The evolution of the term epigenetics has been met with concern as the more modern and fashionable definition is more liable to be abused when interpreting results. As a result there are many who advocate clarity when using the term 'epigenetics' to describe a change in gene expression (Greally, 2018). For the sake of clarity, I would like to specify that I will be referring to epigenetics according to the modern definition.

Modern epigenetics is attractive to biologists as it identifies genes and regulatory mechanisms that could be related to the cause of complex disease (Bernstein *et al.*, 2007). Genome-wide investigations have identified numerous SNPs and haplotypes that contribute towards various disease but are unable to ex-

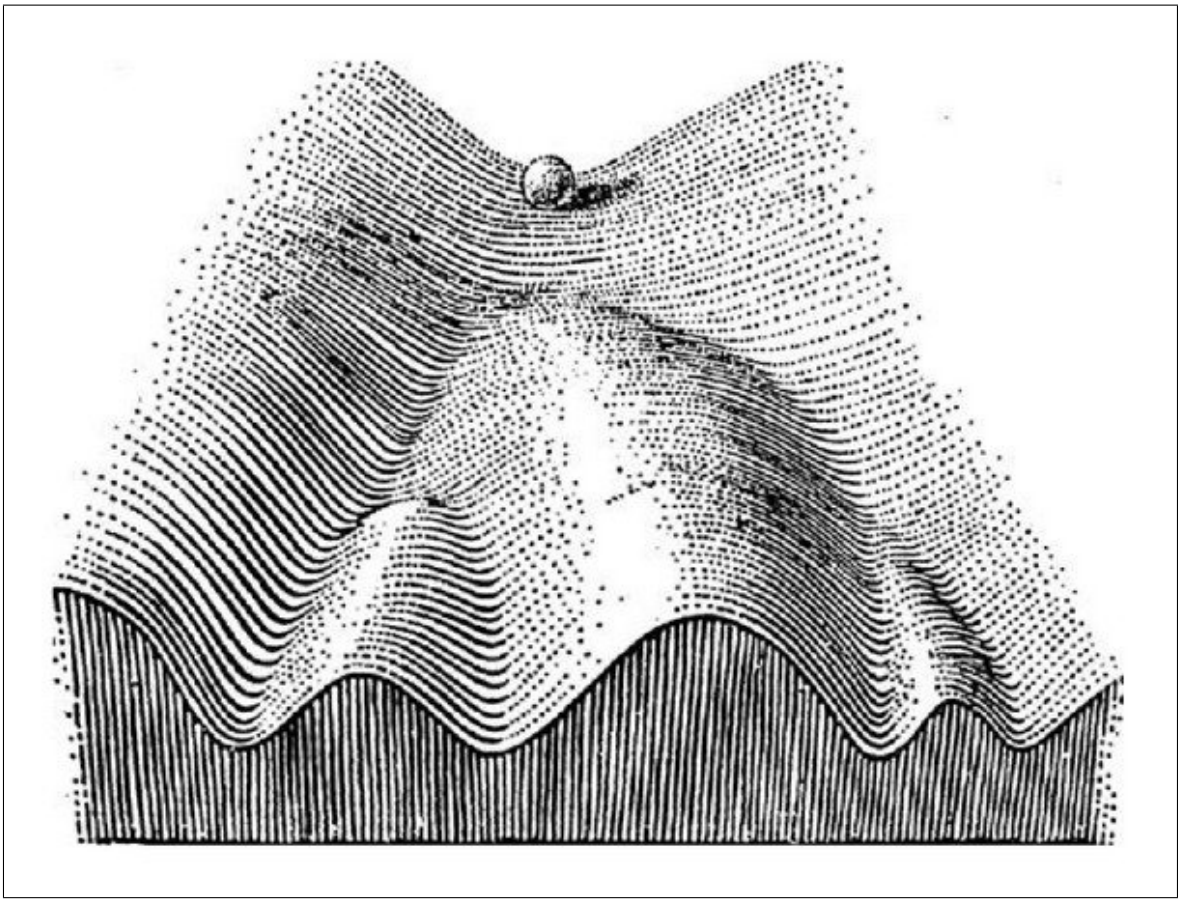


Figure 1.1: Waddington (1957)'s Epigenetic Landscape

plain why certain individuals will develop the disease and others may not. It has been shown that disease can arise when these epigenetic mechanisms are dysregulated (Bestor, 2000; Herman & Baylin, 2003; Feinberg, 2007). This suggests that the layers of information on top of the genome do contribute towards gene expression and is therefore of interest to us to investigate how gene expression changes alongside changes in the epigenome in response to certain stimuli, e.g. the environment.

These layers of information that supposedly influence gene expression are primarily broken down into three categories: DNA modifications, Histone Modifications and non-coding RNAs (Bernstein *et al.*, 2007; Costa, 2008). For this thesis, I will focus on DNA modifications, specifically DNA methylation, as it is considered the easiest epigenetic modification to examine and is popular within the scientific community. This is due to the relative stability and abundance of required material (DNA) and the high-throughput nature of the techniques used to interrogate it. The decision to focus solely on DNA methylation does not detract from the importance that histone modifications or non-coding RNAs may have on gene expression. Briefly, DNA modifications describe any chemical modification that affects a DNA base. Histone modifications describe the chemical modifications that attach to the histone proteins to which DNA itself is wound around and usually alter chromatin state. Non-coding RNAs describe the set of RNA molecules that bind to transcribed DNA (e.g. micro RNA).

1.2 DNA methylation

In most situations, DNA methylation refers to the covalent attachment of a methyl group (CH_3) to the 5C atom of a cytosine nucleotide base (Razin & Riggs, 1980) to form 5-methylcytosine (5mc). DNA methylation in this manner is catalysed by a group of DNA methyltransferase enzymes (DNMTs) which use S-adenosyl-Methionine as the methyl group donor (Bird, 2002; Bestor, 2000). In mammals, this process usually occurs on Cytosines within a cytosine-guanine dinucleotide (CpG) context (Bernstein *et al.*, 2007), aptly referred to as CpG methylation. Mammalian DNA methylation is not limited to 5mc in a CpG context. Non-CpG methylation denoted as CpHpH or CpHpG methylation (where H corresponds to A, T or C), has also been observed in both mammals and plants however the function of Non-CpG

methylation is not understood as clearly as CpG methylation (Lister *et al.*, 2009; Laurent *et al.*, 2010).

There are approximately 28 million CpG sites spread throughout the genome. It has been estimated that as many as 80% of these CpGs are methylated (Wang & Leung, 2004; Saxonov *et al.*, 2006; Ehrlich *et al.*, 1982). Despite the relatively small number of CpGs in the genome, CpGs are often found in high densities often located within promoter regions of genes (Klose & Bird, 2006). The high densities of CpGs are referred to as CpG islands (Gardiner-Garden & Frommer, 1987; Illingworth & Bird, 2009) and account for approximately 7% of all total CpGs in the genome and are usually unmethylated. Considering that CpGs are usually methylated, this leads to the conclusion that there is likely to be some functional significance regarding the methylation state of CpGs within CpG islands (Bird, 1987).

Additionally CpGs can be characterised by their location relative to their nearest CpG island. Simply, CpGs within CpG islands are creatively known as CpG-island CpGs whereas Non-CpG island CpGs are classified into Shores ($<$ 2kb from an island), Shelves ($<$2kb from Shore) and Open Sea or intergenic ($>$ 2kb from Shore) CpGs (Irizarry *et al.*, 2009). These shores and shelves account for CpGs that may not fall into promoter regions but otherwise could be functionally significant as Irizarry *et al.* (2009) demonstrate that there is a strong relationship between gene expression and these distal CpGs.

In mammals, DNA methylation patterns rapidly erase upon fertilisation (Reik *et al.*, 2001). These patterns are not reestablished until the blastocyst stage where differentiation starts to occur. This type of methylation occurs in a *de novo* fashion by the enzymes DNMT3A and DNMT3B (Reik *et al.*, 2001; Jaenisch & Bird, 2003). After DNA methylation patterns are established, the pattern persists through mitosis by DNMT1 which binds to hemimethylated DNA with a high specificity (Bestor, 1992, 2000). It is thought that the DNA methylation patterns are accrued stochastically with this *de novo* mechanism as a response to the environment (Reik *et al.*, 2001).

The removal of 5mc can be described using two mechanisms. The first mechanism is through passive demethylation through down-regulation of DNMT1 results in hemimethylated DNA remaining hemimethy-

lated. Thus after subsequent replication events, the proportion of methylated DNA is diluted. The second method is known as active demethylation where 5mc is converted into 5hmc through oxidation by TET1 enzymes which eventually results in the reestablishment of unmethylated Cytosine (Guo *et al.*, 2011).

As DNA methylation is frequently correlated with gene expression, it is no surprise that it is thought to regulate transcription in some capacity. In some situations, DNA methylation is shown to interfere with the binding of transcription factors as with the case of CTCF (Hark *et al.*, 2000) or that DNA methylation can recruit methyl-binding domain proteins to alter the chromatin state (Bird, 2002). DNA methylation has established roles in Genomic Imprinting (Razin & Cedar, 1994; Li *et al.*, 1993), X Chromosome inactivation (Riggs, 1975; Singer-Sam & Riggs, 2012) and in global demethylation events associated with cancers (Jones & Baylin, 2002; Feinberg, 2007).

It needs to be considered that 5mc is not the only form of DNA modification that exists in the genome. Indeed adenine methylation among other non-methylation based modifications such as 5hmc could have roles in gene regulation. For example, adenine methylation is involved in the silencing LINE-1 transposable elements (Wu *et al.*, 2016). While 5hmc is not only limited to DNA demethylation but also has roles in pluripotency and development (Branco *et al.*, 2011).

1.3 Epigenome-wide Association Studies

If we were to consider the success and high throughput nature of Genome-Wide Association studies (Hindorff *et al.*, 2009; Chanock *et al.*, 2007; Welter *et al.*, 2014; Johnson & O'Donnell, 2009, http://www.ebi.ac.uk/gwas) it is no surprise that EWAS have become popular. Early reviews by Rakyan *et al.* (2011) and Mill & Heijmans (2013) correctly placed EWAS using the microarray technologies as an excellent strategy for the identification of novel association.

Summarising every EWAS performed in the last decade would be a thankless task. Since the introduction of the EWAS design many differentially methylated positions and regions have been identified for tissues

(Davies *et al.*, 2012; Hannon *et al.*, 2015; Varley *et al.*, 2013; Ziller *et al.*, 2013), disease (Feinberg, 2007; Portela & Esteller, 2010), environment (Feil & Fraga, 2012), a variety of socio-economic & lifestyle habits (Breitling *et al.*, 2011; Lee & Pausova, 2013) and peri-natal differences seen in pregnant women (Tobi *et al.*, 2009; Spiers *et al.*, 2015). In addition, since 2009 considerable effort has been made to establish epigenetic consortia to pool together experiments to create highly collaborative resources that could be used to describe the epigenetic landscape of disease and tissue such as The Cancer Genome Atlas, the Epigenome Road Map and the Human Epigenome Project (Bernstein *et al.*, 2010; Kundaje *et al.*, 2015).

The early reviews (Rakyan *et al.*, 2011; Mill & Heijmans, 2013; Michels *et al.*, 2013) championed the EWAS as being a useful strategy to investigate the epigenome and provide numerous recommendations that should be considered going forward. The following recommendations I have briefly summarised as they are important to stress where possible.

1. **EWAS are not limited to Cross-sectional designs** – The epigenome is dynamic and is susceptible to change over time in response to the environment. Therefore it is an excellent candidate for a wide variety of informative designs. Majority of EWAS use a cross-sectional design where both the DNA methylation patterns and an exposure or outcome are obtained at the same point of time. This approach is by far the easiest to perform as it only requires a single time point of investigation and can identify numerous associations depending on the number of samples. This design does not play into the strengths of EWAS as they can benefit from longitudinal or even short-term exposure studies where using multiple time points can elucidate a clearer picture of how the epigenome responds to various traits.

2. **EWAS are sometimes unable to infer a direction of cause** – As the most common form of EWAS makes use of a cross-sectional design; these EWAS are frequently only able to identify when a relationship exists between an exposure or outcome to DNA methylation. These relationships may not necessarily impart any causal evidence to suggest the exposure or outcome is caused by a change in methylation or that the change is caused by the specified exposure or outcome. To establish a direction of cause, the more elaborate and different designs (e.g. longitudinal) are re-

quired. However, it may be possible to infer a direction of cause by using analysis techniques such as Mendelian randomisation (Relton & Davey Smith, 2012).

3. **Choice of Tissue** – It is a well known fact that the epigenome varies between tissue and cell type. Therefore it is essential to select a biologically relevant tissue to examine a biological question when performing an EWAS. The choice of tissue becomes a potential issue when attempting to study an outcome or exposure in a tissue that is difficult to obtain (e.g., brain tissue). This issue is further compounded when trying to utilise a longitudinal design as finding a large enough group of participants willing to undergo frequent invasive procedures can be difficult. As a result majority of the EWAS to date have been performed using DNA obtained from whole blood or blood sub-fractions as it is a relatively easy tissue to obtain. This raises concerns about the validity of results that are obtained from studies that do use surrogate tissues to perform EWAS in place of tissues that are difficult and expensive to obtain. As a result, analyses involving surrogate tissues may require additional validation.

   Further to this some tissues such as whole blood are heterogeneous and contain many different cell types. Samples individually will have different proportions of these cell types which can produce spurious associations should the model be uninformed that such cell-type specific variation exists. Considerable effort has been made in developing strategies to handle cell heterogeneity. Aside from experimentally identifying the proportion of cell types within a given sample, which may not be available to researchers, alternative methods can be used to estimate these proportions (Houseman *et al.*, 2012) and have been widely adopted in EWAS. Houseman *et al.* (2012)'s method uses a reference dataset for blood cell types, but various other reference datasets for different tissues have been collected and reference-free methods are available for tissues that otherwise do not have a reference dataset (Teschendorff *et al.*, 2017).

4. **Additional layers on confounding** – In addition to selecting a relevant tissue and accounting for cellular heterogeneity within tissues, the epigenome is also variable to other environmental exposures. These include and are not limited to age, sex, disease, treatment and lifestyle habits (e.g.,

smoking, alcohol consumption and diet). Furthermore EWAS can also be plagued by systematic batch effects where differences in experimental design lead to a difference in signal and genetic confounding (population structures and familial effects) both of which GWAS are familiar with (Johnson *et al.*, 2007; Devlin *et al.*, 2001).

Usually, these sources of confounding can be handled by including additional exposures as covariates within statistical models. In situations where the exposures have not been recorded, it is possible to make use of methods that could identify surrogate variables that can explain some of the variance within a dataset that then can be included in a model. There are some elements of caution that need to be taken with using many covariates in statistic models as it is possible to over-fit the data and lead to results that may be meaningless.

Failure to take account of confounding and cellular heterogeneity will lead to false positive results which can mislead genuine interpretation and further analysis of results. In GWAS these false positive results can be account for by applying some form of genomic control (Devlin *et al.*, 2001) which entails dividing the test statistics by a factor. In EWAS as the source of the confounding is not limited to population structure or cryptic relatedness a specialised method can be used (bacon) but has yet to be widely adopted (van Iterson *et al.*, 2017)

1.4 Assessing DNA methylation

The methylation state of DNA can be assessed in numerous ways. These range from sequencing-based methods, immunoprecipitation of methyl-binding proteins, methylation-sensitive restriction enzyme assays and hybridisation arrays. Out of these, the sequencing-based technologies are considered to be the ideal experiment to investigate methylation patterns. However these sequencing methods are unable to differentiate between cytosine and 5mc unless the DNA undergoes sodium bisulfite treatment. This sodium bisulfite treatment deaminates cytosine into uracil however the methyl group (and other modifications) in 5mc protects this deamination occurring thus leaving it intact. Sodium bisulfite treated DNA can then

be amplified, which corrects uracil into thymine while the 5mc propagates as cytosine. Thus allowing for differentiation between unmethylated cytosine and 5mc (Frommer *et al.*, 1992). Additionally, next generation sequencing techniques such as NanoPore Sequencing can be used to detect methylated cytosine without the need of bisulfite conversion (Clarke *et al.*, 2009).

It should be noted that bisulfite treatment does not discriminate from the different types of DNA modifications such as 5hmc (Huang *et al.*, 2010). As a result, it is possible that the DNA methylation patterns obtained in this manner are confounded by a mixture of signals from different DNA modifications. For this oxidative bisulfite sequencing can be used to further differentiate between 5mc and 5hmc (Booth *et al.*, 2012, 2013).

Whole Genome Bisulfite sequencing is considered the gold standard method to obtain DNA methylation patterns. It is capable of interrogating the methylome of all of the 28 million CpGs at a single base nucleotide resolution however it is expensive to carry out on hundreds of samples. Alternatives such as reduced representation bisulfite sequencing and hybridisation arrays allow for the same single base resolution but across a smaller coverage for a fraction of the cost. The alternative methods of investigating DNA methylation patterns can also cheaply examine DNA methylation but do not have the same resolution or have as wide coverage compared to the sequencing-based methods.

Ultimately the DNA methylation microarrays became the platform of choice for EWAS. Bibikova *et al.* (2009) describes the repurposing of the commonly used SNP arrays with specific probes that are designed to hybridise to a methylated or unmethylated strand of DNA (following bisulfite treatment). By doing this, it was possible to provide a technology that allowed for identical coverage across many samples in a reasonably high throughput and cost-effect manner. This first iteration of the Infinium BeadChip technology repurposed for DNA methylation was called the HumanMethylation27 BeadChip (27K) microarray and was able to interrogate the methylation state of approximately 27,000 CpG sites located across the genome. These CpG sites were almost exclusively located within proximal promoter regions of nearly 15,000 genes (Bibikova *et al.*, 2009). A few years later the 450K microarray was released (Bibikova *et al.*, 2011) which extended the coverage of the 27K vastly by introduction an additional 450,000 loci

that could be queried. This extension required the introduction of a new probe design (Type II design) which makes use of a single probe and can detect methylation changes in CpGs in regions of relatively low density. The combination of both these probe designs allows for comprehensive genome-wide coverage of the genome for many genes and regions of interest. This microarray has been used extensively in EWAS and was selected as the platform of choice for many studies. The 450K has been succeeded by the EPIC array (Moran *et al.*, 2016) which extends the number of loci scanned by nearly double (up to more than 850,000 loci), many of these loci are of the Type II design and located in regions where the biological may not be as well understood such as enhancer regions.

Biologically speaking DNA methylation is considered a binary trait, either methylated or unmethylated. However in EWAS many DNA molecules are being queried per sample and as a result, DNA methylation is often expressed as a β ratio described as:

$$\beta_i = \frac{Me_i}{Me_i + Un_i + \alpha} \tag{1.1}$$

Where Me is the given methylated signal for a given loci (i), Un is the unmethylated signal and α is an arbitrary offset to handle signals with low readings (usually 100). Conveniently these β values are bound between 0 and 1 which lends itself to easy interpretation where a value of 0 is equivalent to all DNA strands at a given locus being unmethylated and a value 1 corresponds to them all being methylated. From the above formula it is possible to acheive a β value of 0 but because of the addition of α a β value of 1 is never attainable. As the raw signal intensities (Me and Un) are usually in the thousands, the addition of α makes a very little impact in the resultant β values.

Some concerns have been raised over the mathematical properties of β values. Firstly as the β values are bound between two values they cannot be normally distributed and therefore violate the assumptions of statistical tests. Secondly, the β values are heteroscedastic, where the variation of β values for a given loci can vary differently across the range of a given variable. To handle this, Du *et al.* (2010) suggest that M-values, defined as the $\log_2$ ratio between Me and Un intensities, should be used for statistical testing as they are indeed homoscedastic but otherwise directly proportional to β values. These M-values

are technically unbounded and share a linear relationship with β values at intermediate methylation levels (between 0.2 and 0.8) however distort DNA methylation levels at both the high and low values. This approach has a couple of limitations. Firstly M-values are more difficult to interpret over β values as they represent a fold-change in methylation rather than a change in the percentage of methylation. Secondly towards the extremes of methylation M-values can inflate the difference between small values. For example, a change in β value of 0.01 between 0.05 & 0.06 equates to a change of 0.28 between M-values. However, a change of 0.01 between β values of 0.10 and 0.11 equates to a change in M-values of 0.15 despite the difference in βs being the same. This difference in M-values can potentially favour variation in probes which are either highly (un)methylated and penalise loci where there is naturally a greater variation in methylation. It has also been noted that while the overall distribution of β values are characterised with two peaks (bimodally distributed), the β values across single loci are usually unimodally distributed and do not violate the assumptions of the common statistical tests used in EWAS as profoundly as once thought.

1.5 EWAS and Lipids

As previously mentioned, EWAS have been successfully used to explore a variety of exposures and outcomes. Among these are EWAS that focus on metabolic traits that are associated with cardiovascular disease (CVD). Cardiovascular disease is the largest cause of death in Humans in both developed and underdeveloped countries. Typically CVDs refer to the complement of diseases that are related to the heart and circulatory system. CVDs have a large number of risk factors - Genetics, Smoking, Obesity, Diet, Exercise and high cholesterol. GWAS have previously (Kessler *et al.*, 2016; Arking & Chakravarti, 2009) identified numerous SNPs that are associated with CVDs or CVD-based events (Stroke, Heart Attack) and GWAS looking more specifically at blood-lipid levels (Willer *et al.*, 2013) are well established. However, the SNPs identified only explain a small percentage of the relative risk associated with developing these disease. So we once again turn to epigenetics and attempt to identify a mechanistic explanation that would contribute to CVDs or at least high blood-lipid Levels.

The relationship between blood-lipids and epigenetics have been reviewed numerous times (Sayols-Baixeras *et al.*, 2016a; Dekkers *et al.*, 2016a; Mittelstraß & Waldenberger, 2018) which firmly sets the pretense that blood-lipid EWAS have been successful and produce numerous reproducible results. EWAS looking at blood-lipid concentrations were relatively slow to get off the ground with the first one to be performed in 2014 by Petersen *et al.* (2014) who investigated a variety of metabolic traits and identify a surprisingly small number of associations considering the number of traits the authors had tested (639 traits). Particularly notable results from Petersen *et al.* (2014)'s study include associations of CpGs within DHCR24 and ABCG1 with Total Cholesterol. A few more studies looking at blood-lipid concentrations were also published in the same year. Irvin *et al.* (2014) and Frazier-Wood *et al.* (2014) both identified the association of CPT1A with LDL-C and TG concentrations in CD4+ T cells. These results were later reproduced in peripheral whole-blood by Gagnon *et al.* (2014) firmly establishing a definitive relationship between DNA methylation and blood-lipid levels.

The next largest study was performed a year later by Pfeiffer *et al.* (2015), who identified numerous associations for different lipid traits, notably an association to ABCG1 with HDL-C and associations with ABCG1, SREBF1, TXNIP and CPT1A with triglyceride concentrations. Similar findings were also found in the study by Sayols-Baixeras *et al.* (2016b) who also reproduced the inverse relationship between TG and HDL with ABCG1.

A distinct feature by all of the studies up to this point is that they were all performed using a cross-sectional design. As a result, none of the authors were realistically able to determine whether or not the elevated blood-lipid levels had contributed towards the changes in methylation. The study by Dekkers *et al.* (2016b) made use of Mendelian randomisation to attempt to identify a direction of cause which in turn provided evidence suggesting that an increase in HDL-C and TG concentrations that lead to the change in methylation at the frequently reported ABCG1 locus.

The most recent and largest studies by Hedman *et al.* (2017) and Braun *et al.* (2017) replicate these results of the past blood-lipid EWAS very well and also report numerous novel findings that had not been reported by previous EWAS. Hedman *et al.* (2017) identify 25 novel associations between DNA

methylation and blood-lipid measurements. While the findings were novel with respects to blood-lipid concentrations, there was considerable overlap between the novel results and other metabolic traits such as adiposity and type 2 diabetes. Braun *et al.* (2017) reproduce the previous associations between HDL-C and TG with genes such as DHCR24, ABCG1, SREBF1 and CPT1A. Pathway analyses of these genes identifies enrichment for lipid, sterol and cholesterol biosynthesis and transport (Hedman *et al.*, 2017) demonstrating that the results from these EWAS are identifying CpGs within genes that have some biological importance relating to lipid biology.

There is considerable overlap in the results of EWAS looking at blood-lipid levels and other metabolic traits such as BMI, Waist Circumference (WC) and Type 2 Diabetes. The EWAS of BMI by Mendelson *et al.* (2017) had reported that DNA methylation of CpGs within SREBF1, ABCG1 and DHCR24 (and others) were also related to BMI. These results have been reported in many other independent BMI EWAS (Aslibekyan *et al.*, 2015; Demerath *et al.*, 2015; Al Muftah *et al.*, 2016; Mamtani *et al.*, 2016). These are also reported in the large scale meta-analysis of BMI EWAS by (Wahl *et al.*, 2017) which was also able to reproduce the previous associations of blood-lipid associated CpGs to be associated with BMI. Studies looking at waist circumference identified: CPT1A and ABCG1 Wilson *et al.* (2017); Arner *et al.* (2015) And studies into type 2 diabetes identified CpGs in TXNIP (Al Muftah *et al.*, 2016; Florath *et al.*, 2016), ABCG1, SREBF1, PHOSPHO1 and SOCS3 (Kulkarni *et al.*, 2015; Chambers *et al.*, 2015; Dayeh *et al.*, 2016). It is clear that there is considerable overlap between various metabolic traits and the epigenome and it is likely that any additional research will be valuable in the contribution to the current understanding of the epigenetic mechanisms and how they could contribute towards CVDs.

1.6 Aims of this Thesis

1. Chapter 2 explores how quality control of data affects downstream results. In addition to exploring a variety of quality control methods, I develop and test a set of data-driven quality control tools which can be used in conjunction with pre-existing methodologies.

2. Chapter 3 addresses the large memory requirements that are associated with the R programming language which can stifle the analysis of large datasets. By drawing inspiration from the tools that were developed for GWAS, it was possible to extend these frameworks to create a workflow that allows for the low-memory computation of DNA methylation microarray data. This approach can scale into the tens of thousands of samples without leading to any problems associated with memory. The development of these tools is timely as the EPIC array has been released and essentially doubles the size of every dataset going forward. Therefore the improvement on the current methodologies used to quality control, normalise and to perform statistical testing is likely to be well received by the scientific community.

3. Chapter 4 describes an EWAS between DNA methylation and blood-lipid measurements (TC, HDL-C and TG) from a cohort of 1,193 participants from the Understanding Society: UK Household Longitudinal Study which were assayed on the newly released EPIC microarray. This analysis served two purposes, firstly to demonstrate how the tools designed in both Aims 1 and 2 can be used in an EWAS setting. Secondly, to reproduce the findings from past EWAS performed on the 450k and identify novel findings which are unique to the EPIC microarray.

4. Chapter 5 presents two examples of preliminary analyses of tens of thousands of samples that are publicly available can be used to investigate biological questions. Using the software I developed in Aim 2, I demonstrate that large-scale analyses are possible and can be used to produce a number of insightful discoveries.

Chapter 2

Quality control of DNA methylation microarrays

This chapter aims to demonstrate how the quality control steps performed in EWAS can affect downstream results irrespective of the overarching biological question. Recently, there have been frequent calls for a need for "better reporting standards" with respects to quality control (Min *et al.*, 2018). However, there has been a lack of development and demonstration of such tools that contribute towards reproducible quality control. While the auto-generated HTML reports such as those produced by ChAMP or RnBeads are sufficient for diagnosing problems within a dataset. These documents only provide a small description of what steps have been undertaken with respects to quality control and are rarely provided as supplementary material alongside any published EWAS, despite the calls for better reporting standards.

The majority of quality control tools focus on using the inbuilt control probes available on microarrays to identify low-quality samples. As a result, these tools can be prone to identifying samples which may not be outliers. As an alternative to control-probe based methods, data-driven methods can provide a conservative approach to identifying outlying samples and probes. Moreover, a combination of the two methods can have a reasonable impact in reducing the number of false positive findings within EWAS.

To rectify the lack of data-driven methods available in the field of EWAS. I have developed a suite of tools that slot seamlessly into the previously established workflows provided by other software frequently used in EWAS. These tools will allow others to perform both reproducible and robust quality control on their data and will be able to report the exact methods used when presenting their results.

2.1 Introduction

Performing an EWAS requires the well-thought-out design of a study, the collection of biologically relevant samples & data, careful handling of such data, appropriate statistical testing and interpretation of the results. As excellently described in the reviews by Rakyan *et al.* (2011); Mill & Heijmans (2013); Michels *et al.* (2013), considerable thought has been placed into the methodological design of EWAS as a whole. As one of the goals of these reviews was to convey the message that reproducible results ultimately leads to success, many scientists have contributed their expertise towards the development of many software and workflows that help facilitate such reproduction. Conveniently, these software fit into a generalised workflow as described below:

1. **Data Import** – The process of converting the raw data (.idat files) into a biologically interpretable format. During this step, the raw signal intensities from the microarray are converted to methylated and unmethylated signals and then converted into β values.

2. **Quality Control** – The process of removing bad samples and probes. Either through consideration of the control probes on the microarray or through consultation of probe-lists generated by previous studies.

3. **Normalisation** – The process of removing unwanted, systematic, variation between samples. This step is often coupled with additional preprocessing methods such as removal of background noise, adjusting the differences between Type I & Type II probes and correcting positional effects. Typically this is usually applied to the methylated and unmethylated signal intensities and then the β

values are recalculated accordingly. Type I probes were introduced on the 27k microarray Bibikova *et al.* (2009) which required two probes per loci to detect unmethylated or methylated DNA. Type II probes were introduced on the 450k microarray (Bibikova *et al.*, 2011) and only required a single probe to detect methylation state. Because the Type I and Type II probes are functionally different they have slightly different distributions which require seperate normalisation.

4. **Statistical Testing** – The process of obtaining differentially methylated positions (DMPs) or regions (DMRs), usually using linear regression or an ANOVA to test individual probes or clusters of probes

5. **Interpretation** – The process of investigating DMPs and DMRs in the context of the given biological question.

The focus of most development efforts within these five broad steps are towards improving the normalisation and statistical testing of data because they have the most impact on the downstream results. These focused efforts have resulted in dozens of different normalisation methods and a handful of different statistical testing protocols which are useful for EWAS. This focus makes sense as the choice of statistical test and the model used will often define what results are obtained from a study. In a similar aspect, how one chooses to normalise the data will affect the amount of detectable variation there is between samples (Fortin *et al.*, 2014b).

In contrast, there is little development or comprehensive investigation into how data can be quality controlled and how this can affect the downstream results of EWAS. Furthermore, the reporting standards of quality control, compared to normalisation or statistical testing, ranges from sparse to none at all in published studies; leaving many uninformed of the decisions that lead to the exclusion of samples from analysis. Despite the recent calls for better reporting standards (Min *et al.*, 2018), there are still only a small number of examples where the quality control of a dataset is presented alongside the results of an EWAS.

Currently, there are more than two dozen R packages that can be used for the analysis of DNA methylation microarray data. A general summary of these packages are presented in Table 2.1 which describes what functionality is present in some of the more popular R packages used for analysis. However, I feel it is necessary to provide a more detailed description of some of these packages to help contextualise the types of analyses that are available for EWAS.

- **minfi** is the most popular R package used in EWAS. It provides an excellent complement of tools and data-structures that are widely used by many of the other R packages and are also useful for most EWAS in general. Minfi offers two forms of quality control, the first using the shinyMethyl R package and secondly offering a routine quality control pipeline wrapped into a single tool (minfiQC). Minfi offers a large variety of normalisation methods including Illumina, SWAN (Maksimovic *et al.*, 2012), quantile, ssNOOB (Fortin *et al.*, 2016) and functional normalisation (Fortin *et al.*, 2014b).

  Briefly, Illumina normalisation provides background substraction and correction according to control probes. SWAN normalises probes according to technical differences between Type I and Type II probes according to CpG content (e.g. CpG island, Open Sea). Quantile normalisation within minfi first quantile normalises Type II probes and then interpolates the Type II quantiles onto the Type I probes and then normalises Type I probes. Functional normalisation uses the internal control probes to identify technical variation, this approach is more considerate than quantile normalisation as quantile normalisation is known to eliminate variation between samples we would normally expect large variation (e.g cancerous samples compared to healthy samples). Lastly, ssNOOB makes use of the out-of-band intensities from Type I probes to estimate background signal and additionally corrects for dye-bias.

  Minfi also provides functions for statistical testing, namely bumphunting (Jaffe *et al.*, 2012) and block finding (Aryee *et al.*, 2014) alongside the ability to directly annotate to a reference genome.

Table 2.1: Summary of the functionality of a selection of R packages used for the analysis of DNA methylation microarray data

Package	Data Import	QC	Normalisation	Statistical Testing	Reference
minfi	Yes	Yes	Yes	Yes	Aryee *et al.* (2014)
RnBeads	Yes	Yes	Yes	Yes	Assenov *et al.* (2014)
ChAMP	Yes	Yes	Yes	Yes	Morris *et al.* (2014)
wateRmelon	Yes	Yes	Yes	No	Pidsley *et al.* (2013)
MethylAid	No	Yes	No	No	van Iterson *et al.* (2014)
shinyMethyl	No	Yes	No	No	Fortin *et al.* (2014a)
ewastools	Yes	Yes	No	No	Heiss & Just (2018)
methylumi	Yes	No	Yes	No	Triche *et al.* (2013)
MissMethyl	No	No	Yes	Yes	Phipson *et al.* (2015)
sesame	Yes	Yes	Yes	No	Zhou *et al.* (2018)

Additionally, it provides a reference based method for the estimation of cell-type proportions using the method described in (Houseman *et al.*, 2012) for whole blood, cord blood and prefrontal cortex.

- **ChAMP** distinguishes itself from minfi by providing a more comprehensive approach to analysis and a more rigid workflow. In terms of quality control, it automatically removes 'bioinformatically determined' poor-quality probes (Zhou *et al.*, 2017) and produces an HTML report detailing the results of some of the quality control performed on the data. It offers a narrower selection of normalisation methods, limited to Peak-based Correction (Dedeurwaerder *et al.*, 2011), SWAN, Functional Normalisation and BMIQ (Teschendorff *et al.*, 2013) (similar to minfi quantile without the quantile normalisation steps). Within ChAMPs workflow it strongly recommends performing SVA (Leek *et al.*, 2017) and attempts to automatically correct for batch effects. In terms of statistical testing, it also facilitates the detection of DMRs and DMPs with the addition of further examination of results with gene set enrichment analysis.

- **RnBeads** offers a very similar set of tools to what ChAMP, offers but is not limited to the analysis of microarray data. RnBeads allows for the analysis of Bisulfite sequencing data (either RRBS or WGBS). Distinctly, RnBeads allows for the entire analysis, from data import to interpretation of results, to be performed using a single R function which can make it appealing to researchers who may not be bioinformatically inclined. RnBeads does offer more normalisation options than ChAMP allowing it to be a versatile software.

- **wateRmelon** offers a variety of functionality to many other packages through the use of generic methods which apply to any pipeline (e.g. minfi). Conventionally it does not provide any methods for statistical testing as it was designed solely for pre-processing data. In terms of normalisation, it offers many variations of quantile normalisation that are useful in a variety of circumstances. The authors however strongly recommend dasen normalisation for almost all circumstances as it is the most gentle normalisation method provided according to a variety of metrics derived from the microarrays.

- **MethylAid** provides an extensive quality-control pipeline which provides various graphical plots to help visualise how each sample behaves according to various control-probe based metrics van Iterson *et al.* (2014). It fits seamlessly with the minfi workflow as it directly depends on the data-structures provided by minfi. As a result, including it in any analysis that involves minfi is easy. However, it does not support other software which can potentially limit MethylAid's utility. The plots are created using the shiny R package (Chang *et al.*, 2018), which provides a graphical interface for the users to explore the data without the need to create their own plots.

- **shinyMethyl** offers a similar quality control experience to MethylAid with a different variety of plots (Fortin *et al.*, 2014a). shinyMethyl also distinguishes itself from the other R packages as it utilises the shiny R package. These shiny plots differ to how MethylAid presents the quality control plots, but the results are often comparable.

- **ewastools** is a recently published R package focusing on the quality control of data. In contrast to shinyMethyl or MethylAid, the ewastools package uses methods identical to the quality control procedures recommended by Illumina's BeadStudio software (Heiss & Just, 2018). Briefly, it includes over 17 different quality control checks which suggest ewasTools offers a comprehensive set of quality control. In contrast, MethylAid offers five quality control checks. In addition to various quality control, ewastools also offer methods to identify mislabelled samples through SNP agreements and sex checks.

- The **methylumi** R package - introduced the concept of out-of-band (OOB) background correction Triche *et al.* (2013) which makes use of the Type I probe signals ignored in regular analysis and uses these OOB signals as a function of background noise instead of the control probes dedicated to detect background signal. Although it has not been functionally updated to handle the new EPIC arrays, the data import function provided in methylumi has been extended in wateRmelon (through my own efforts) and other functionality has been assimilated into minfi (e.g. ssNOOB normalisation).

- The **MissMethyl** R package (Phipson *et al.*, 2015) provides another workflow much like the ones provided by ChAMP and RnBeads. Notably, the distinction of the missMethyl package is that it provides a new way to test differential variation (DiffVar) using a Levene's test and also provides gene set analysis which is only provided thus far by ChAMP.

- A recently published package: **sesame** seeks to remove signals from the DNA methylation microarrays which are caused by genomic deletions within individuals which can confound with the results of EWAS (Zhou *et al.*, 2018). These artefacts are corrected using the new normalisation method (pOOBah) which masks the artefacts while maintaining the biological variation that is of interest between samples. In addition to providing a new analysis technique, the authors provide a fully functional preprocessing platform that is quite versatile.

This summary shows that there is a wide range of software available for the analysis of the DNA methylation data. Each aspect of the general workflow described earlier appear to be well represented with a lot of focus being placed on the normalisation and statistical testing of data. Many R packages do provide some form of quality control. However, these quality control tools make decisions based on the control probes that are available on each microarray which may not be an accurate indication of the true quality of a sample. MethylAid and shinyMethyl both offer interactive GUIs for the exploration of data that has been imported by minfi. Ewastools offers a relatively simple set of quality control that can only be applied to data read in by the ewastools package and therefore not immediately applicable to other workflows. ChAMP and RnBeads both offer HTML reports with vastly similar quality control procedures to that of MethylAid or shinyMethyl but are also limited to their respective workflows.

All of these methods focus on control probes and therefore can be considered as one dimensional because they do not examine the aspect of the data that is being implicitly tested (the β values). Therefore an opportunity to develop data-driven based quality control methods presents itself. Data-driven tools are useful because they consider the entire complement of features that are present on the array instead of examining a small selection of probes.

Here I introduce some tools to ameliorate this one-dimensional nature of the quality control tools available. These tools are packaged within the latest version of the wateRmelon R package and nearly all of the tools can be used on any resultant β matrix produced by minfi, ChAMP, RnBeads or other software. The tools I describe here include: outlyx (a robust outlier detection method), bscon (a fast tool that checks the control probes to estimate the quality of DNA applied to the microarray), pwod (an outlier detection tool that checks each probe separately) and qual (an experimental outlier detection tool that considers the degree of transformation a sample undergoes during normalisation). In addition to the quality control tools, I extend the functionality wateRmelon by including accessory functions that include a data import function (readEPIC) capable of reading in EPIC array data, sex prediction (predictSex), age prediction (agep) and cell type composition estimators to bring wateRmelon up to speed with other popular R packages.

The aim of this chapter is to determine whether or not rigorous quality control of data leads to a decrease in test-statistic inflation. To properly examine this aim I first need to answer a few questions: What is rigorous quality control? What is test statistic inflation and how do we measure it? And how does quality control affect downstream results?

2.1.1 What is rigorous quality control?

One may argue that a well thought out pipeline that arbitrarily removes low-quality samples and questionable probes is sufficient for most datasets. For the most part, such an approach would not come under much scrutiny, providing it is reported. However, simply looking at the control probes and or the median signal intensities may not convey the full picture with respects to a samples outlying nature. The quality of a sample can be questionable according to control probes but have a perfectly reasonable β distribution. Likewise, a sample can have a wildly erratic β distribution but can look perfectly normal when considering the control probes. Because of this nuance, it is likely better to use a comprehensive approach, one that uses both control-probe and data-driven methods, when quality controlling data instead of relying on a single aspect of quality control.

Another well-reported form of quality control in EWAS is probe filtering. This process involves the removal of features (probes) from analysis that are determined to have a poor signal, either through tools such as pfilter (Pidsley *et al.*, 2013) or from lists of probes that are known to cross-hybridise or are affected by underlying SNPs in the probe sequence. Such probe lists are available for both the 450K (Zhou *et al.*, 2017) and EPIC microarrays (Pidsley *et al.*, 2016) and used by default in some R package workflows (e.g. ChAMP).

It is important to remember that quality control is not just limited to identifying outliers and filtering problematic probes. It also includes checking the sanity of the data such as identifying mislabelled samples, checking for familial relationships and looking for potential batch related problems are also necessary for comprehensive quality control. These sanity checks can be performed using specific software (e.g. omicsPrint (Van Iterson *et al.*, 2018) or ewastools) or manually by inspecting principal components or making multidimensional scaling plots.

2.1.2 What is test statistic inflation

Test statistic inflation is a phenomenon that has affected GWAS and EWAS for a long time. In genome-wide analyses where hundreds of thousands of statistical tests are being performed, it becomes necessary to adjust the test-statistics to satisfy the multiple testing thresholds. Often this adjusted is done by dividing the test statistics by the number of tests being performed (Bonferroni Correction) or by converting test statistics into a false discovery rate such as the method suggested by Benjamini & Hochberg (1995). Test statistic inflation can be observed, even after this adjustment, when the observed number of significant results obtained is greater than the expected number of significant results. This increase suggests that there is a bias towards the lower tail of p-values within a given set of results therefore indicating the potential for false positives.

In GWAS this inflation is caused by population structure or cryptic relatedness, where immeasurable sub-populations within sample groups drive variance towards spurious associations. A popular remedy

for inflation in GWAS is to divide the test statistics prior to multiple testing adjustment by an inflation factor λ_{GC}. This factor is defined as the ratio between the median test statistic and the median expected test statistic derived from an empirical null of a χ^2 distribution equivalent to a value of 0.456 (Devlin & Roeder, 1999; Devlin *et al.*, 2001). Dividing test-statistics by this inflation factor increases (providing that $\lambda_{GC} > 1$) the observed p-values such that the corrected p-values are comparable to the expected distribution of p-values (the empirical null distribution). This approach works in these scenarios as it will doctor the upper-tail of the observed distribution (providing that $\lambda_{GC} > 1$)

In EWAS the source of this inflation is more complicated. Due to the nature of the epigenome and how it is measured, epigenetic data is subjected to many sources of confounding including age, environment, cell heterogeneity and batch effects. Additionally it had been shown that λ_{GC} is limited when the trait being investigate is associated with multiple small effects (Voorman *et al.*, 2011) which is usually the case in EWAS. To remedy these additional sources of confounding van Iterson *et al.* (2017) proposed a novel method of estimating λ for EWAS (henceforth referred to as λ_{bacon}). Distinctly, the λ_{bacon} method implements a Bayesian outlier model which assumes that there are a small number of genuinely associated findings and calculates the inflation factor with these findings excluded. This means that λ_{bacon} is mostly independent of the relatively few associations that an EWAS generates. In addition to computing the inflation factor, van Iterson *et al.* (2017) also make the argument that the test statistics produced by EWAS are also subject to a bias which leads to a shift in the distribution of effect sizes. Despite such development, the practice of controlling for test statistic inflation in EWAS is not yet commonplace. Nonetheless, it is possible to use either λ_{GC} or λ_{bacon} as a way to estimate the amount of test statistic inflation that exists in any given analysis.

When trying to account for test statistic inflation, one should try their best to remove batch effects, outliers and other sources of confounding. While appropriate quality control should effectively handle outliers, efforts have been made to reduce the other sources of inflation. Correcting for batch effects can be done using ComBat (Johnson *et al.*, 2007) or by including experimental variables as covariates (e.g. Slide and plate numbers). Removal of unwanted variations can be done using sva (Leek *et al.*, 2017), ruv (Gagnon-Bartsch, 2018) or CATE (Wang & Zhao, 2015) but have been demonstrated not to be effective

in removing all confounding (van Iterson *et al.*, 2017). Alternative workflows such as ChAMP or the workflow proposed by Lehne *et al.* (2015) suggest including various numbers of principal components in the model which is another strategy that is employed by GWAS.

2.1.3 Does meaningful quality control decrease test statistic inflation

Given that test statistic inflation in the context of EWAS is driven by numerous sources, it is reasonable to assume that the application of a quality control method can decrease the test statistic inflation and improve genuine signals. In the example of ewastools, Heiss & Just (2018) look at a wide variety of datasets but limit their analysis to only describe the number of samples that fail quality control measures (reflective of Illumina's suggestions). Heiss & Just (2018) identify numerous samples in each dataset which suggests that a majority of the datasets could benefit from quality control prior to analysis. Therefore a reasonable extension of these analyses would be to consider how different quality control methods differ in the number of samples identified and how the removal of flagged samples can affect down-stream results. Although I am limited by the fact the deposited data may not have all of the required information to reproduce the exact analysis used in the original studies, I should be able to perform analyses to a reasonable standard while comparing three different quality control methods (wateRmelon, MethylAid and ewastools).

2.2 Methods

There are many publicly available datasets and also a large number of quality control pipelines I can examine. Therefore, I decided to select three quality control pipelines that allow for the precise control of other features of analysis (e.g. normalisation and statistical testing). By comparing the differences in the number of samples flagged by each method and by performing the same statistical analysis, I will be able to examine how different quality control methods fare in reducing test statistic inflation.

I decided to compare the MethylAid package (an interactive GUI quality control R package) which also

includes a part of minfis quality control pipeline, ewastools (a command-line tool derivative of the BeadStudio software of Illumina) and wateRmelon (the data-driven tools I will describe in this study). The reasoning behind choosing to focus on both MethylAid and ewastools are dedicated software for the quality control of data and provide will perform functionality similar to that of most software such as ChAMP or RnBeads.

In MethylAid, outliers are determined based on five checks and additional between-sample checks can be manually verified. MethylAid's outlier tests include looking at the median signal intensities, bisulfite conversion efficiency, the overall quality of samples in both sample dependent and independent control probes (using the non-polymorphic and hybridisation probes) and a measure of background noise on a per sample basis according to the negative control probes. Ewastools use these same quality control probes but tests each sample based on the guidelines suggested by Illumina. In total, ewastools uses 17 different metrics to quality control samples.

Despite the large selection of quality control methods available, none of these are particularly data driven and therefore the opportunity to develop and present some data-driven tools is described here. Data-driven methods for detecting outliers are attractive as it is entirely possible for a potentially outlying sample to appear completely normal or even well-performing according to the control probes. Conversely, samples flagged by control probe methods may appear to fail but will otherwise generate a usable signal and thus could lead to removing data that otherwise do not need removal.

One data-driven approach can include checking the distribution of β values on a per sample basis and remove samples that have a distinctly different shape (See Figure 2.1 for example). Another method for outlier detection can also include the use of dimensional reduction techniques such as principal components analysis or multidimensional scaling and plotting two of the dimensions in a scatter plot. Both methods are particularly useful as they are quite robust and can identify samples which are considerably different from the rest of the data. However, both of these methods are not reproducible as they require manual selection and verification to determine outliers.

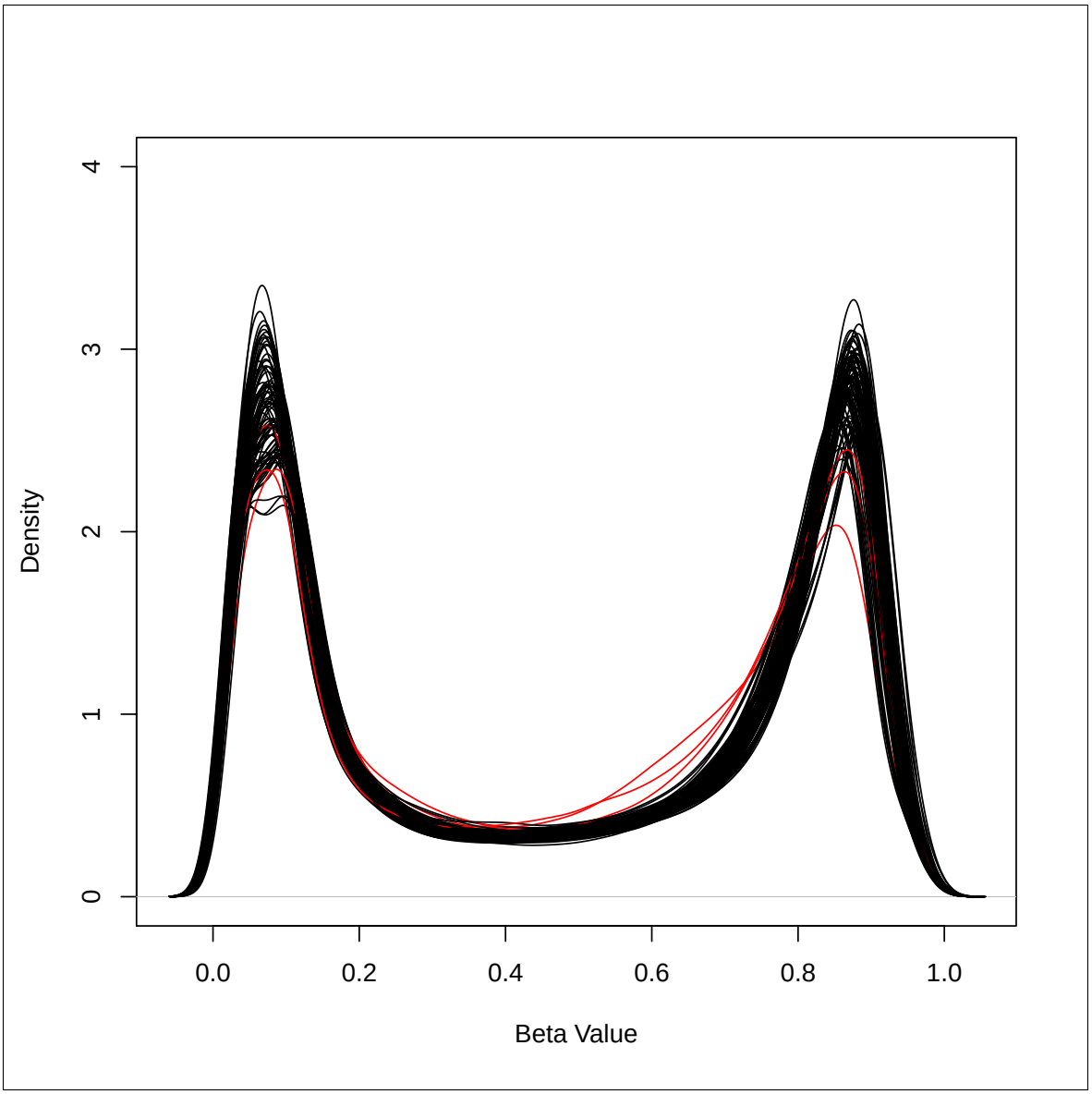


Figure 2.1: Example of using β distributions to identify outliers. Potential outliers are coloured in red and are characterised by lower peaks at β values of 0 and 1 alongside an increased band around a β of 0.5.

The outlyx function uses both principal components analysis and Mahalanobis distances to identify outliers within datasets. By default, it considers the first principal component as it is the largest source of variance within DNA methylation data. The Mahalanobis distances are computed by using the pcout function from mvoutlier package (Filzmoser *et al.*, 2008). Samples are flagged based on how far they away from the bulk of the data in terms of interquartile range and also based on the final weight according to the pcout function. The final weight ranges from 0 to 1 where values $<$0.25 are considered outliers by the original authors, this final weight is computed from two other metrics and a value of $<$0.25 is only ever acheived if a sample scores a 0 in atleast 1 of the two tests (Filzmoser *et al.*, 2008). Overall, outlyx is fairly conservative (depending on the thresholds chosen) as only samples that fail both tests will be considered outlying. This approach yields a robust and reproducible method that is both easily interpreted using outlyx's inbuilt plotting functionality and is unaffected by both swamping and masking effects.

The next tool, bscon, functions similarly to how the other control probe based metrics perform. The distinct difference is that bscon translates the signal from bisulfite conversion probes into a percentage. This distinction allows for a more natural interpretation of sample quality compared to other quality control metrics based on bisulfite conversion probes (e.g. MethylAid's raw signal method and ewastool's ratio method (Unmethylated Intensity/ Methylated Intensity)). As a result, it is possible to set meaningful thresholds (e.g. $< 80\%$ bisulfite conversion) to screen for outliers.

Qual seeks to exploit an area of quality control that is almost always overlooked in EWAS and can be important in diagnosing downstream problems with samples that may not present themselves as outlying by the current quality control tools. The thought process behind qual is to measure the amount of change a sample requires to fit with the expected distribution following normalisation. As normalisation attempts to correct for systematic variation, it is possible that it drastically changes the raw signals of a sample such that it fits with the remaining dataset. This 'violence' can be explained with the root mean square (RMSD) and the standard deviation (SDD) of the difference ($\Delta\beta$) between the normalised and raw β values for a given sample, $\Delta\beta_j = \beta_j^{Normalised} - \beta_j^{Raw}$. Samples that are subjected to a low degree of violence are expected to have small RMSD and SDD values, likewise a sample that has been subjected to a high degree of violence will have high RMSD and SDDs. The effectiveness of qual is dependent on

both the normalisation method used and whether or not the application of normalisation was meaningful. A meaningless normalisation for example could involve normalising two datasets together from different tissues, where the differences are always going to be very large.

Lastly, I suggest another probe filtering method that does not aim to remove testable probes from analysis but rather prune the individual probes for outliers within their own β distribution which may be caused by SNPs that may not be accounted for according to existing probe filtering methods. The aptly name tool, pwod ('p'robe-'w'ise 'o'utlier 'd'etection), identifies probe-level outliers based on signals that lie outside of more than 4 interquartile ranges from the upper and lower quartiles of each single probes distribution. The extremely conservative threshold is used to ensure that only obvious probe-level outliers are removed.

These tools: outlyx, bscon, qual and pwod - make up a thoughtful quality control pipeline that considers many aspects of quality control to remove samples from analysis. These tools can be used by themselves or as part of a preexisting methodology and are highly reproducible as the results and reporting of these tools are straightforward.

To test the usefulness of these tools, they were compared to the quality control tools provided by MethylAid and ewastools as they are representative of the other quality control pipelines that are available. The datasets used to test the methods are detailed below, but the general analysis of each dataset is as follows. All data was read in using the readEPIC function to parse idat files into methylumiset objects. Datasets were quality controlled according to the default parameters for each method. For both MethylAid and ewastools, any flagged samples were removed from the dataset then low-quality probes identified by pfilter were removed from analysis. Data was then normalised using dasen normalisation. For watermelon, outliers from bscon and outlyx were removed prior to dasen normalisation, samples flagged by qual were then removed followed by pfilter then the raw methylated and unmethylated intensities (with samples identified by qual removed) was normalised using dasen one more dasen normalisation again. After normalisation all quality control pipelines were additionally subjected to pwod.

2.2.1 Datasets

The datasets used for this analysis were obtained from GEO and an additional 1,193 samples were used from the Understanding Society UK household longitudinal study which were assayed on the EPIC array. Despite initially looking at over 70 data-sets, only a selection of these were used for further analysis due to lack of information or inappropriate annotations to produce a sensible model or to reproduce the model the original authors had described in their analysis.

2.2.2 Measuring Test Statistic inflation

Test statistic inflation was quantified as the genomic inflation factor λ_{GC}. Briefly, this is derived as the median observed test-statistic divided by the median expected value from an empirical null distribution, computed from a χ^2 distribution. For this study I will only be using λ_{GC} as a measure of inflation and will not be dividing the test-statistics by the factor prior to calculating the number of bonferroni significant hits. As I also intend to compare the number of bonferroni significant hits between the methods the number of hits identify before and after genomic control should be comparable between methods.

2.2.3 Statistical Analysis

The datasets that were selected for statistical testing attempted to follow the model used by the original author however the quality control and normalisation method used in the original studies were not used. Because of this, the results from these reproduction analyses may not be exact reproductions due to missing variables and difference in upstream processing. All models included age, sex, slide number and array position as covariates with cell-type composition estimates included when the samples were obtained from whole blood. Each model (for each dataset) will be run for each method of quality control (no quality control, wateRmelon, MethylAid and ewastools). Test statistic inflation and the number of genome-wide significant results will be compared across all tests to examine how each quality control method affects downstream results.

2.3 Results

To ascertain how the tools described in this study can be used in an analysis it is useful to demonstrate how each tool functions Firstly, outlyx is the first data-driven outlier detection tool that uses dimensional reduction techniques to identify outliers according to two separate tests. By using two tests it is a both highly robust and conservative that is not affected by swamping and masking effects. As seen in Figure 2.2 the outlyx function provides an inbuilt plotting function which provides users with a useful plot that demonstrates how each sample looks with respects to the rest of the data. In general, the second test (using Mahalanobis distances) appears to be more likely to identify outliers. By using two tests it is possible to ensure that only genuinely outlying samples are identified. This approach does require a considerable amount of computing resources to produce results however it can be sped up at the cost of some accuracy by using a smaller subset of probes.

In regards to which thresholds would be most appropriate for data I observe that the final weight produced by the pcout function tends to score samples poorly quite frequently. Due to how the pcout function is designed a weighted score of $<$0.20 is acheived when a sample scores a value of 0 in either of the two tests that the pcout function uses. This can be seen in Figure 2.4 and Figure 2.3b where there is a distinct cluster or change in shape of the slope at a value of 0.20. To avoid this I chose a value of 0.15 to select samples which are distinctly different from the bulk of the data. Despite this a threshold of 0.15 will still select approximately 24% of samples (Figure 2.4). By using the interquartile based method ($>$2 IQRs) in addition to the mahalanobis method we manage to flag $<$1% of samples to be outliers and may warrant the removal from analysis as the IQR method is indeed the limiting factor for outlyx where 2 IQRs only flags around 1.5% of samples (Figure 2.4 and Figure 2.3a).

The bscon tool works similarly to all the other control probe based metrics described in this study. One distinction with bscon is that the output is in the form of a percentage. This makes it incredibly easy to make an assumption about the quality of the sample. In general, the tool will output a value of around 95% if the sample is of very good quality (A score of 100% is unlikely due to background noise). From examination of all the data analysed in Table 2.2 we can see that appoximately 90% of all samples have a bscon value $>$80% (Figure 2.5 and Figure 2.6). Although a low bisulfite conversion value may determine

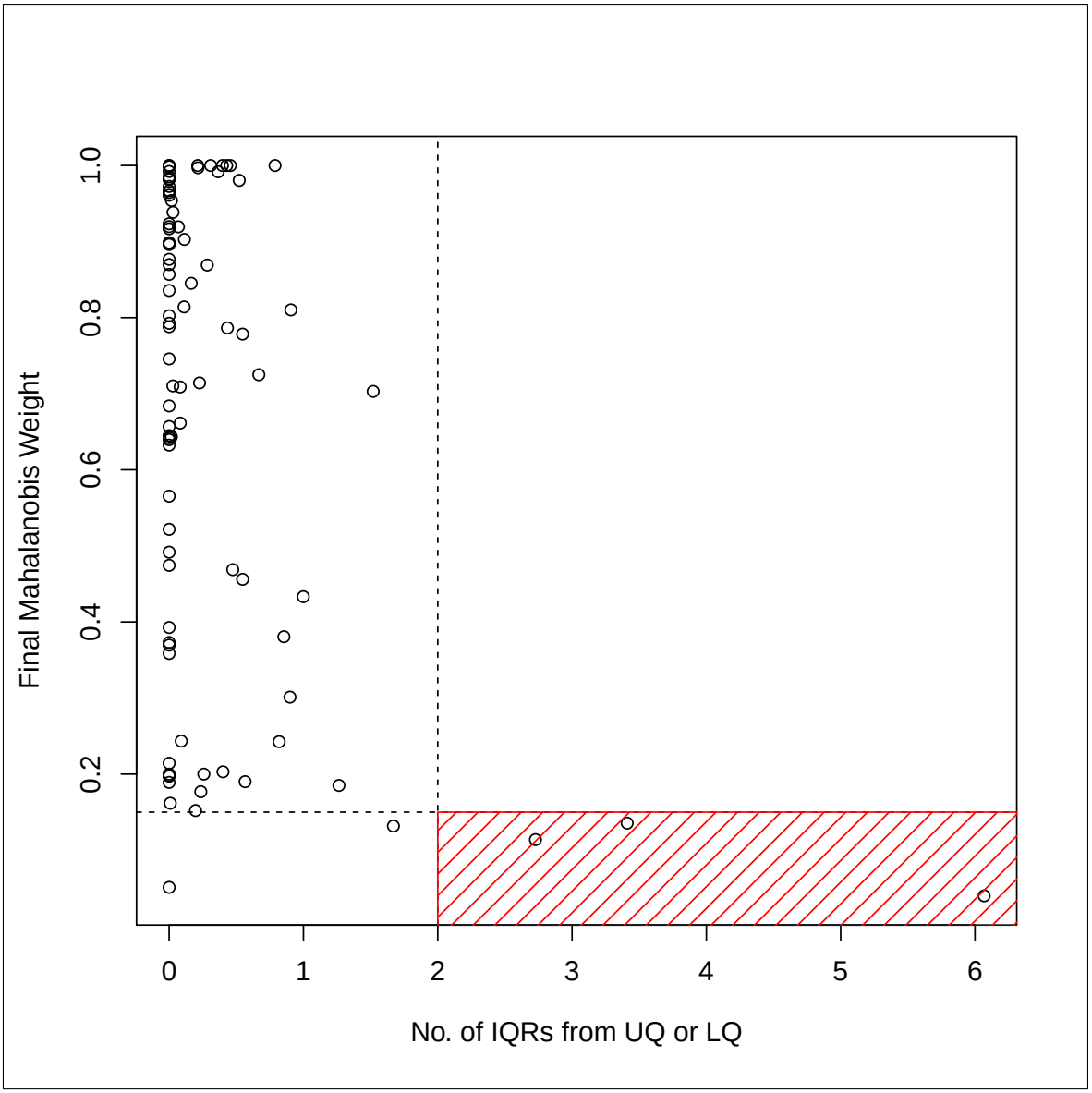


Figure 2.2: Example of the output from the outlyx tool. Performed on 93 samples obtained from DNA from cerebellum brain tissue assayed on the 450k microarray. No. of IQRs from Upper or Lower quantiles are calculated from loading values from PC1. Final Mahalanobis Weight is derived from a modified version of the pcout function. Outliers can are determined as outlying based on where each point (a single sample) lies on the plotting area. Samples within the red squares are considered outlying.

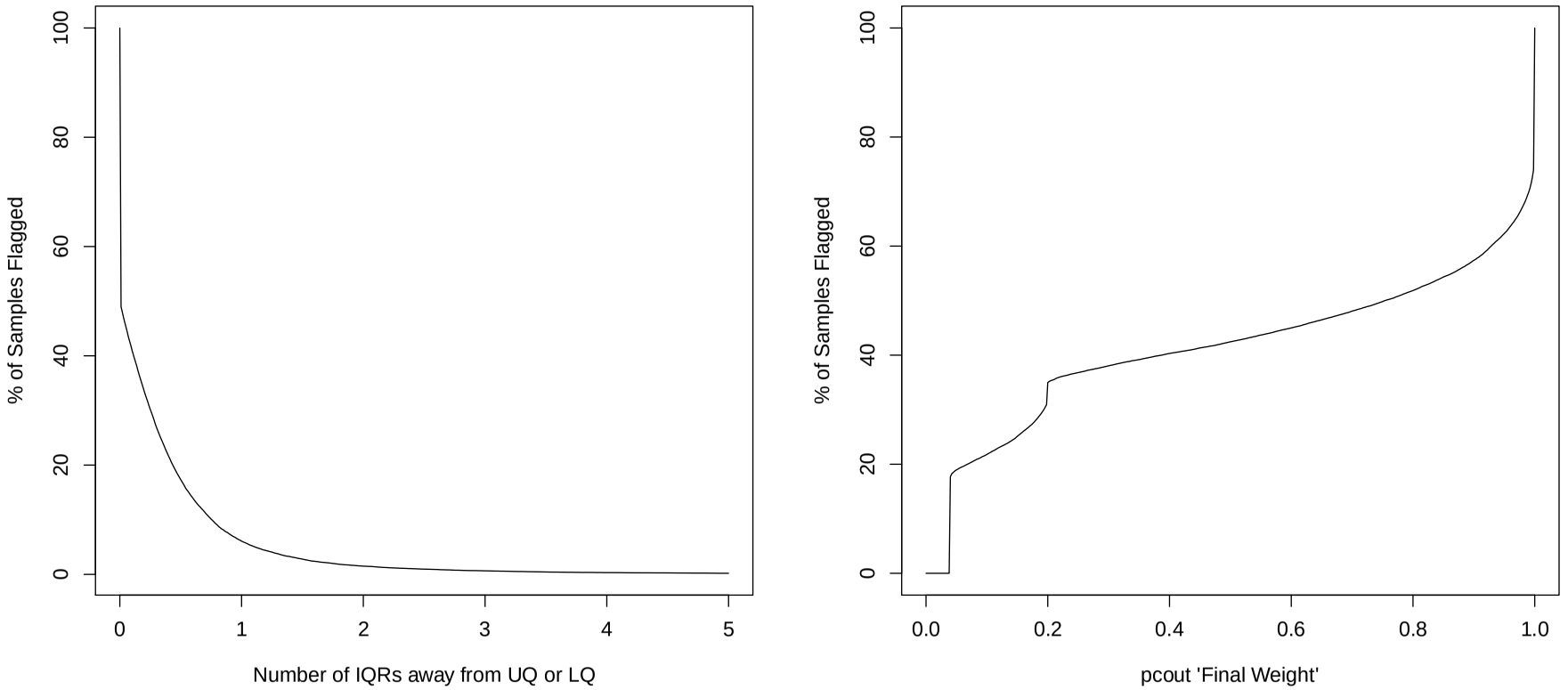


Figure 2.3: Distributions of the number of samples flagged by outlyx using at different thresholds of **(a)** Number of IQRs away from upper or lower quantiles and **(b)** Mahalanobis Final Weight when performed on each dataset described in Table 2.2.

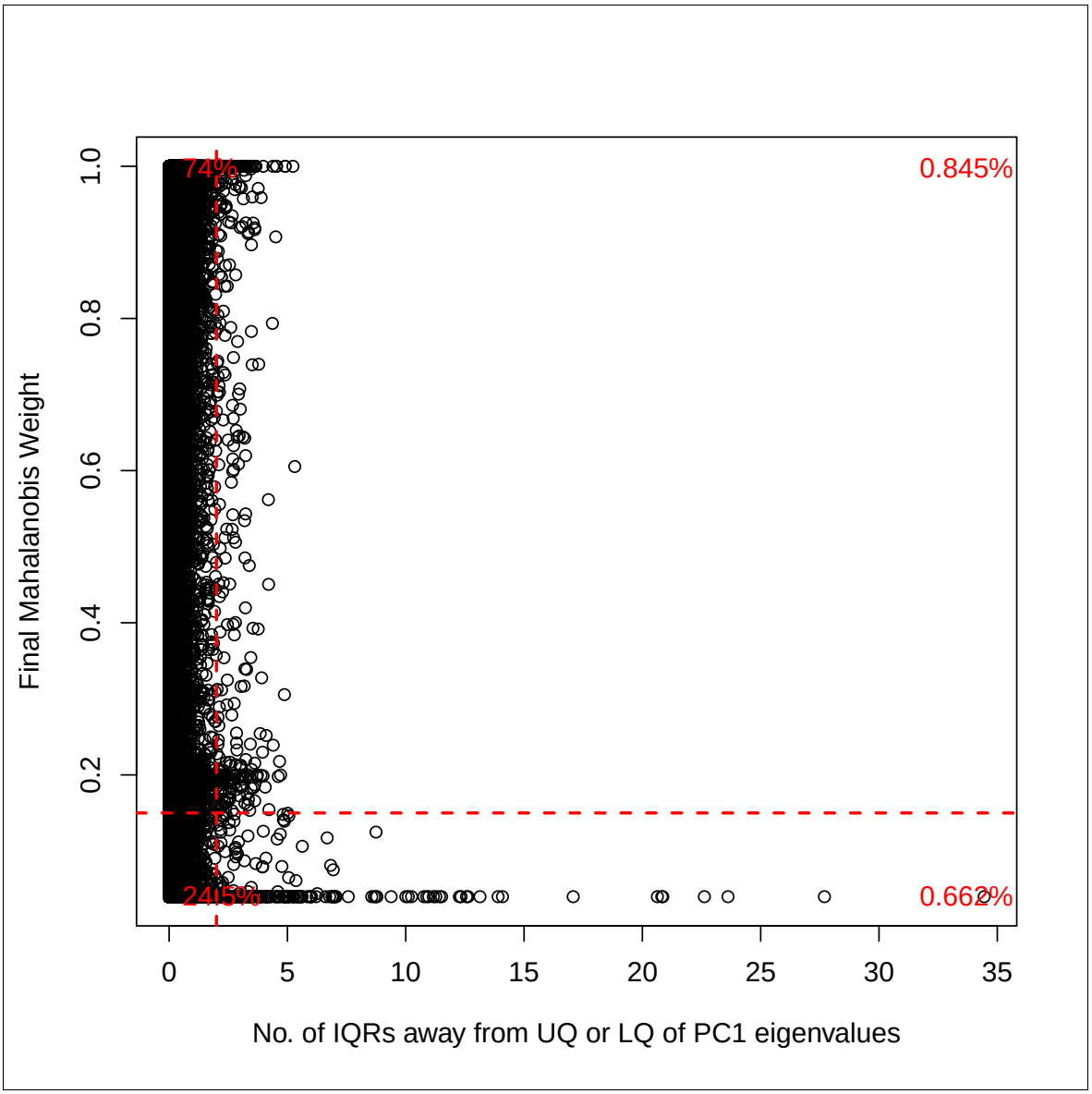


Figure 2.4: Overall number of outliers detected by outlyx when applied to the numerous datasets described in Table 2.2. Red-dashed lines indicate thresholds of 2 IQRs and a final weight of <0.15.

if a sample is outlying a low bisulfite conversion value may indicate potential lab problems which should be looked into further.

Qual is an interesting tool that has the potential to be a useful tool in the bioinformaticians toolbox. Qual measures the degree of normalisation violence that a sample undergoes during normalisation. This is often an unaddressed area of EWAS as the assumption is that the data after normalisation has had most of the technical problems have now been accounted for. Out of all the software available only RnBeads offers similar functionality but do not expand on it in any way. Here we characterise this violence using the RMSD and SDD of the difference between normalised and raw values on a per sample basis. This allows relatively easy interpretation, by plotting the two metrics in a scatter plot (Figure 2.7) will provide a pattern where samples that change very little, cluster around the origin while samples that have had a lot of violence are further away. From examination of various thresholds (Figure 2.8) we see that at a value of around 0.05 there is a sharp elbow in the number of samples flagged by both RMSD and SDD. I found that a value of > 0.05 for both RMSD and SDD appear to capture around 5% of samples (Figure 2.9) and serves to function as a conservative threshold. However, whether samples should be removed from analysis or a more suitable normalisation method should be used depends entirely on the research question.

In addition to looking for outliers on the sample level, it is possible to identify individual outlying signals on the probe level. In most circumstances, the β distribution of a single probe has a unimodal distribution except for probes that have a SNP underlying the probe sequence which can yield a multimodal distribution. These probes that could exhibit multimodal distributions are usually removed form analysis as they are often identified in the probe lists. In the probes that are considered normal and unimodally distributed there is not a consistent check to identify whether or not there are any outliers on the probe level. Indeed it is difficult to ascertain why a signal on a probe would be extremely different from the rest of the data but, it is thought that either SNP heterozygotes or minor allele frequencies are a likely cause of this. To aid in this, pwod considers the quartiles of each probe and removes any signal that is more than 4 IQRs away from the upper and lower quartiles (Figure 2.10). These large boundaries are to ensure that no genuine variations caused by the experiment design would influence as even the largest effect sizes seen in studies only contributes to a minor difference in methylation between groups.

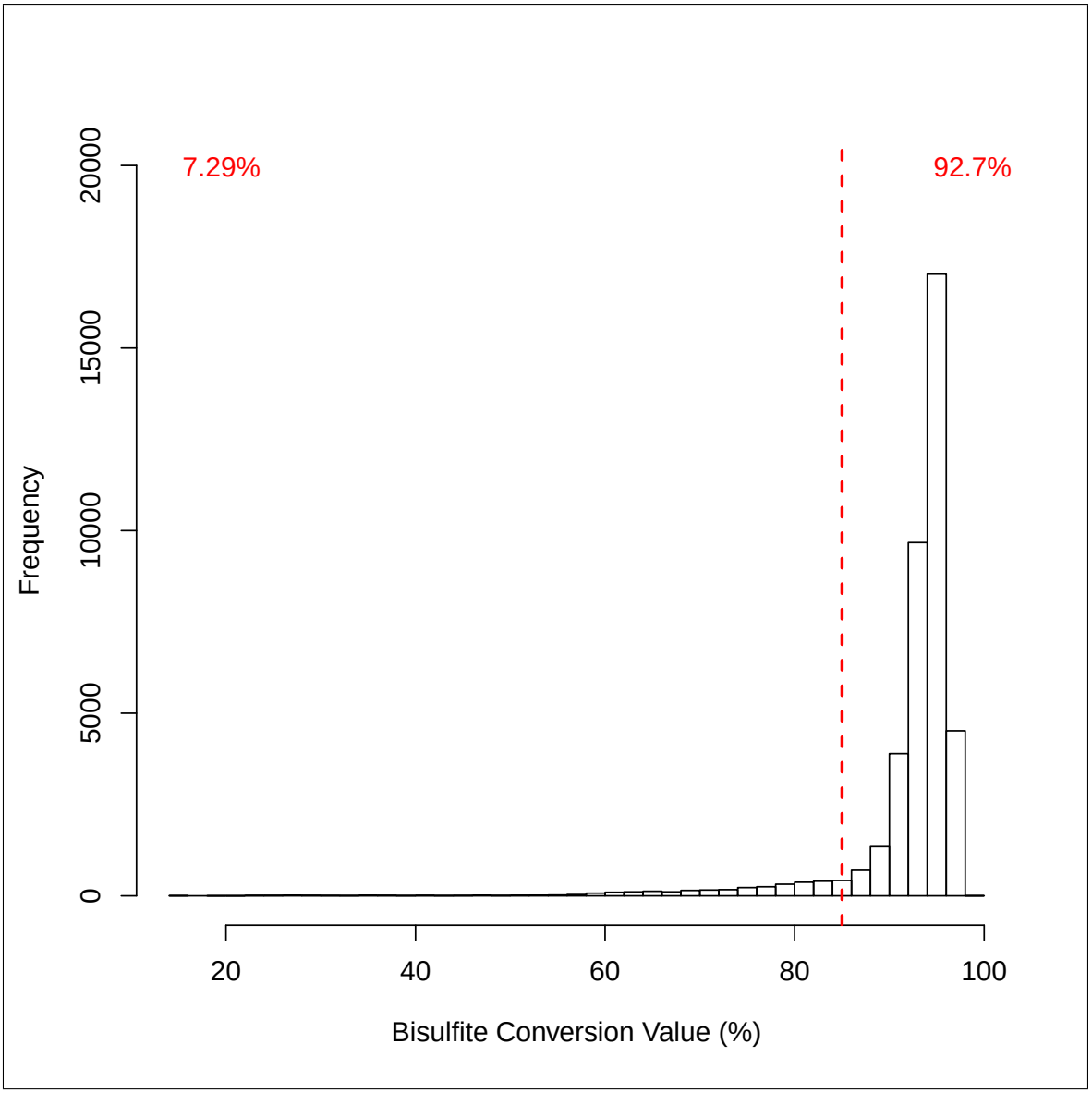


Figure 2.5: Example usage of the bscon function performed numerous datasets described in described in Table 2.2. Computation of the bisulfite conversion percentage per sample using bscon allows for easy determination of low-quality samples. In general samples of good quality will have a bisulfite conversion rate of $\sim$ 95% depending on the source of DNA. Samples that have a bisulfite conversion percentage less than a certain threshold (shown here as 80%) should be considered as outlying and removed.

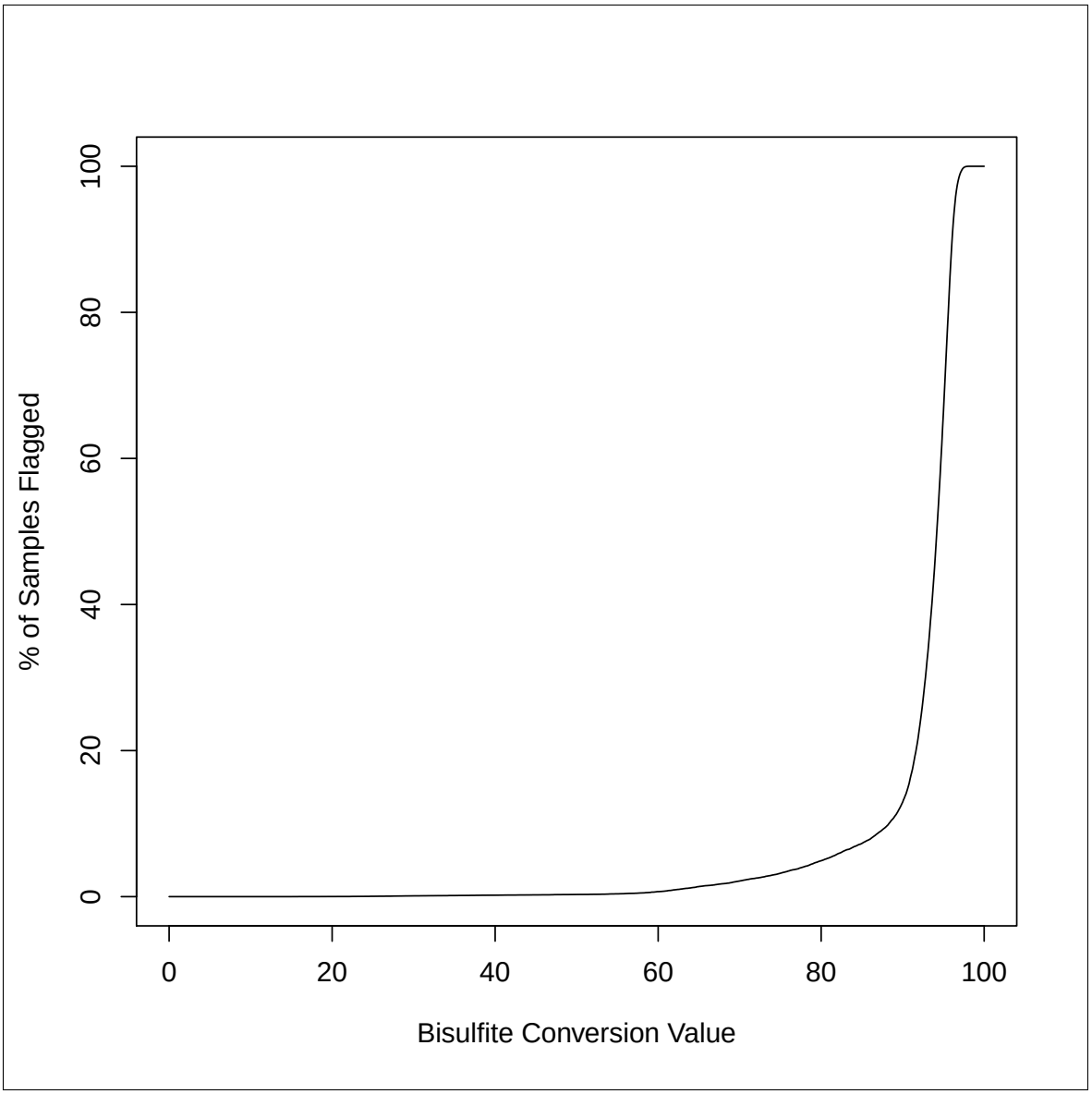


Figure 2.6: Number of samples flagged by different thresholds of bscon when applied to numerous datasets described in described in Table 2.2.

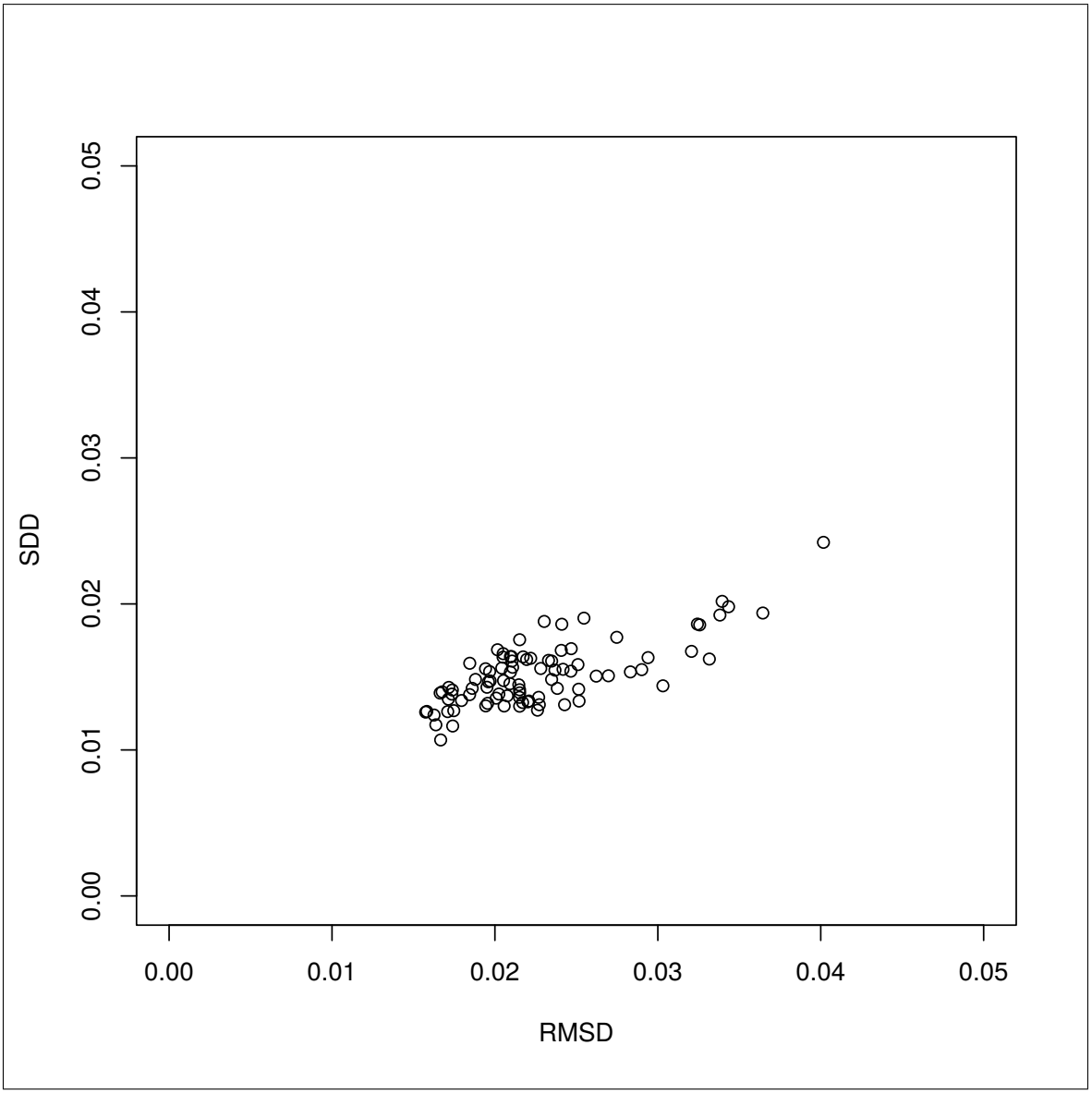


Figure 2.7: Example usage of the qual function from wateRmelon. Performed on 93 samples obtained from DNA from cerebellum brain tissue assayed on the 450k microarray, normalised using the dasen method. Differences between normalised and raw betas values are characterised using two metrics (RMSD and SDD), samples that undergo the largest amount of change are characterised by having a large RMSD and SDD (see Figure 2.9 for examples), in general samples that have a RMSD and SDD > 0.05 should be treated carefully.

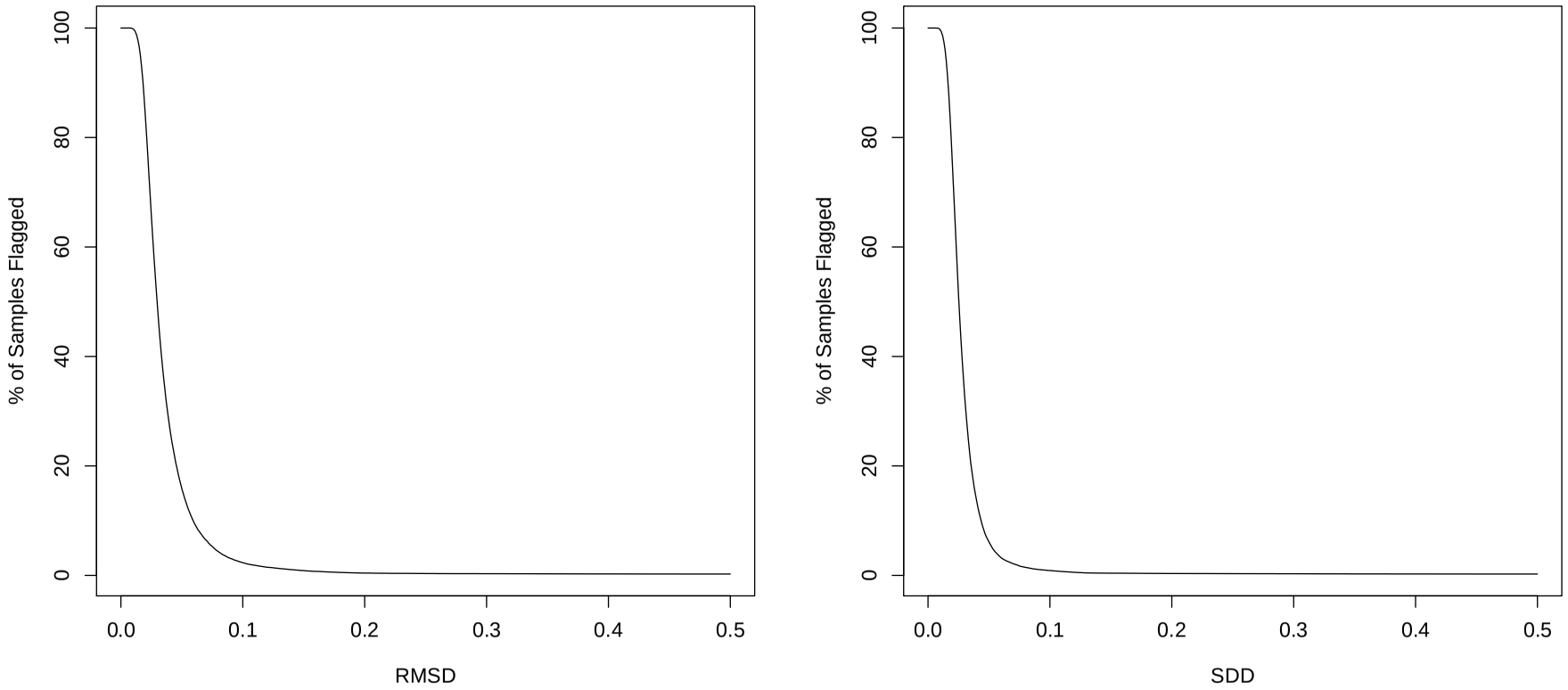


Figure 2.8: Distributions of the number of samples flagged by qual when applied to numerous datasets described in described in Table 2.2 using different thresholds of **(a)** RMSD and **(b)** SDD

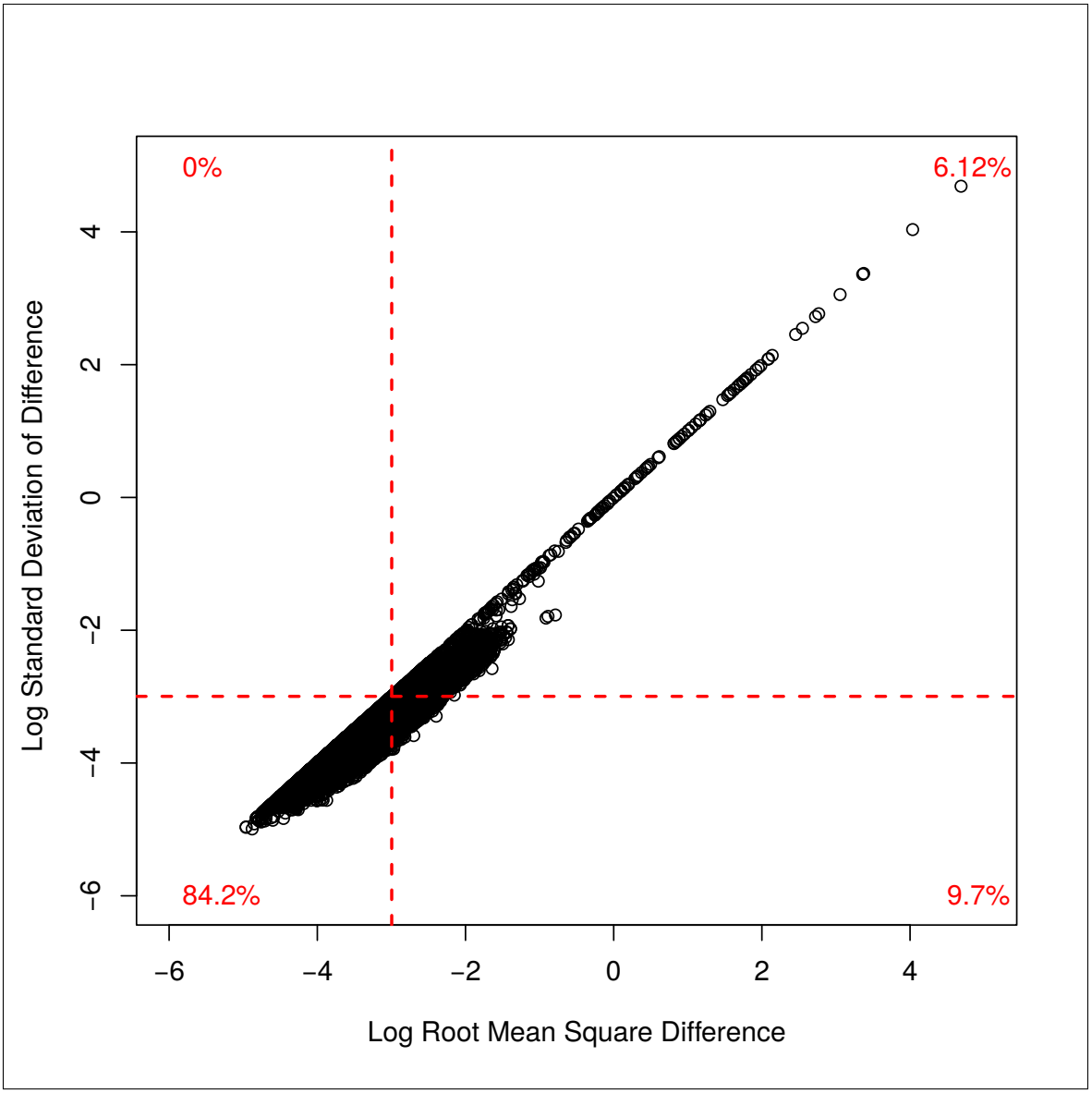


Figure 2.9: Scatter plot of the outputs of qual (RMSD) and (SDD) when applied to numerous datasets described in described in Table 2.2. Outputs are log-transformed due to some RMSD and SDD values being exceedingly large. Red dashed lines represent RMSD and SDD thresolds of 0.05 e^{-3}. Each quadrant are described by the percentage of samples that exist within that area. For example, 6.81% of samples would be flagged by the qual tool when using the value of 0.05 as a threshold.

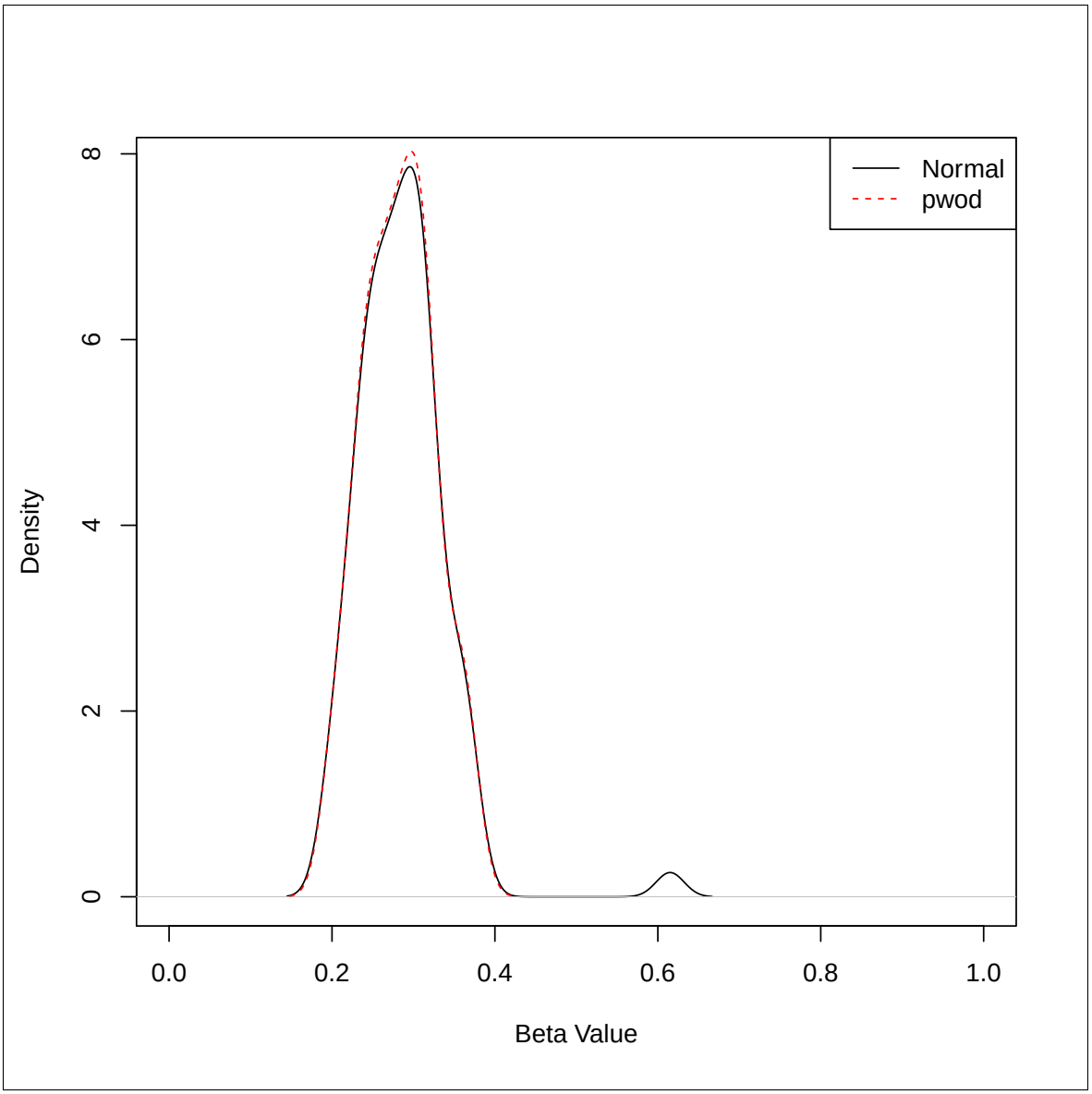


Figure 2.10: Density plot of a single CpG site describing the differences between raw (black) and pwod-treated data (red). It can be seen that the distribution of both raw and pwod beta values are identical except for a single, outlying signal lying far from the mean of the bulk distribution. This process is applied to every probe available in the data-set.

To assess how each of the quality control workflow performs I first determined how many samples were flagged as outlying within each dataset according to the three methods being examined. To do this, I used over 70 datasets that were publicly available on GEO. These datasets were assayed on the 450K microarray except for the large dataset assayed on the newly released EPIC microarray (UKHLS). Each dataset was quality controlled using the default thresholds as described by each workflow.

Table 2.2: Summary of the number of flagged samples using default thresholds by ewastools (with Non-Polymorphic probes), MethylAid and wateRmelon (with bscon)

Dataset	No. Samples	ewastools (+ NP probes)	Methylaid	wateRmelon (+ bscon)
GSE42861	689	2 (16)	0	4 (5)
GSE43976	95	0 (8)	0	0 (0)
GSE51032	845	3 (91)	204	77 (79)
GSE51057	329	0 (15)	5	0 (2)
GSE52980	13	0 (1)	0	0 (10)
GSE55491	24	0 (0)	0	1 (1)
GSE59065	296	0 (7)	134	5 (5)
GSE59524	24	0 (0)	0	0 (0)
GSE60655	36	0 (0)	0	1 (3)
GSE61107	48	0 (1)	3	0 (1)
GSE61279	110	82 (82)	55	4 (80)
GSE61454	269	129 (248)	188	3 (85)
GSE61496	312	8 (26)	4	4 (6)
GSE62219	60	9 (48)	55	7 (13)
GSE63106	62	0 (0)	0	0 (0)
GSE63695	97	0 (39)	40	47 (39)
GSE65058	24	2 (2)	2	0 (0)
GSE65638	16	0 (0)	0	0 (0)
GSE66459	22	0 (0)	0	0 (0)
GSE67393	117	10 (29)	64	3 (11)
GSE67419	24	0 (0)	0	0 (0)
GSE67444	70	4 (22)	12	0 (0)
GSE68777	40	5 (13)	33	0 (2)
GSE69852	6	0 (0)	0	0 (0)
GSE70478	38	0 (0)	1	2 (2)
GSE72120	72	0 (5)	25	4 (4)
GSE72354	34	14 (14)	0	0 (0)
GSE72364	12	0 (0)	0	0 (11)
GSE72556	96	25 (51)	70	12 (15)
GSE73115	180	2 (6)	0	0 (0)
GSE73412	74	12 (12)	7	1 (1)
GSE73626	18	0 (1)	1	0 (0)
GSE73745	24	0 (0)	0	0 (2)
GSE74432	122	0 (0)	8	13 (13)
GSE74548	174	11 (33)	67	5 (7)
GSE76503	48	0 (0)	21	0 (0)
GSE79064	18	0 (0)	1	0 (0)
GSE79257	137	62 (71)	28	6 (42)
GSE79329	34	0 (0)	0	0 (0)
GSE80261	216	5 (35)	51	11 (17)
GSE81846	16	0 (0)	5	0 (15)
GSE82084	36	10 (10)	7	5 (6)
GSE85042	71	0 (0)	0	0 (1)
GSE85506	47	0 (1)	0	0 (0)
GSE85568	115	0 (3)	13	0 (0)
GSE87571	732	1 (15)	0	7 (8)
GSE87582	21	0 (0)	0	0 (1)
GSE87655	6	2 (3)	0	0 (5)
GSE88883	100	8 (8)	0	0 (0)
GSE89474	10	1 (1)	5	0 (0)
GSE90871	24	0 (0)	2	0 (0)
GSE93266	75	11 (11)	1	0 (0)
GSE94462	16	4 (6)	3	1 (1)
GSE97362	235	35 (35)	47	1 (2)

GSE98056	69	23 (25)	49	0 (0)
GSE98203	88	43 (45)	0	1 (1)
GSE98876	71	16 (16)	42	0 (0)
GSE99553	84	23 (23)	1	0 (0)
GSE99863	257	29 (29)	3	6 (6)
GSE100940	24	0 (0)	0	0 (1)
GSE102177	36	5 (5)	0	0 (0)
GSE102504	25	7 (7)	11	0 (0)
GSE103413	67	1 (6)	1	3 (7)
GSE103769	57	0 (0)	5	0 (0)
GSE103911	65	5 (5)	0	0 (0)
GSE104087	40	0 (0)	0	1 (1)
GSE104287	48	3 (4)	0	2 (2)
GSE104472	72	19 (19)	0	1 (1)
GSE104812	48	15 (15)	0	0 (0)
GSE105124	108	84 (84)	0	0 (108)
GSE105798	11	4 (4)	0	0 (0)
GSE107737	24	12 (12)	0	0 (0)
UKHLS	1193	13 (1193)	95	14 (19)

The number of outliers flagged by each workflow is described in Table 2.2. There appears to be considerable variation in the number of samples flagged between each method. Overall it appears that the tools described in this thesis (wateRmelon) are more conservative (flag fewer samples) when compared to both MethylAid and ewastools. MethylAid and ewastools performed relatively the same and tended to flag samples in high numbers on separate datasets. The disparity between different methods suggests that a comprehensive and careful approach to quality control is likely needed in all situations.

Each quality control workflow presents its own disadvantages. Ewastools tends to flag outliers exclusively on the Non-Polymorphic control probes which is seen in both GSE61454 and UKHLS where almost all samples were flagged as outliers. Heiss & Just (2018) do caution that these quality control checks are to be used to flag samples for further investigation as the samples could still be usable. MethylAid has a tendency to identify outliers based on the hybridisation efficiency control probes as seen in cases such as GSE51032 and GSE59065 where many of the samples are flagged. Lastly, the tools in wateRmelon are mostly data-driven they suffer in performance if they are run on large datasets. Additionally, they are functionally dependent on the composition of the dataset being analysed. Specifically, these tools are sensitive to datasets that contain mixtures of tissues (outlyx) or are dependent on the normalisation method that has been applied to the data (qual).

The datasets that are described in Table 2.2 only represent a small fraction of the data available on public repositories. While it would be interesting to consider all of the datasets that are publicly available it would be tedious and ultimately uninformative to examine all datasets that are of usable quality with respects to quality control. As a result, I have chosen to focus on a selection of datasets that displayed a reasonable variation in the number of samples flagged by each method to examine how removing the flagged samples according to each quality control method will affect the downstream results.

For the following analyses, I perform statistical models on a selection of the quality controlled datasets with each of the quality control methods. I decided to not exclude samples based on the non-polymorphic control probes for the ewastools method because of the tendency of these probes to flag the majority of samples within a given dataset. The datasets that were chosen had models that were relatively close to the

original analyses performed in their respective studies where the data originated from. In circumstances where no EWAS had been previously published or the original model was not possible to reproduce due to missing variables the models was then assumed. In all circumstances, the final model used for each dataset was either a fixed-effect linear regression model or an ANOVA depending on the variable of interest.

Test statistics were compared for the following workflows: No quality control, ewastools (without NP probes), MethylAid and wateRmelon (with bscon). All quality-control methods explored additionally had pwod applied to compare with the dasen normalised β matrix. To evaluate whether or not the models were inflated or indeed the test statistics had been improved a few measures were considered. Genomic Control (λ_{GC}) was used to quantify the general degree of test statistic inflation despite the limitations described earlier and the number of significant results (unadjusted and Bonferroni corrected) were also examined.

Table 2.3 summarises the results of the statistical testing for each dataset using each quality control workflow. Application of wateRmelon yielded the largest relative decreases in test statistic inflation and the number of significant probes when compared to the other quality control pipelines. All forms of quality control did affect the downstream results usually a decrease in test statistic inflation which shows that any amount of quality control is likely to have an impact on a study. A key point that needs to be considered is that the amount of test statistic inflation and the number of significant probes identified in a model will be dependent on what is being examined. For example, if the model is looking at differences between sex, then it makes sense that there will be thousands of genome-wide significant hits as there are thousands of loci located on the sex chromosomes on the microarray.

The number of probes tested in each model tested does vary between dataset and the quality control method used. However, within this study, the number of probes will not vary considerably ($<$ 1000 probes) between the different quality control methods but may not be comparable with the number of probes analysed in the original study the dataset comes from.

Table 2.3: Summary of results following various quality control on different data-sets. Variable is defined as the chosen variable to explore in a given model (either linear regression or an ANOVA). Processing is determined by what type of normalisation and probe filtering each dataset underwent before statistical analysis. 'norm' indicates the application of pfilter followed by dasen normalisation from wateRmelon. 'pwod' indicates the application of pfilter, dasen and then application of pwod to the normalised β values. All analyses were conducted on either normalised or normalised + pwod β values. The number of significant hits $p < 0.05$ is included for unadjusted p values and adjusted p values following Bonferroni correction. λ_{GC} is defined as $\frac{median(\chi^2_{unadjusted})}{0.455}$

Dataset (Publication)	Variable (Test)	Method	No QC			ewastools			MethylAid			wateRmelon		
			$p_{unadj.} < 0.05$	$p_{bon.} < 0.05$	λ_{GC}	$p_{unadj.} < 0.05$	$p_{bon.} < 0.05$	λ_{GC}	$p_{unadj.} < 0.05$	$p_{bon.} < 0.05$	λ_{GC}	$p_{unadj.} < 0.05$	$p_{bon.} < 0.05$	λ_{GC}
GSE42861 (Liu *et al.*, 2013)	Rheumatoid Arthritis (lm)	norm	37759	5	1.235158	37475	5	1.231318	37759	5	1.235158	36880	5	1.216171
		pwod	37763	9	1.235514	37559	7	1.233362	37763	9	1.235514	36357	6	1.206290
GSE51032 (N/A)	Breast Cancer (lm)	norm	48048	0	1.481856	46621	0	1.440293	57071	0	1.592060	47168	0	1.429385
		pwod	42998	0	1.349631	43015	0	1.346276	54100	0	1.552292	43748	0	1.382146
GSE59065 (Tserel *et al.*, 2015)	Age (lm)	norm	236209	78066	8.290527	236209	78066	8.290527	188248	22624	4.869290	241321	80373	8.608858
		pwod	245386	81030	9.218022	245386	81030	9.218022	187443	23003	4.818709	251882	83403	9.693425
GSE61454 (Bonder *et al.*, 2014)	BMI (lm)	norm	39768	37	1.482159	16593	7	0.841826	28305	0	1.353793	24557	11	1.040576
		pwod	38953	27	1.444656	16430	3	0.834339	28828	0	1.364034	24213	5	1.017024
GSE61496 (Tan *et al.*, 2014)	Birth Weight (lm)	norm	31627	0	1.325080	23578	0	1.075219	21233	0	0.958290	21002	1	0.943780
		pwod	26346	0	1.060213	24191	0	1.020424	22697	0	0.968867	23058	1	0.977962
GSE63695 (Rushton *et al.*, 2014)	Osteoarthritis (ANOVA)	norm	7428	0	1.282734	7428	0	1.282734	9240	0	1.383491	9242	0	1.385023
		pwod	7456	0	1.280342	7456	0	1.280342	9093	0	1.365895	9060	0	1.368008
GSE67393 (Inoshita *et al.*, 2015)	Sex (lm)	norm	77183	8028	1.878653	71246	7493	1.767316	37716	5390	1.119859	71961	7526	1.778840
		pwod	77331	8087	1.882557	71253	7536	1.769167	38059	5392	1.126068	72211	7566	1.784041
GSE74432 (Choufani *et al.*, 2015)	Sotos Syndrome (ANOVA)	norm	136837	7168	3.239918	136837	7168	3.239918	139788	8163	3.333198	127017	6434	2.958290
		pwod	138769	7174	3.322662	138769	7174	3.322662	141405	8170	3.402514	128855	6440	3.025200
GSE74548 (Kok *et al.*, 2015)	Effect of Treatment (lm)	norm	20344	0	0.967834	17502	0	0.852270	25985	0	1.069511	17754	0	0.887979
		pwod	20765	0	0.983169	17800	0	0.861339	25977	0	1.069468	18164	0	0.900661
GSE80261 (Portales-Casamar *et al.*, 2016)	FASD (lm)	norm	60420	29	1.453326	75315	32	1.689033	88966	17	2.064697	88928	52	1.815312
		pwod	70708	33	1.797575	75939	32	1.892840	79805	18	1.990227	81680	50	1.769820
GSE87571 (Johansson *et al.*, 2013)	Age (lm)	norm	255668	99587	9.888464	255370	99312	9.850863	255668	99587	9.888464	255848	99090	9.885893
		pwod	260200	100817	10.388363	259887	100522	10.344170	260200	100817	10.388363	260211	100227	10.375626
GSE99863 (N/A)	Season of Birth (lm)	norm	82612	3	2.165199	79815	2	2.142527	86286	3	2.238093	83457	3	2.178623
		pwod	79818	3	2.085377	71630	2	1.967699	82981	3	2.153933	79875	3	2.086365
UKHLS	Blood TG Levels (lm)	norm	52746	12	1.027468	62601	17	1.168107	53152	11	1.050953	53385	15	1.082187
		pwod	57439	15	1.123828	58606	16	1.134621	57682	11	1.129119	52051	14	1.069660

The models performed here do vary from the original manuscripts slightly. In the case of GSE42861, the original study by Liu *et al.* (2013) identified around 50,000 genome-wide significant (after Bonferroni correction) results. The only real difference between these analyses is that I used dasen normalisation and additionally included the slide and microarray position as covariates within the model. Upon removing these terms from the model, I find that the number of genome-wide significant hits for the no quality control model increases to $>$ 20,000 probes. This falls short of the 50,000 probes identified by Liu *et al.* (2013) which suggests that the disparity between these results is likely due to the normalisation methodologies. Indeed the dasen normalisation method attempts to correct for positional effects in addition to correcting for Type I and Type II differences which could be driving these large numbers of significant results. Considering that more than five years have passed since the Liu *et al.* (2013) paper was published it may be worth revisiting these results using the understanding we have gained during this time to see if it is possible to glean any new or potentially missed understanding.

Out of the three quality control methods explored in this study, the best performing method appears to be wateRmelon as it is firstly the most conservative (preserves the most samples) method out of the three methods. In addition to keeping the largest number of samples it did decrease test statistic inflation and also changed the number of genome-wide significant hits. This trade-off between keeping the largest number of samples and improving the results demonstrates that data-driven methods are superior in terms of quality controlling data. In addition, the application of pwod, regardless of the quality control method, further reduced the amount of test statistic inflation and the number of genome-wide significant results in approximately half of the cases. Considering that pwod only requires a β matrix to function it is general enough that it can be applied to any dataset before statistical testing and should yeild and improvement in results.

2.4 Discussion

In this study, I attempt to identify what is the most effective quality control method for use on DNA methylation microarray data. While it remains to be fully determined which quality control workflow

should be implemented for analysis. I present evidence to suggest that data-driven methods, particularly those aimed towards identifying outlying samples, should be considered as part of any robust and reproducible quality control pipeline.

Firstly it should be noted that all of the quality control pipelines, including those not compared in this study, will be effective in handling test statistic inflation in the majority of cases. However, the data-driven methods I have developed appear to be superior to the control-probe based methods. A comprehensive approach (one that uses both control-probe based and data-driven methods) is likely to be the most effective way to quality control data as a broad approach will identify potential outliers that would have otherwise been missed.

The tools I introduced in this chapter are (outlyx, qual, pwod) are general enough that they can be applied to any β matrix that has been produced by other software packages. This quality means that these tools can easily fit into other preexisting workflows without too much hassle. Whereas using quality control methods such as ewastools requires a highly specific pipeline that may not facilitate all necessary functionality to perform an EWAS. Although I do not compare my tools with the quality control tools described in other popular R packages, I feel that due to the high similarity between the tools examined in this study that the conclusions of this study are still applicable. I demonstrate that my tools are the most effective in reducing the amount of test statistic inflation and reduce the number of spurious associations while also preserving the largest number of samples for downstream analysis. Although these results are dependent on the variable of interest, there is a marked decrease in significant results coupled with a decrease of λ_{GC} in almost all datasets when wateRmelon quality control is applied.

The number of genome-wide significant hits in this study were determined using Bonferroni correction. This was used over applying a false discovery rate to control for multiple testing because Bonferroni correction is a simple, conservative method of determining the number of genome-wide significant hits there were in each model. As I was interested in determining which quality control methods were most effective in reducing test statistic inflation and improving genuine results the use of Bonferroni correction is appropriate because in circumstances where the number of genome-wide significant hits increased us-

ing a different methodology could suggest that the effect size of a genuine association was able to reach genome-wide significance.

The models used to estimate how effective each quality control workflow may not have been identical reproductions to the original publications. From experience, I decided to include cell-type composition estimates (where appropriate) and the slide number and microarray position for each sample as covariates. While the originating authors may not have gone as far to include these as covariates, it has been shown that cell heterogeneity should be accounted for when it presents itself (Teschendorff & Relton, 2018). As a result, I sometimes obtained wildly different results compared to the original studies. In the case of Liu *et al.* (2013) I identified a handful of genome-wide significant probes while Liu *et al.* (2013) obtained $>$ 50,000. Upon removing the slide numbers and positions from the model, the number of genome-wide significant hits increased to approximately half of what was identified in the original study suggesting that the remaining difference was due to the difference in normalisation methodology. Indeed the normalisation method I used was considerate of array position while also adjusting the probes for Type I and Type II differences while Liu *et al.* (2013) used the default normalisation methodology as provided by Illumina. Despite these methodological differences it is surprising that the such a large discrepancy between these two analyses which are for the most part identical. These discrepancies in analyses could be interesting to follow-up as we understand more about the 450K microarray and how to perform EWAS it may be beneficial to revisit old studies and attempt to reproduce the results using new methodologies.

Although the data-driven methods performed better than their control-probe counterparts they are not without drawbacks. These tools require an entire β matrix to function (or two in the case of qual) which can take up a large amount of computer memory and could take a long time to run. However, computers are getting larger and faster so these problems may not be as large of limitations as they appear to be. In addition, as these tools are data-driven, it is possible to speed up the tools considerably by sub-sampling the number of probes to be used at the cost of accuracy. In addition to potentially having a long compute time, these tools by their very definition are dependent on the contents of the data. For example, outlyx will perform poorly on datasets that contain multiple tissue types however it appears to perform fairly well on heterogeneous tissues such as blood. Likewise, qual will perform differently depending on the type

of normalisation that is being applied to the dataset and gentle normalisation methods may not impart such large changes on the data thus potentially requiring different thresholds.

Outlyx appears to be the most robust outlier tool that is currently available for DNA methylation data. Although it is derivative of manual data checking using principal components it distinguishes itself in that it is a robust and reproducible method that is robust to masking and swamping effects that will detect outliers without the user needing to investigate thresholds. While it performs poorly on mixed-tissue or mixed disease datasets, it performs remarkably well on heterogeneous tissues such as whole blood. In addition, because it only requires a β matrix is it able to be implemented in all of the current workflows (e.g. RnBeads, ChAMP, minfi) without too much hassle. Lastly, outlyx provides a simple and elegant plotting functionality that visually shows how each sample appears according to the tool.

Qual is an interesting tool as its utility has yet to be fully realised. As part of the wateRmelon quality control workflow, it works well to identify samples which have undergone drastic changes during normalisation. This aspect of quality control is almost always ignored, so it is difficult to compare this to other types of quality control as it is unique. I exclude samples using qual based on a term coined as 'normalisation violence' which is proposed to be some type of confounding that is introduced during normalisation. However, this is speculative as I have not examined the differences in downstream results without the application of qual. In the future, I would like to perform a more thorough and comprehensive examination of both quality control and normalisation to establish whether or not this normalisation violence is a confounding influence on results.

Identifying outliers on the probe level appears to have some effect on downstream results. The pwod function I developed here showed a decrease in test statistic inflation in approximately half of the datasets examined. Although it is not as large of a difference in comparison with the application of outlyx or qual, there is still a reasonable enough difference in the results to consider applying pwod in most circumstances. Considering that the alternatives to probe-filtering are currently pfilter and removing probes according to probe-lists (Zhou *et al.*, 2017) I believe that pwod is a useful addition to the probe-level quality control of data. A distinction of pwod to other probe-filtering methodologies is that pwod does

not remove probes from analysis but instead removes observations from each probe. This allows for many more loci to be tested while potentially removing some of the genetic confounding that could be driving spurious associations.

The analysis performed in this study was focused on how upstream quality control leads to better results. Other aspects of analysis can lead to differences in downstream results. It is important to consider that both the choice of the normalisation method and the choice of the statistical test being performed are likely to affect the results especially in conjunction with quality control. Use of sophisticated analyses such as bumphunting (Jaffe *et al.*, 2012) are thought to yield highly robust results as they consider comethylation implicitly during the statistical testing of differential methylation. Additionally, the identification of surrogate variables Leek *et al.* (2017); Wang & Zhao (2015); Gagnon-Bartsch (2018), removal of batch effects Johnson *et al.* (2007), thorough probe filtering (Zhou *et al.*, 2017; Pidsley *et al.*, 2016) and the use of reproducible pipelines (Lehne *et al.*, 2015, RnBeads, ChAMP) can also drastically improve results but also potentially mask genuine results that have small effect sizes.

The new method for applying genomic control described by van Iterson *et al.* (2017) in the BACON R package could vastly improve the results of EWAS. Because the application of genomic control is not yet widely adopted in EWAS because of the additional sources of confounding, λ_{bacon} could fulfil this role in the future. In this study, I chose to not apply any form of genomic control on the results and solely used λ_{GC} as a measure of test statistic inflation. In consideration, it is possible that using λ_{bacon} as the measure of test statistic inflation may have been more appropriate to use as λ_{bacon} is somewhat considerate of the additional confounding that EWAS are subjected to.

Ultimately the quality control of data requires a careful and considerate approach to ensure that there are no technical problems associated with data. I explored how a variety of quality control tools focused on identify outlying samples can affect downstream results and determined that a combination of both data-driven and control-probe based methods would likely yield the best results. I demonstrate that a data-driven approach to quality control leads to the largest improvements in the reduction of test statistic inflation while conserving the largest number of samples for analysis. It is up to those who are responsible

for the analysis of the data to perform reasonable quality control and to report the methods used in some degree so that analysis in the future can be reproduced accurately.

2.5 Conclusion

There is no definitive quality control pipeline. While it may seem obvious, it is imperative that some form of quality control is applied to the data before statistical analyses. I demonstrate that there is a great deal of variation in the results of data when different quality control pipelines are applied. This is expected as there will be small differences in the number of samples and the number of probes being analysed. While all quality control pipelines were effective in attenuating some test statistic inflation, it was the tools described here which performed the best. In most situations, the quality control tools provided solely by the software used to read in data are not able to identify all problematic samples. I recommended a comprehensive approach that considers one-of or some of the data-driven tools that I have described in this study. The tools I have developed are general enough that they can be applied to any processed β matrix and therefore can easily fit into pre-existing workflows. I suspect that these tools could be welcome additions to any bioinformaticians toolbox and should be heavily considered. I stress that the reporting of the quality control of data is as important as reporting the type of normalisation method used and will vastly improve the reproducibility of any study should it be included in the future.

Chapter 3

Bigmelon: tools for analysing large DNA methylation datasets

Bioinformatics, 2018, 1–6
doi: 10.1093/bioinformatics/bty713
Advance Access Publication Date: 23 August 2018
Original Paper

OXFORD

Gene expression

Bigmelon: tools for analysing large DNA methylation datasets

Tyler J. Gorrie-Stone [1,*], Melissa C. Smart[2], Ayden Saffari[3,4,5], Karim Malki[6], Eilis Hannon[7], Joe Burrage[7], Jonathan Mill[7], Meena Kumari[2] and Leonard C. Schalkwyk[1,*]

[1]School of Biological Sciences, University of Essex, Colchester CO4 3SQ, UK, [2]Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK, [3]Department of Psychological Sciences, Birkbeck, University of London, London WC1E 7HX, UK, [4]Department of Non-Communicable Disease Epidemiology, [5]MRC Unit, The Gambia and MRC International Nutrition Group, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK, [6]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London SE58AF, UK and [7]University of Exeter Medical School, University of Exeter, Exeter EX2 5DW, UK

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on November 23, 2017; revised on June 4, 2018; editorial decision on August 14, 2018; accepted on August 20, 2018

Abstract

Motivation: The datasets generated by DNA methylation analyses are getting bigger. With the release of the HumanMethylationEPIC micro-array and datasets containing thousands of samples, analyses of these large datasets using R are becoming impractical due to large memory requirements. As a result there is an increasing need for computationally efficient methodologies to perform meaningful analysis on high dimensional data.

Results: Here we introduce the bigmelon R package, which provides a memory efficient workflow that enables users to perform the complex, large scale analyses required in epigenome wide association studies (EWAS) without the need for large RAM. Building on top of the CoreArray Genomic Data Structure file format and libraries packaged in the gdsfmt package, we provide a practical workflow that facilitates the reading-in, preprocessing, quality control and statistical analysis of DNA methylation data.

We demonstrate the capabilities of the bigmelon package using a large dataset consisting of 1193 human blood samples from the Understanding Society: UK Household Longitudinal Study, assayed on the EPIC micro-array platform.

Availability and implementation: The bigmelon package is available on Bioconductor (http://bioconductor.org/packages/bigmelon/). The Understanding Society dataset is available at https://www.understandingsociety.ac.uk/about/health/data upon request.

Contact: tgorri@essex.ac.uk or lschal@essex.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

DNA methylation is the most easily analyzed, and probably the most stable epigenetic mark. There are multiple site-specific assay methods for DNA methylation based on bisulfite conversion, and currently the most used genome-wide method are micro-arrays made by Illumina, based upon genotyping technology. This has made Epigenome-Wide Association Studies (EWAS) (Rakyan *et al.*, 2011) possible, analogous to genome-wide association studies

(GWAS). EWAS have been dominated by the use of the Illumina Infinium HumanMethylation450 BeadChip micro-array, or 450K array (Bibikova *et al.*, 2011), which allows for the interrogation of DNA methylation levels of more than 450 000 loci at a relatively low-cost. The 450K has been used widely and as of July 2017, data from more than 60 000 arrays have been deposited onto the Gene Expression Omnibus (under GPL13534). The 450K has since been superseded by the HumanMethylationEPIC BeadChip micro-array (EPIC). The EPIC array has substantial overlap with the 450K and extends genome coverage to almost twice the number of loci (Moran *et al.*, 2016). With this increase in size of data it is apparent that current methodologies are not suitable for handling the large memory requirements necessary for analysis.

Analysis of DNA methylation array data is usually performed using one of three software packages: Minfi (Aryee *et al.*, 2014), ChAMP (Morris *et al.*, 2014) and RnBeads (Assenov *et al.*, 2014), all available on Bioconductor (Gentleman *et al.*, 2004). Minfi provides tools for the reading-in of raw data files, normalization, mapping of DNA methylation data to the genome and the identification of differentially methylation positions and regions. The ChAMP package extends the minfi package but also seeks to integrate other analyses and incorporates a selection of useful tools such as batch correction and gene enrichment analysis into a rigid workflow. RnBeads also offers a similar workflow to ChAMP but is not limited to DNA methylation micro-array data and can additionally analyze sequencing data. RnBeads also seeks to guide users through analyses with sequential functions and can even perform an entire analysis pipeline within a single function. Other packages worth mentioning include MethylAid (van Iterson *et al.*, 2014) and wateRmelon (Pidsley *et al.*, 2013), which focus on the quality control and preprocessing of DNA methylation data respectively. WateRmelon is extremely compatible with minfi, ChAMP and RnBeads and provides a variety of useful normalization methods and quality control tools. MethylAid thoroughly examines the control probes located on DNA methylation micro-arrays and presents users with a collection of graphics that help diagnose problematic samples. Downstream analysis of any resultant processed data is performed on a probe-by-probe basis with tools such as limma (Ritchie *et al.*, 2015) or with a variety of methods to identify differentially methylated regions such as bumphunter (Jaffe *et al.*, 2012) or block-finding (Hansen *et al.*, 2011).

Analysis of DNA methylation data from the raw format (.idat files) first requires the parsing of data using the illuminaio package (Smith *et al.*, 2013) and conversion into a useful format. Using the minfi package as an example: idat files are read into R, into memory, as an RGChannelSet object and subsequently can be converted into a MethylSet object using a specified normalization methodology or left unprocessed whilst simultaneously matching probes to identifiers. This MethylSet object essentially contains two matrices corresponding to methylated (M) and unmethylated (U) intensities. Statistical analysis of DNA methylation data mostly involves β values which are the ratio between the Methylated and total signals, defined as $\beta = \frac{M}{(M+U+\alpha)}$, where α is an arbitrary value to offset low intensity values (usually 100). Assuming all three steps (RGChannelSet $\rightarrow$ MethylSet $\rightarrow$ β matrix) are performed within a single R session it would not be unreasonable to assume that there are three copies of the same information stored in memory. If such analysis was performed on a dataset consisting of 1000 450K arrays we can expect to require at least 16 GB of memory (Supplementary Materials S1) to simply load and convert data from raw format to a biologically interpretable output before any statistical analysis has been performed. The memory requirements may be mitigated through careful memory management and garbage-collection however taking such steps would require reloading data into memory if they are needed at a later point in time.

All the R packages described require data to be first loaded into memory prior to any analysis. This can become an issue when handling particularly large datasets as this would take up a considerable amount of time and memory (depending on the computer) to load into R. Presently, this is not an issue as the average size of an experiment using 450K arrays is around 100 samples (400 Mb β matrix size). Out of the 900 experiments deposited onto GEO (as of July 2017), only 27 of these have sample sizes larger than 500 and these larger studies (Hannon *et al.*, 2016; Jaffe *et al.*, 2016; Liu *et al.*, 2013) may have been presented with analytical challenges during down-stream analysis. Furthermore, large-scale analyses that involved the aggregation of numerous datasets such as the ones used in creating the epigenetic clock (Horvath, 2013) or exploring repositories such as Marmal-Aid (Lowe and Rakyan, 2013) may have been severely limited by the need to load all the data into memory as this would have made analysis computationally expensive.

Recent efforts have been made to handle this problem, notably with the release of the meffil R package (Min *et al.*, 2017). The meffil R package allows the parsing of data one sample at a time and offers a single form of normalization (functional normalization) but is still limited by the fact that end result, β values, are stored in memory. In addition to this, meffil does not permit for the (raw) methylated and unmethylated intensities to available alongside the β values which can be useful in certain analyses. Furthermore meffil does not allow for interactive preprocessing of data prior to normalization, a feature that is highly important in our experience of EWAS studies.

This feature of analysis, coupled with the release of the EPIC array means that data will be increasing in size and current methodologies may not be suitable for the analysis of large datasets. To combat potential memory constraints imposed by DNA methylation analysis we introduce the bigmelon R package which includes memory-efficient tools for reading-in, quality control, exploring data and provides a practical workflow. In a well-run large-scale genomics project the data is examined, quality-controlled and stored as experimental batches are produced, rather than at the end. Bigmelon is the only existing package that is designed to facilitate a workflow of incremental data addition and analysis.

2 Approach

The bigmelon package makes use of the genomic data structure file format (.gds format) implemented in the gdsfmt package (Zheng *et al.*, 2012). Originally designed for the storage of SNP micro-arrays used in GWAS, the.gds format is a hard-disk representation of data with libraries that support efficient access. The gdsfmt package is also used by the GWASTools package (Gogarten *et al.*, 2012) and the SNPRelate package (Zheng *et al.*, 2012) which provide tools for principal components analysis and identity-by-descent algorithms that are integral in GWAS for adjusting for population structure and cryptic relatedness. In a similar manner, bigmelon is an extension of both gdsfmt and wateRmelon that enables the analysis of high dimensional DNA methylation data. The design objective of bigmelon is to provide the tools necessary for a complete workflow, these include quality control, normalization and statistical testing but also provide methods for further evaluation and analysis. Tools are additionally provided for estimating covariates such as age (Horvath, 2013), sex and whole blood cell-type proportions

(Houseman *et al.*, 2012). Another heavily used tool for evaluating and exploring data is principal components analysis, and an efficient sampling approach to doing this on a large datasets is provided. Finally, the package is designed to facilitate incremental analysis, so that small batches of data can be readily looked at for quality control and even allow for first pass analyses as data is produced.

3 Materials and methods

A summary of the bigmelon workflow is described in Figure 1, the workflow can be broken down into three main parts: data import, quality control & preprocessing and analysis. Further descriptions of each section are as follows:

1. **Data import**:
   Much like other packages described, bigmelon offers the ability to read data into R into the gds file format using the iadd or iadd2 functions. The output of these functions is a hard-disk representation of an object that closely resembles the methylumiset object from the methylumi package (Triche *et al.*, 2013). For large data-sets these functions support memory-efficient batch processing. minfi (RGChannelSet, MethylSet) and methylumi (MethyLumiSet) objects can also be converted into gds format using the eset2gds function.
2. **QC & preprocessing**:
   Once data is in a gds file, it is possible to do thorough quality control using a number of memory-efficient tools. These include checking for outliers (outlyx), array quality (bscon), principal components analysis and age predictions, which can reveal mislabelling and other problems. After problematic samples are removed the data can be normalized. A range of quantile normalization methods are available as in wateRmelon. We introduce (qual), a quality measure based on the magnitude of changes introduced by the normalizer. This can identify further problematic samples which can degrade the quality of the dataset, for example introducing test-statistic inflation.
3. **Analysis**:
   One way to analyze the data is to extract the β values or subsets of them from the gdsfmt object and analyze them with any of the conventional methods. Bigmelon also facilitates conversion to MethylSet and MethyLumiSet objects using the gds2mset or gds2mlumi functions. This of course will be limited by the available memory. The core of EWAS analysis probewise analysis, and this is can be done relatively fast using minimal memory with apply.gdsn, and can also be parallelized using clusterApply.gdsn More complex analysis methods can be adapted for use with bigmelon objects. We provide a guide to doing this using bumphunter, and a bumphunter method is provided in the package.

3.1 Datasets

To demonstrate the capability of the bigmelon package we analyze two large datasets. The first consists of 1193 individuals from the *Understanding Society:* UK Household Longitudinal Survey. The goal of *Understanding Society* is to assess long-term and short-term effects of social and economic change on a variety of outcomes. Social and economic data are recorded through questionnaires and additional information including biomarker data and genotyping micro-arrays have also been obtained. Biomarker and relevant questionnaire data are available at https://www.understandingsociety.ac.uk/about/health/data upon request. 500Ng of whole blood DNA from each individual was treated with sodium bisulfite using the EZ96 DNA methylation kit (Zymo Research, CA, USA) following manufacturer's standard protocol. DNA methylation intensities were assess using Illumina Infinium HumanMethylationEPIC BeadChips (Illumina Inc, CA, USA) in the Laboratory of Professor Jon Mill (University of Exeter). DNA methylation levels were assessed on an Illumina HiScan System (Illumina). This data-set is used to demonstrate the complete workflow described in Figure 1.

The second dataset is the Marmal-aid database (Lowe and Rakyan, 2013). Marmal-aid is the largest, most readily available dataset for DNA methylation consisting of 14 586 450K arrays. Originally it was collated to be used as a reference database for many cancerous and noncancerous tissues as it contains rich detail about each array (Tissue, Disease State, Sex and Age) but it can also serve as a useful resource for software performance on very large datasets.

3.2 Comparisons of memory usage

To test the difference in memory usage during analysis we the normalizeQuantiles function from limma (used on the Marmal-Aid dataset) with the bigmelon optimized versions (qn.gdsn). Bigmelon contains many optimized versions of functions used to normalize data and reproduce the results of the analysis precisely but differ in how the computations are handled. The aim of testing the difference in memory usage is to demonstrate that it is possible to execute memory expensive computations without much cost of speed. Memory usage was recording using an in-house bash script (Supplementary Materials S2) to monitor the memory usage of a specified R process at regular intervals during the normalization process.

3.3 Data accession

To estimate how much time it takes to retrieve that data from the hard disk into memory, the time taken to retrieve random portions of data from the Marmal-Aid dataset using the microbenchmark package (Mersmann, 2015).

All analyses were performed using R 3.4.0 on a machine with 500GB RAM (necessary for conventional analysis).

4 Results

4.1 Bigmelon provides a convenient workflow

Data import: the functions iadd and iadd2 conveniently read in raw data (idat files), and can append new data to an existing gdsfile, which is the key mechanism allowing an incremental workflow. We go through an analysis of Understanding Society data-set to demonstrate the steps shown in Figure 1.

Quality Control: The outlyx function is a robust outlier detection tool that identifies outlying samples without supervision (Fig. 2). Within the original 1193 samples it can be seen that 6 samples are outlying (Fig. 2A), and removal of the most-outlying sample yields no change in the results for the remaining samples (Fig. 2B), suggesting that the tool is not susceptible to swamping/masking effects. Similarly, removal of all outlying samples does not unmask further candidates. (Fig. 2C) further demonstrating the robustness of the quality-control procedure. Due to the unsupervised nature of this tool, it can also be used to check datasets after quality control.

Atypical arrays are most likely the result of DNA quality or processing faults, and the control probes on the array offer some information on this. bscon calculates a bisulfite conversion value, which would ideally be 100%. In some datasets this may be lower,

but certainly particularly low samples are an indication of trouble. Supplementary Figure S1 shows the output of bscon on the Understanding Society dataset. Here we select a conservative threshold of 85% bisulfite conversion, and six samples were identified as having low-quality, from these only one of these was also identified as being atypical using outlyx.

Minor systematic differences between arrays introduced by sample quantity or other technical variations are readily normalized away, and quantile normalization based methods are excellent for imposing identical distributions on vectors that are similar in the first place. The objective of EWAS is to detect a relatively small number of true differences on a homogeneous background. We introduce a function qual that measures 'normalisation violence' required to bring an individual array into line. The properties of the measure have not been fully explored, but a reasonable cutoff of 0.05 for root mean squared deviation identifies 6 potentially bad arrays in this dataset (Supplementary Fig. S2).

In summary, out of 1193 samples we began with, 18 were removed for failing qc criteria, 6 from outlyx, 7 from qual and 5 from bscon as detailed above. Each of these involve thresholds that may need to be that may need to be adjusted in some cases but in the main they can be used as automated filters. Additional qc and sanity checks are equally important but require more human intervention. Principal component analysis often reveals stratification, samples with the wrong labelled sex and other problems. In Supplementary Figure S3 we present the first and second principal component loading values which clearly show two clusters which can be used to guess the sex of samples, in our experience we have found that the number of probes required to produce such a plot is small and in some cases <1% of the total number of probes on a micro-array will produce a biologically interpretable result. It is for that reason the principal components method packaged in bigmelon allows for a random selection of probes to be used instead of the full data-set. Age prediction (Supplementary Fig. S4) can also be used to check whether or not samples aligned with their supposed phenotypic data. In addition to offering age prediction using Horvath's coefficients we also allow the option to compute ages using Hannum's coefficients (not shown) (Hannum *et al.*, 2013).

Cell-Type composition estimation (Supplementary Fig. S5) has been optimized by imposing methylated and unmethylated quantiles onto the reference dataset instead of normalizing the reference and biological dataset together as it was felt that given a large enough number of samples, the addition of the reference dataset would not have an effect on the precision of the cell-type estimates. When compared to minfi it can be seen that the cell counts calculated using by normalizing data together do not vary much from the cell counts calculated from the alternative method and correlated highly

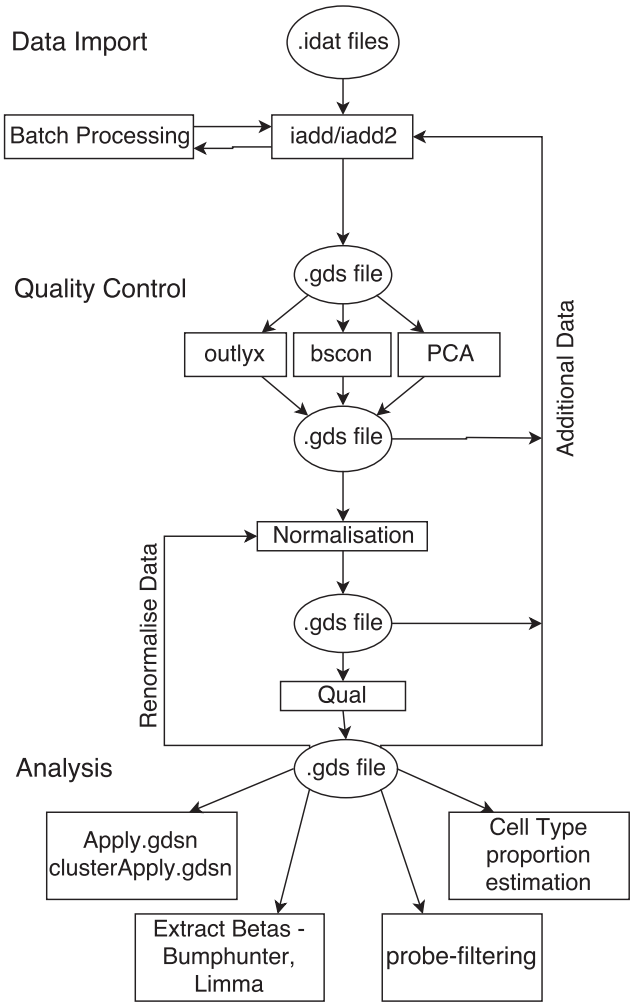


Fig. 1. Example of bigmelon workflow. The workflow is broken up into three parts: Data Import, Quality Control and Analysis. Quality-control and analysis boxes propose examples that can be used at each stage of the analysis

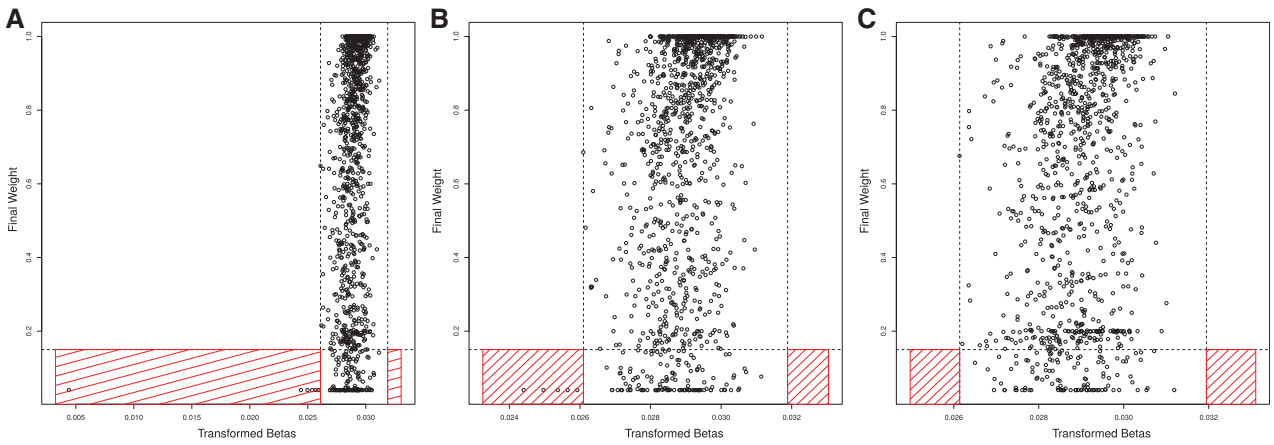


Fig. 2. Demonstration of outlyx on Understanding Society Dataset ($n = 1193$). (**A**) The results of outlyx used on all samples, (**B**) The results of outlyx with an obvious outlier removed and (**C**) the results of outlyx with all outliers removed

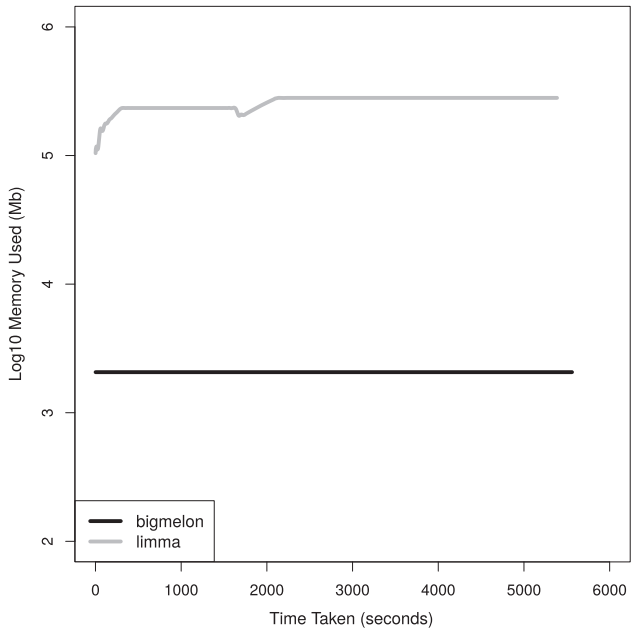


Fig. 3. Comparison of quantile normalization on 52 GB β matrix from Marmal-aid dataset ($n = 14\,586$) using limma::normalizeQuantiles function and bigmelon::qn.gdsn, computation was performed on a single core computer with 500GB of memory

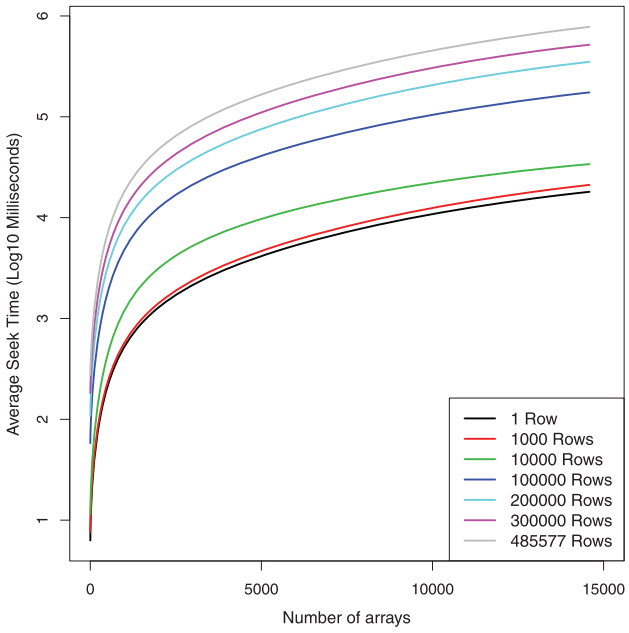


Fig. 4. Median time spent randomly accessing different sized portions of data from the Marmal-Aid data-set ($n = 14\,586$) stored in.gds file format

together (Root Mean Squared Differences between minfi and bigmelon estimated cell counts range from 0.020 to 0.006).

4.2 Bigmelon uses less memory

When comparing the memory usage of bigmelon to other software (limma and wateRmelon) it can be seen that there is at least a hundred-fold difference in memory usage at any given time throughout analysis (Fig. 3). These improvements in memory efficiency are mostly dependent on the size of the data that are being analyzed however this demonstrates that there is vast improvement using two large biological data-sets. This improvement suggests it may be possible to carry out a complete analysis workflow on a low-end computer (e.g. a workstation with just over 2 GB memory) as a full analysis only requires 600 MB of memory at any given time. In this comparison the performance of limma is identical to that of wateRmelon and minfi, as all use the same normalizeQuantiles function. This is further demonstrated in Supplementary Figure S6 where we assess the time it takes to quantile normalize varying data-sizes on a modest workstation where it quickly runs out of memory and resorts to thrashing to complete analysis. This reflects how both minfi, wateRmelon and other R packages would perform.

4.3 Random access is fast

Despite being stored on the hard-disk access is still relatively fast (Fig. 4). The median seek-time, using the Marmal-Aid dataset as a benchmark, ranged from 6.2 ms when seeking a single data-point randomly from the gds file to 13 min, when seeking all the data (458 877 rows, 14 586 columns). Additionally, accessing full rows and columns from hard-disk take on average 14 and 0.3 s respectively. It however must be noted that it appears the time required for accessing any amount of data is dependent on the number of samples being accessed at the time, for example accessing all data for a 500 sample dataset will only take 22 s.

5 Discussion

We have demonstrated how the bigmelon package resolves a severe limitation that is associated with current methodologies in EWAS. The bigmelon package facilitates the reading-in, quality-control, preprocessing and statistical analysis of DNA methylation microarray data with an additional selection of useful tools. Through storage of data on the hard-disk it is possible to circumvent majority of memory constraints and allow the analysis to be performed on most computers. Additionally, due to the nature of the workflow (Fig. 1) it is possible to append data to pre-existing gds files allowing users to analyze data as it is produced. The workflow has similarities to the workflows presented in minfi and ChAMP, and there is a reasonably simple transition path from these to bigmelon.

Currently, bigmelon does not support all of the generalized clustering methodologies used for the identification of differentially methylated regions, although we do have an implementation of bumphunter. Bigmelon allows for the seamless transition to and from minfi or methylumi data structures (MethylSet and MethyLumiSet objects), offering a route to using specialized tools if enough memory is available. To assist in the writing optimized functions for users with highly specific analyses we have provided a guide to writing functions for bigmelon that covers most of the important aspects to writing memory efficient code (Supplementary Materials S3). We plan to implement as many analyses as we see fit and will strive towards implementing many existing methodologies in the future.

6 Conclusion

The bigmelon package offers users the ability to easily handle and analyze large DNA methylation datasets (both 450K and EPIC) without the need of huge RAM or powerful computers however can reap the benefits of powerful computers as the gds file format supports parallel computing. The bigmelon package trivializes the

compilation, exploration and analysis of extremely large datasets and should prove integral for the analysis of DNA methylation data in the future.

Acknowledgements

The authors acknowledge the use of the High Performance Computing Facility (Genome) and its associated support services at the University of Essex in the completion of this work.

Funding

Understanding Society is funded by the Economic and Social Research Council (ES/N00812X/1). MK is funded by Essex University and ESRC (ES/M008592/1). MS is funded by the ESRC (ES/M008592/1). LCS and JM are funded by Medical Research Council (MR/K013807/1). TGS is funded by Essex University.

Conflict of Interest: none declared.

References

Aryee,M.J. *et al.* (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

Assenov,Y. *et al.* (2014) Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods*, **11**, 1138–1140.

Bibikova,M. *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Gogarten,S.M. *et al.* (2012) GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, **28**, 3329–3331.

Hannon,E. *et al.* (2016) An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol.*, **17**, 176.

Hannum,G. *et al.* (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell*, **49**, 359–367.

Hansen,K.D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.

Horvath,S. (2013) DNA methylation age of human tissues and cell types. *Genome Biol.*, **14**, R115.

Houseman,E.A. *et al.* (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, **13**, 86.

Jaffe,A.E. *et al.* (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.*, **41**, 200–209.

Jaffe,A.E. *et al.* (2016) Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat. Neurosci.*, **19**, 40. 7.

Liu,Y. *et al.* (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.*, **31**, 142–147.

Lowe,R. and Rakyan,V.K. (2013) Marmal-aid – a database for Infinium HumanMethylation450. *BMC Bioinformatics*, **14**, 359.

Mersmann,O. (2015) *microbenchmark: Accurate Timing Functions.*, R Package, https://CRAN.R-project.org/package=microbenchmark.

Min,J. *et al.* (2017) Meffil: efficient normalisation and analysis of very large DNA methylation samples. doi: 10.1101/125963.

Moran,S. *et al.* (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, **8**, 389–399.

Morris,T.J. *et al.* (2014) ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics*, **30**, 428–430.

Pidsley,R. *et al.* (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, **14**, 293.

Rakyan,V.K. *et al.* (2011) Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, **12**, 529–541.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

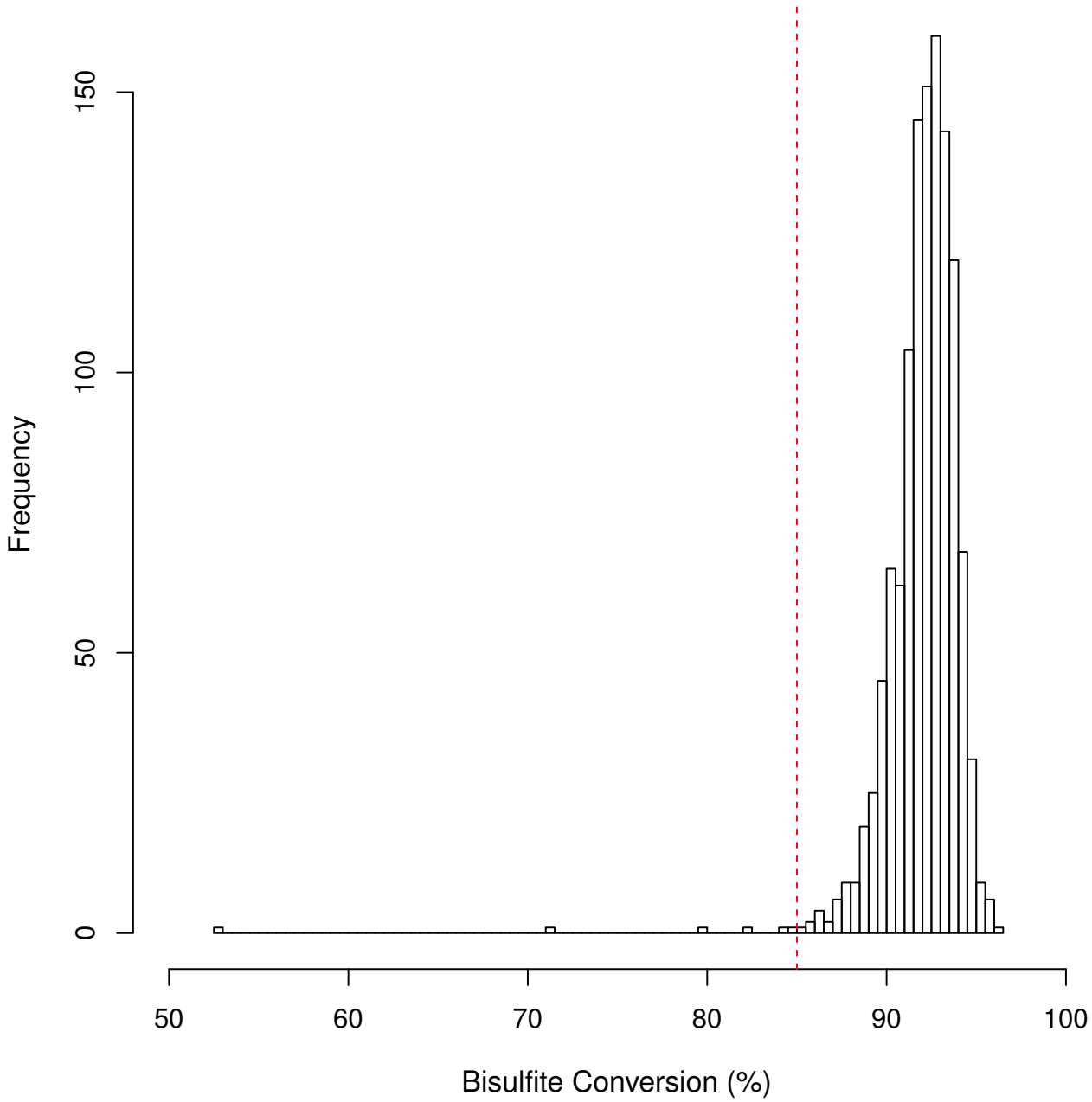
Smith,M.L. *et al.* (2013) illuminaio: an open source IDAT parsing tool for Illumina microarrays. *F1000Research*, **2**, 264.

Triche,T.J. *et al.* (2013) Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.*, **41**, e90.

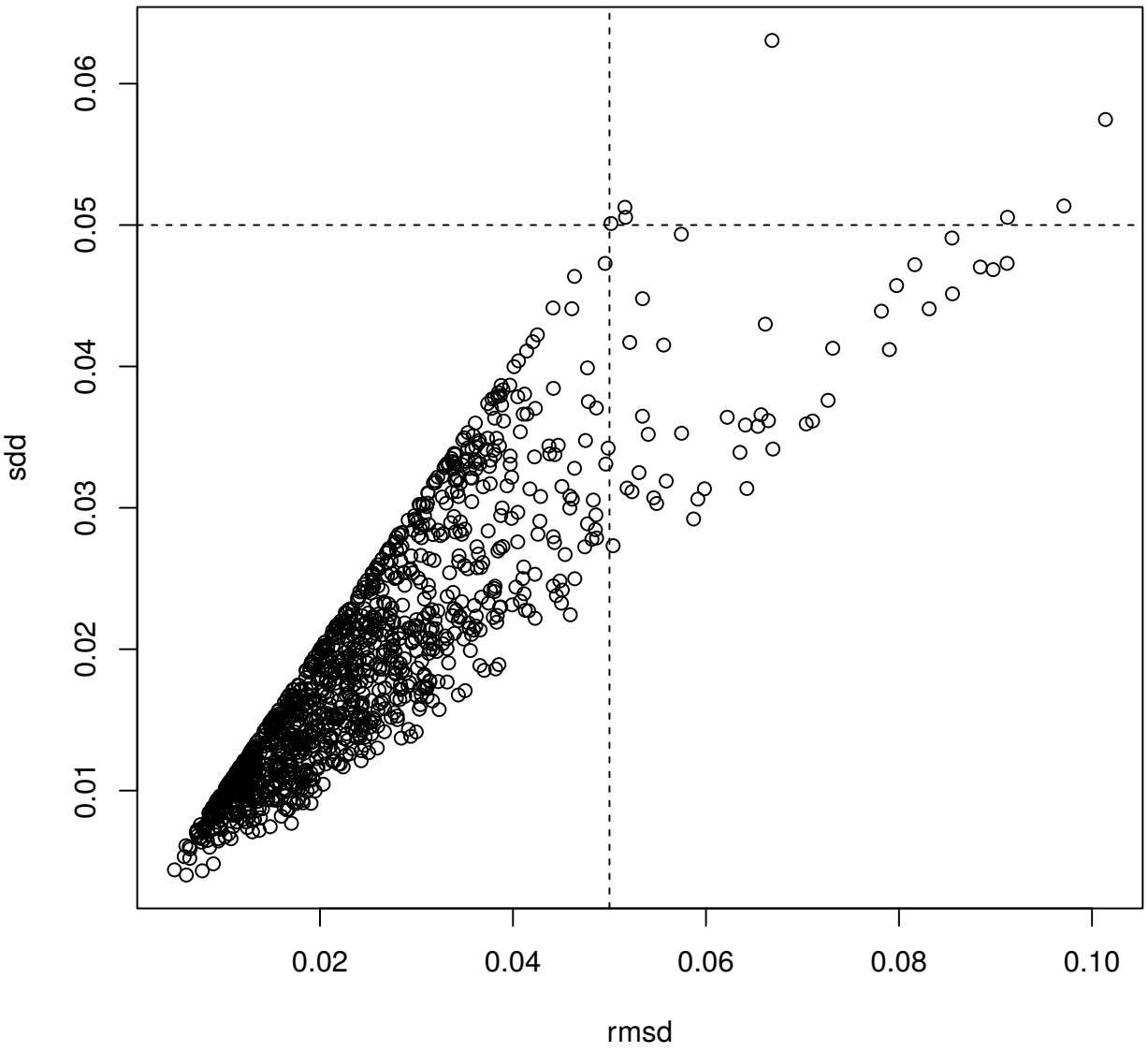
van Iterson,M. *et al.* (2014) MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics*, **30**, 3435–3437.

Zheng,X. *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.

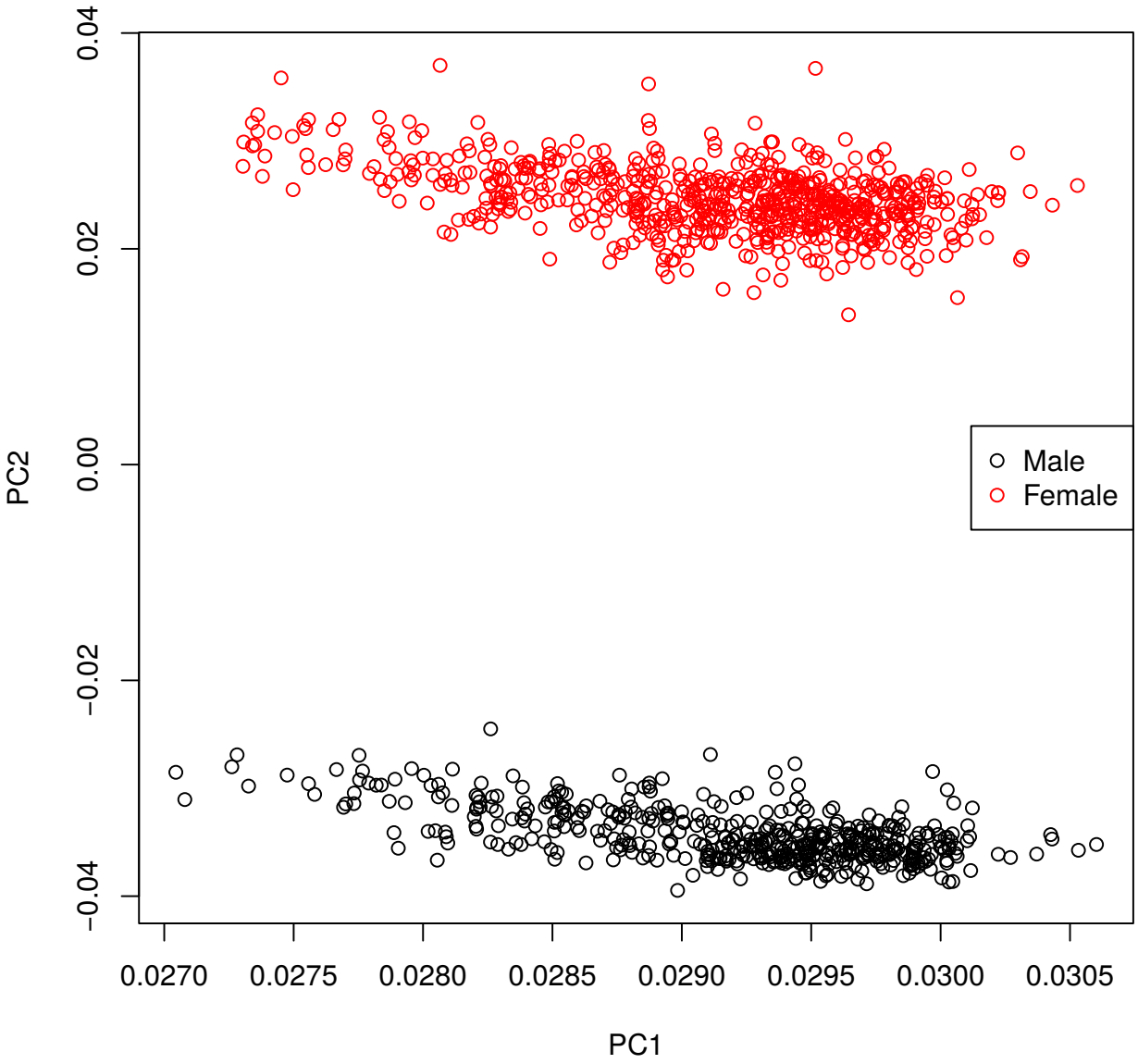
3.1 Supplementary Figures



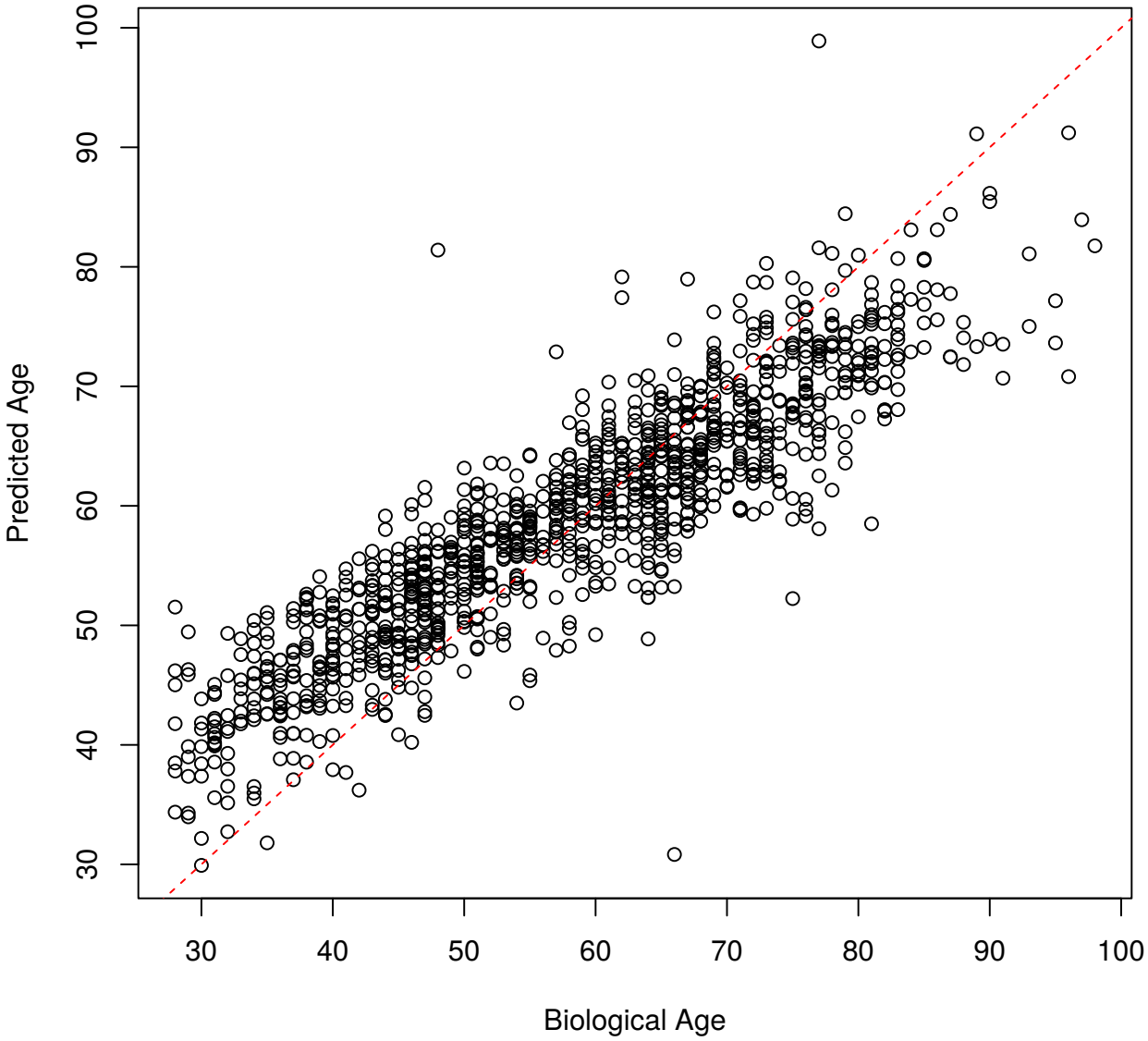
Supplementary Figure 1: Histogram of bisulfite conversion percentages from *Understanding Society:* UK Household dataset as estimated by the `bscon` function in wateRmelon. Conservative threshold of 85% represented with red-dashed line is used to filter out low-quality samples.



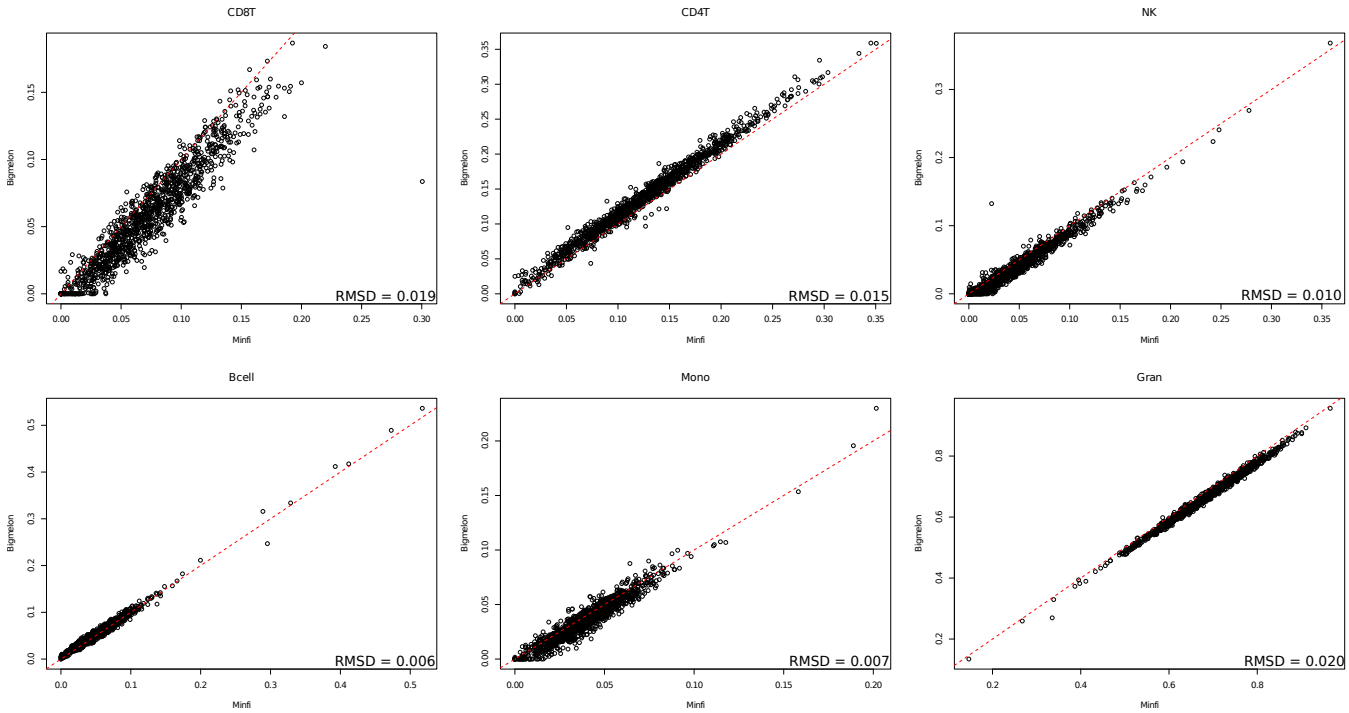
Supplementary Figure 2: Differences between `dasen` normalised and raw β values from *Understanding Society:* UK Household dataset as calculated by the `qual` function in wateRmelon package. Samples above thresholds of 0.05 (dashed lines) Root Mean Square Difference (rmsd) or Standard Deviation of Difference (sdd) were excluded from further analysis as they represent data that have undergone the most change during normalisation. Each data point represents a single sample.



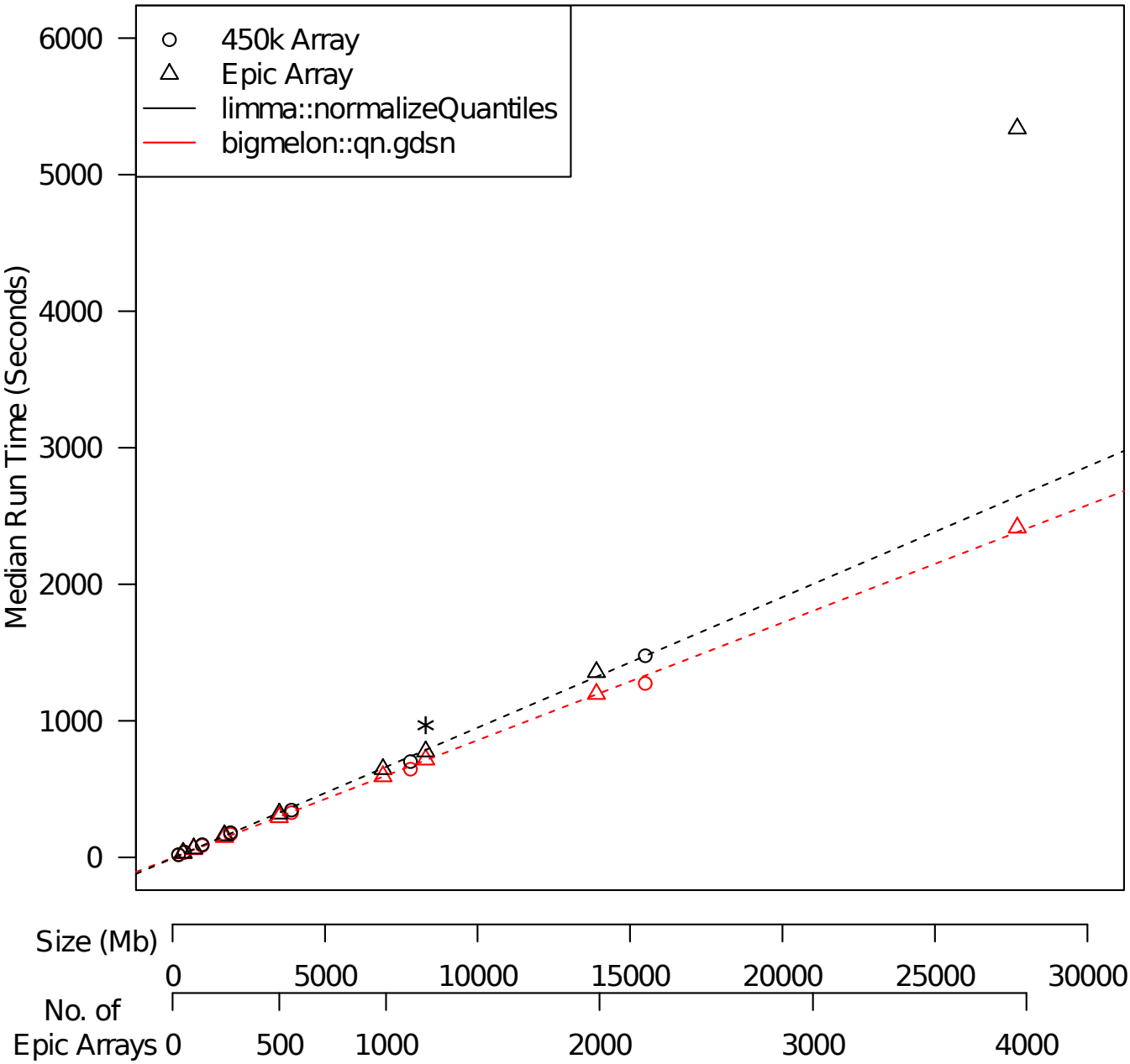
Supplementary Figure 3: Scatter plot of loading vectors from Principal Components 1 and 2 from *Understanding Society:* UK Household dataset calculated from a random 1% of data. Data-points correspond to individual samples and coloured by annotated sex.



Supplementary Figure 4: Example of age prediction on the *Understanding Society:* UK Household dataset, calculated by the `agep` function in the wateRmelon package, red dashed lines represented a perfect fit between biological and predicted age.



Supplementary Figure 5: Comparison between bigmelon and minfi methods of cell-type composition estimations. minfi cell-type compositions estimated using the estimateCellCounts function and bigmelon cell-type composition estimations determined using estimateCellCounts.gds function. estimateCellCounts.gds functions similar to the minfi version however differs only when normalising the biological data with the reference dataset, by normalising reference dataset using the biological dataset quantiles rather than normalising the two datasets together. Root mean square difference was calculated for each predicted cell-type to show overall good precision despite differing methodologies.



Supplementary Figure 6: Median run time of quantile normalisation methods of bigmelon and limma on increasing numbers of microarrays. $*$ denotes the size (in Gb) of roughly 1200 HumanMethylationEPIC arrays. Analysis was performed on a workstation with 128Gb of memory. Speeds remain comparable up until 25Gb where limma::normalizeQuantiles function runs out of memory, on workstations with less memory the breaking point is much less.

3.2 Explanation of Supplmentary Materials

Supplementary Material 1 describes a breakdown of how much memory an EWAS could theoretically take up in memory using the object.size function in R. Using linear algebra it is possible to then estimate the amount of memory required to merely load the data into R prior to any statistical testing. Additionally, this material also includes a description of how functions in the bigmelon R package have been written to exploit storing data on the hard-disk instead of within computer memory.

Supplementary Material 2 contains the bash script that was used to monitor the memory usage of a specified process running on a linux system. Due to flaws with the inbuilt methods to record memory usage within R, which tracks the sizes of objects within an environment the outputs of the original method over-estimate the concurrent memory usage of an R function which makes it difficult to determine the precise amount of memory being used by a single R process. The bash script monitors a given job id and outputs the % of memory used by the process on a given machine. This allows memory usage to be monitored over-time. Output of % of memory consumed can be converted into the amount of RAM used by multiply the output by the computers memory.

Supplementary Material 3 describes a detailed guide on how I would write memory-efficient functions using the bigmelon R package. Here I demonstrate the process of implementing a memory-efficient method of bumphunter which originally includes a number of matrix multiplications which often require large amounts of computer memory. Due to the length of this material it is located within the Appendices.

3.3 Supplementary Material 1

Concurrent memory usage

Given how standard work-flows currently exist it is necessary to parse the raw idat files in pairs and convert them into meaningful values. This can consume a lot of memory before any meaningful analysis has even taken place. For example: Parsing a single pair of idat files (450k) within minfi requires a minimum 76 MB of memory with each additional pair of idat files contributing an additional 4.8 MB to the total object size (of the RGChannelSet, not including detection p-values or beadcounts). Typical work-flow dictates that data is further processed into a MethylSet (either through processing or not) which then additionally requires at least 60 MB of memory plus 7.4 MB per sample. Lastly, analysis is often performed on Beta values which are the ratio between Methylated and Unmethylated values. Which in total requires 30 MB plus 3.7 MB per sample.

The consumption of memory can be largely avoided through garbage collection and deleting unneeded memory objects after processing. However assuming that after going through a pipeline a resultant beta matrix containing 1,000 samples there is approximately (3.7GB and 6.6GB of memory are being used to store the beta matrix alone for 450k and EPIC array respectively.) These estimations in data-sizes are estimates that do not include other encoded data within the idats such as detection-p values and bead counts. Which also influence the total memory requirement for reading in data.

Linear functions of memory usage for minfi given increasing n for 450K and EPIC micro-arrays. Generalised object sizes determined using object.size() function on increasing sizes of objects from minfi. N.B. This is estimated using RGChannelSet instead of RGChannelSetExtended which contain information that can be used to compute detection-p values and bead counts and take up more memory.

450K::

$$\begin{aligned} M_{RG-450k} &= n4.8 + 76 \\ M_{Mset-450k} &= n7.4 + 60 \\ M_{\beta-450k} &= n3.7 + 30 \\ M_{Total-450k} &= M_{RG-450k} + M_{Mset-450k} + M_{\beta-450k} \\ M_{Total-450k} &= n4.8 + n7.4 + n3.7 + 166 \end{aligned} \tag{3.1}$$

EPIC::

$$\begin{aligned} M_{RG-EPIC} &= n8 + 126.7 \\ M_{Mset-EPIC} &= n13.3 + 105.7 \\ M_{\beta-EPIC} &= n6.6 + 52.9 \\ M_{Total-EPIC} &= M_{RG-EPIC} + M_{Mset-EPIC} + M_{\beta-EPIC} \\ M_{Total-EPIC} &= n8 + n13.3 + n6.6 + 285.3 \end{aligned} \tag{3.2}$$

How bigmelon acheives memory efficiency

As the bigmelon R package stores data within a .gds file format which is stored on the hard-disk it is possible to access smaller portions of data quickly. For example, if we are interested in the β values of a single sample within a data set of 500 samples, we would require 1.9 Gb of memory to first store the entire dataset in memory and then randomly access the specified set of data. In bigmelon, we would technically only require 33 Mb of memory as the dataset is stored upon a hard-disk and the queried range is then loaded into R. Using this, it is possible to perform operations by iterating across the rows or columns of a dataset to perform tasks using as little memory as possible. This comes at a cost of speed as read from hard-disk multiple times will add up over time.

For a specific example we can take a look at performing quantile normalisation. Normally this is performed by sorting each column of data and then calculating the mean of each row which yields us the quantiles (ordered from lowest to highest). These are then used to replace the original values according to the rank of each original value, performed in a sample-wise manner.

Within bigmelon, instead of storing the sorted data within R it is possible to access each sample one

at a time, sort the data and store the sorted values in the form of a rolling sum. After performing this operation on all samples we can arrive at the row means by dividing each value by the number of samples we used. Then the data can be accessed once again (one sample at a time) and the values can be replaced by the quantiles and written back onto the hard-disk during the second pass over the data. The advantage of using bigmelon is that this process shouldn't use more than 200 Mb of memory regardless of the number of samples that are being processed however if we were using the former method we would end up a lot of memory (as described by Figure 3 within the Chapter.)

3.4 Supplementary Material 2

```
#!/bin/bash
while [ true ]
do np=$(ps -ef | grep $1 | wc -l)
if [ "$np" -gt 0 ]; then
    top -n 1 -b | grep $1 | awk -v x=10 '{print $x}' >> $2
    sleep 1
else exit
fi
done
```

Chapter 4

Lipids, Drugs and Rock & Roll

4.1 Introduction

Cardiovascular disease (CVD) refers to the set of diseases associated with the heart and circulatory system. Collectively CVDs contribute towards the largest cause of death in human beings in both developed and under-developed countries, with recent estimates accounting for around 31% of total deaths each year (Mendis *et al.*, 2011). CVDs are considered a complex disease and have many risk factors. These risk factors include age, smoking status, diet, exercise, genetic variation and blood-lipid levels.

From an examination of the number of deaths attributed to each risk factor, it can be seen that many of the metabolic traits are among the top causes of death (Figure 4.1). Considering that there are many angles of the epidemiology of CVDs to explore, I decide to focus on the relationship between DNA methylation and blood-lipid levels for this study.

These blood-lipid levels refer to four main measurements: High-Density Lipoprotein Cholesterol (HDL-C), Low-Density Lipoprotein Cholesterol (LDL-C), Triglycerides (TG) and Total Cholesterol (TC). In general, an increase in blood-lipid levels (or a decrease in HDL-C concentration) are often associated with both CVDs and atherosclerosis. Atherosclerosis describes the process of an accumulation of plaques made up of fatty materials (cholesterol and triglycerides) within arteries. The build-up of these plaques leads

to an increase in blood pressure and gives rise to myocardial infarction or stroke, usually resulting in death.

All of the associated risk factors of CVDs have previously been investigated in rich detail in both genome-wide and epigenome-wide contexts.

GWAS have identified more than 100 SNPs to be associated with blood-lipid concentrations (Willer *et al.*, 2013). The studies that looked at the genetic contribution towards elevated blood-lipid levels established the expected genetic risk of CVDs does exist. GWAS can explain why some individuals exhibit increased blood-lipid concentrations but provide little information towards understanding the complex interplay between the genetic and environmental factors that contribute to the development of CVDs. As a result, epigenome-wide approaches are now considered because they can provide insight into how non-genetic components can contribute towards disease.

Recently, numerous studies have explored the relationship between DNA methylation and metabolic traits (See Section 1.5 for an in-depth review of literature). EWAS investigating blood-lipid levels are not as popular as other metabolic traits such as Obesity or Type II Diabetes (Wahl *et al.*, 2017; Petersen *et al.*, 2014) but many studies (Hedman *et al.*, 2017; Braun *et al.*, 2017; Pfeiffer *et al.*, 2015; Dekkers *et al.*, 2016b) and reviews (Dekkers *et al.*, 2016b; Mittelstraß & Waldenberger, 2018) have been presented on blood-lipid levels over the last five years. All of the EWAS on blood-lipids to date have been performed on the 450K microarray.

Because Illumina has recently released the EPIC microarray (Moran *et al.*, 2016) the opportunity to extend and reproduce the results from previous studies on the new platform has presented itself. This is coupled with the generation of a large data-set, assayed on the EPIC array, organised by the Understanding Society: UK Household Survey - which is rich in information of socio-economic and lifestyle factors in addition to a number of biomedical measurements. In addition to extending previous results, this study will be able to serve as an excellent demonstration of both the quality control and memory optimisation methods that I have described in Chapters 2 and 3.

In this study, I present the first EWAS between blood-lipid concentrations (TC, TG and HDL-C) and DNA methylation performed on the new EPIC microarray. This EWAS was carried out on 1,193 samples

obtained from individuals taking part in the Understanding Society UK Household study. The results replicate many of the existing findings from past EWAS and present additional novel findings, providing an interesting opportunity for future work to validate and study further.

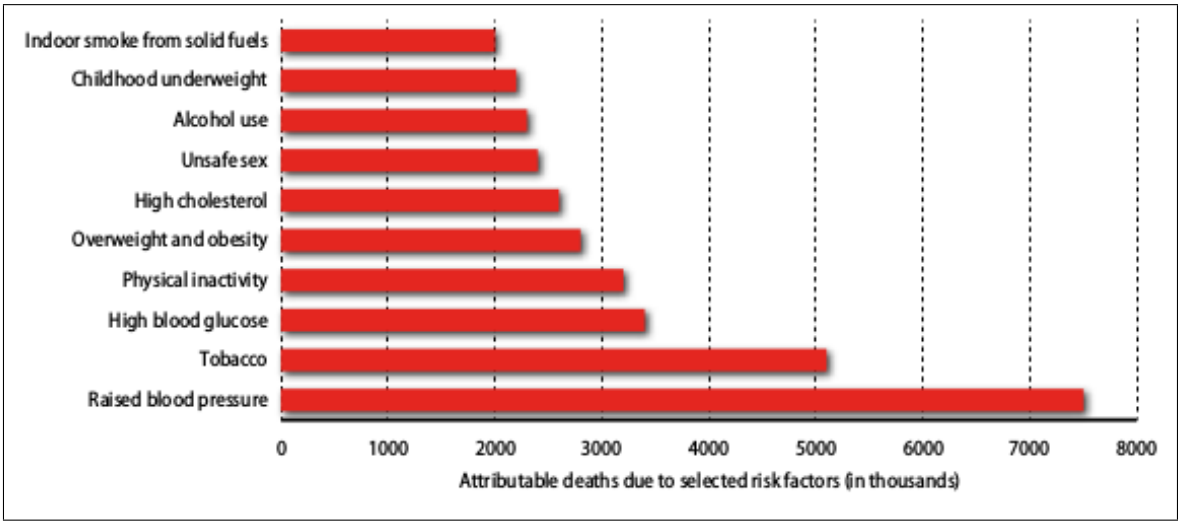


Figure 4.1: Ranking of 10 selected risk factors on cause of death - adapted from Mendis *et al.* (2011)

4.2 Methods

4.2.1 Discovery Cohort

The British Household Panel Survey began in 1991. In 2010 it was incorporated into the larger UK Household Longitudinal Survey (Knies, 2015). Annual interviews have been collecting socio-demographic information. This survey was coupled with biomedical measures and blood samples that were collected during 2011 to 2012. Respondents of the BHPS were eligible to give a blood sample if they had taken part in the previous main interview in English, were over the age of 16 years old, lived in the United Kingdom (excluding NI), were not pregnant and additionally met other conditions detailed in the Understanding Society user guide (Benzeval *et al.*, 2014).

DNA methylation profiles were obtained from DNA extracted from whole blood from 1,193 eligible individuals who had consented to both blood sampling and genetic analysis during 2011-2012 and had been present at all annual interviews between 1999 to 2011. Additionally, samples whose time between blood sampling and processing that exceeded three days were not considered for analysis. Eligibility requirements for genetic analyses meant that the epigenetic samples were restricted to participants of white ethnicity.

4.2.2 DNA Methylation Measurements

500ng of whole blood DNA from 1193 individuals were treated with sodium bisulfite using the EZ96 DNA methylation kit (Zymo Research, CA, USA) following the manufacturer's standard protocol. DNA methylation was assessed using Illumina Infinium HumanMethylationEPIC BeadChips (Illumina Inc, CA, USA) (Moran *et al.*, 2016). DNA methylation levels were quantified on an Illumina iScan System (Illumina, CA, USA). Samples were randomly assigned to chips and plates to minimise batch effects. A fully methylated control (CpG Methylated HeLa Genomic DNA; New England BioLabs, MA, USA) was included in a random position on each plate to facilitate sample tracking, resolve experimental inconsistencies and confirm data quality. Raw signal intensities were parsed into R and converted into β values using the bigmelon R

package.

4.2.3 Quality Control

Data was quality controlled using the tools described in Chapter 2. Poor quality and low represented probes were removed from the data using pfilter. Data outliers and low-quality samples were identified and removed using outlyx and bscon ($<$ 85% bisulfite conversion). Data was then normalised using the dasen quantile normalisation method. Following this, the difference between normalised and raw β values was estimated using the qual function. Samples found the have a RMSD $>$ 0.05 and a SDD $>$ 0.05 were removed from analysis. The raw signal intensities, following removal of samples identified by qual, were re-normalised using qual. After quality control, a total of 857071 probes and 1,175 samples remained for further analysis. Details of how the dataset was quality controlled are described in full detail in Chapter 3 and descriptions of the tools is provided in Chapter 2.

4.2.4 Lipid Measurements

Blood-lipid measurements were obtained from non-fasting whole-blood samples. TC, TG and HDL-C were measured directly while LDL-C measurements were calculated using Friedewald method (Friedewald *et al.*, 1972).

$$Cholesterol_{LDL}(mmol\ dm^{-3}) = Cholesterol_{Total}(mmol\ dm^{-3}) - Cholesterol_{HDL}(mmol\ dm^{-3}) - \frac{Triglyceride(mmol\ dm^{-3})}{2.19} \tag{4.1}$$

4.2.5 Discovery EWAS

Multivariate fixed-effect linear regression models were used to assess the relationship between DNA methylation and each blood-lipid trait (TG, HDL-C, TC and LDL-C) using age, sex, plate number and cell-type composition (CD8 T Cells, CD4 T Cells, Natural Killer cells, B Cells, Monocytes and Granulocytes) estimates were included as covariates in the following model.

$$\begin{aligned} DNAm_i \sim & Lipid\ Phenotype + Drug\ Status + Age + Sex + \\ & Plate\ Number + CD8T + CD4T + NK + BCell + Mono + Gran \end{aligned} \tag{4.2}$$

The cell-type estimates were computed using estimateCellCounts function from minfi using a reference based deconvolution method described by Houseman *et al.* (2012) using a reference dataset obtained from 450K microarrays.

Linear regressions were chosen because they are generally quite robust to violations in their assumptions whereas other tests may not be as forgiving. For statistical analysis, all blood-lipid measurements were scaled to a mean of 0 and a SD of 1, except for TG concentrations which were natural log transformed. These TG concentrations were log transformed as the distribution was otherwise skewed towards the left (and therefore exhibited a log-normal distribution) while the other blood-lipid measurements were otherwise normally distributed.

The statin-use model was similar to the above model but did not include any blood-lipid measurements. Participants that were taking non-statin lipid-lowering medication were further excluded from all analyses. In total 1,173 samples were used in this discovery EWAS.

Prior to statistical testing, all outlying observations on a per probe basis were removed using the pwod function from wateRmelon.

Genome-wide significant hits were assessed based on a bonferroni corrected p-value <0.05

4.2.6 Sensitivity Analysis

To test the robustness of the genome-wide significant results identified in the discovery EWAS, I performed a sensitivity analysis by re-performing the previously described model including additional covariates which could influence the results. The additional covariates include metabolic traits (BMI, waist circumference), lifestyle factors (smoking status, alcohol consumption) and the first 10 principal components obtained from previously obtained genotyping data. Further models were performed with removing genetically similar participants (singletons only) and removing all participants that were known to be using lipid-lowering medication.

4.3 Results

The characteristics of the Understanding Society: UK Household study cohort is described in Table 4.1. Out of the 1,193 samples that were initially assayed on the EPIC microarray a total of 1,173 remained following quality control. Overall the cohort consists of a majority of samples obtained from female participants who are on average one year younger than the males in the cohort. More males participants were using lipid-lowering medication which could explain the overall lower cholesterol measurements (TC, LDL-C, HDL-C) in males when compared to females while the TG levels remain higher.

EWAS for TC, LDL-C, HDL-C and TG blood concentrations were performed on 857,071 CpG sites located throughout the genome. A total 4 (TC), 0 (LDL-C), 42 (HDL-C) and 23 (TG) differentially methylated CpGs were identified in each initial analysis with estimated effect sizes ranging from 0.002 to 0.013, -0.010 to 0.011 and -0.025 to 0.015 per standard deviation increase for TC, HDL-C and TG concentrations respectively. Effect size estimates were not sensitive to smoking, alcohol-use and genetic variation;

however, some CpGs displayed sensitivity to diet-related measurements (BMI and Waist circumference). There were no genome-wide significant findings reported from the LDL-C EWAS.

The metabolic traits that were examined in this study display a wide variety of correlation Figure 4.2. TC and LDL-C shared the highest correlation alongside a high correlation between WC and BMI. HDL-C concentrations showed small correlations with TC, WC, BMI and shared an inverse correlation with TG concentrations. These correlations make some sense as LDL-C concentrations are determined as the remainder of TC concentrations after the subtraction of both HDL-C and TG measurements. BMI and WC are well known to correlate. The inverse correlation between HDL-C and TG could be explained by diets that involve higher HDL-C may be sparse in TG or vice versa. Despite the small correlations it seems appropriate to treat each of these blood-lipid traits independently in models to avoid any attenuation of genuine effect.

4.3.1 Total Cholesterol

The EWAS between DNA methylation and TC concentrations revealed a total of 4 genome-wide significant CpG sites (Table 4.2). All four of the CpGs identified were present on the 450K microarray and therefore had the opportunity to be identified in previous studies. All of the associations present a positive increase in methylation state at each CpG site with increasing TC concentrations although the effect sizes are very small, ranging from $+0.002$ to $+0.013$ per standard deviation of TC concentration.

4.3.2 Triglycerides

A total of 23 genome-wide significant probes were identified to be associated with log-transformed triglyceride concentrations and DNA methylation (Table 4.3, 4.4 and 4.5). 19 of the 23 CpG sites identified in this analysis were annotated to a total of 12 genes. Effect sizes ranged from -0.022 to $+0.015$. 9 of 23 probes identified are specific to the EPIC microarray and are therefore novel.

Table 4.1: Characteristics of the Understanding Society: UK Household Longitudinal study cohort

Characteristic	Male	Female	Total Sample
n	488	685	1173
Mean Age [SD] (years)	58.86 [14.90]	57.29 [15.01]	57.95 [14.98]
Mean TC [SD] (mmol/L)	5.21 [1.18]	5.51 [1.13]	5.38 [1.16]
Mean HDL-C [SD] (mmol/L)	1.34 [0.41]	1.66 [0.47]	1.53 [0.47]
Mean TG [SD] (mmol/L)	2.03 [1.20]	1.65 [0.97]	1.80 [1.09]
Mean LDL-C [SD] (mmol/L)	2.94 [1.04]	3.10 [0.97]	3.02 [0.99]
% using Lipid-Lowering Medication	23.77%	17.37	19.27%
Waist Circumference [SD] (cm)	100.63 [11.51]	90.74 [13.88]	94.87 [13.83]
BMI [SD]	28.24 [4.41]	28.34 [5.84]	28.30 [5.30]

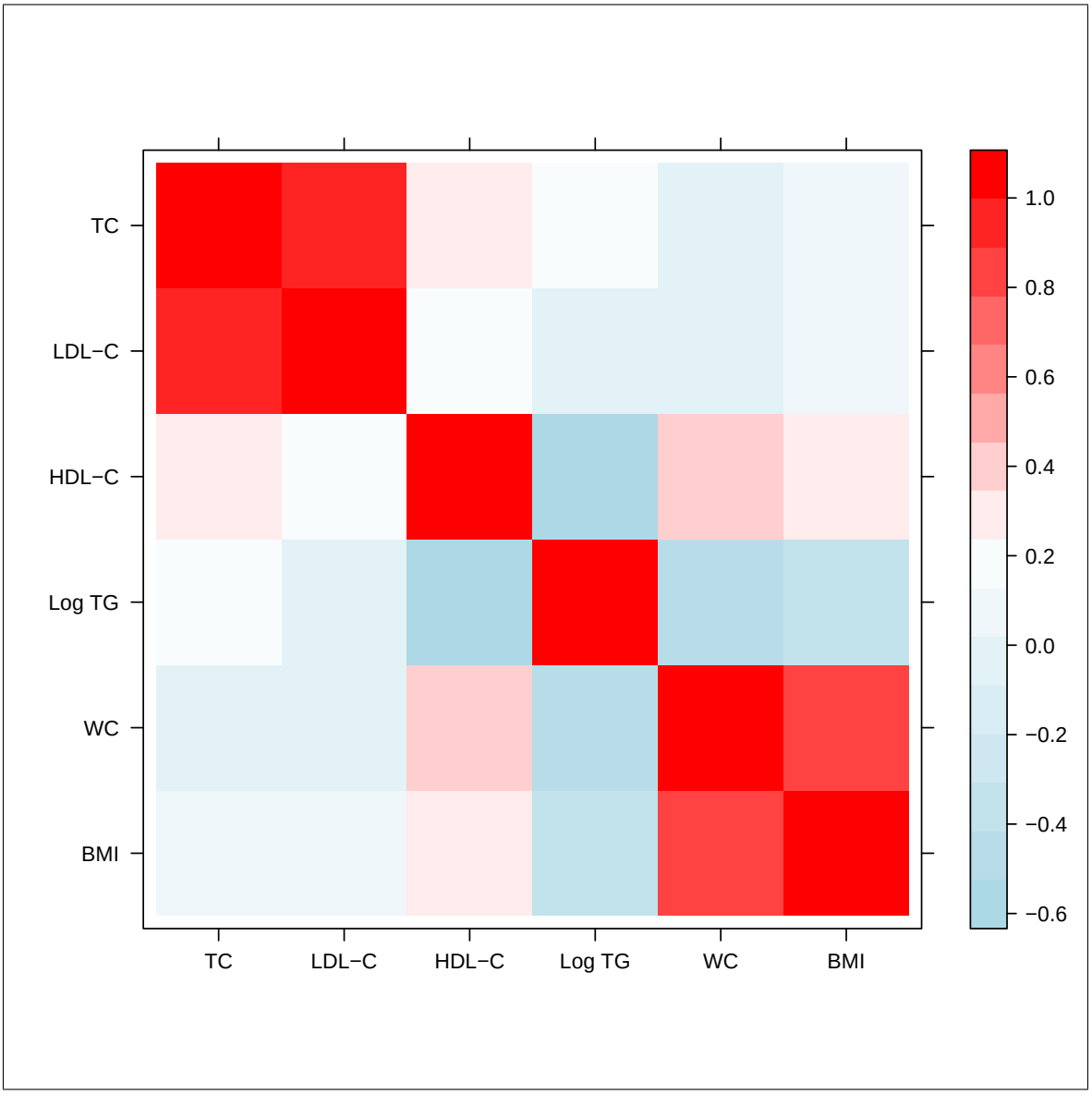


Figure 4.2: Heatmap of Pearson Correlations between metabolic traits: Total Cholesterol (TC) concentration, LDL-C concentration, HDL-C concentration, Log Transform Triglycerides (TG) concentration, Waist circumference (WC) and Body Mass Index (BMI). Obtained from 1,193 participants from the Understanding Society: UK Household longitudinal study.

Sensitivity analysis of significant probes shows that most results are robust to sources of potential confounding including genetic relatedness. Some probes are affected by the inclusion of diet-related variables (BMI and Waist Circumference), as shown by the decrease of effect size estimates (Figures 4.6 & 4.7).

Investigation of quantile-quantile plots from the discovery model and models that included the BMI and WC as covariates show that there is slight test statistic inflation in the discovery model ($\lambda = 1.04$). Including BMI and WC as covariates do nominally reduce this inflation; however, because the inflation is very close to a value of 1 I am comfortable that the discovery model is appropriate.

Table 4.2: Top 4 significant loci from the Total Cholesterol discovery EWAS

Probe ID	Gene Name	CHR	Location	Epic	Effect Size	adj p-value	Previous Association
cg03440556	SCD	10	102107757	FALSE	0.008079	0.002569	
cg07839457	NLRC5	16	57023022	FALSE	0.01314	0.03872	Hedman *et al.* (2017)
cg10073091	DHCR24	1	55352301	FALSE	0.002419	0.03909	
cg09978077	SREBF2	22	42229983	FALSE	0.002264	0.04548	Sayols-Baixeras *et al.* (2016b); Hedman *et al.* (2017)

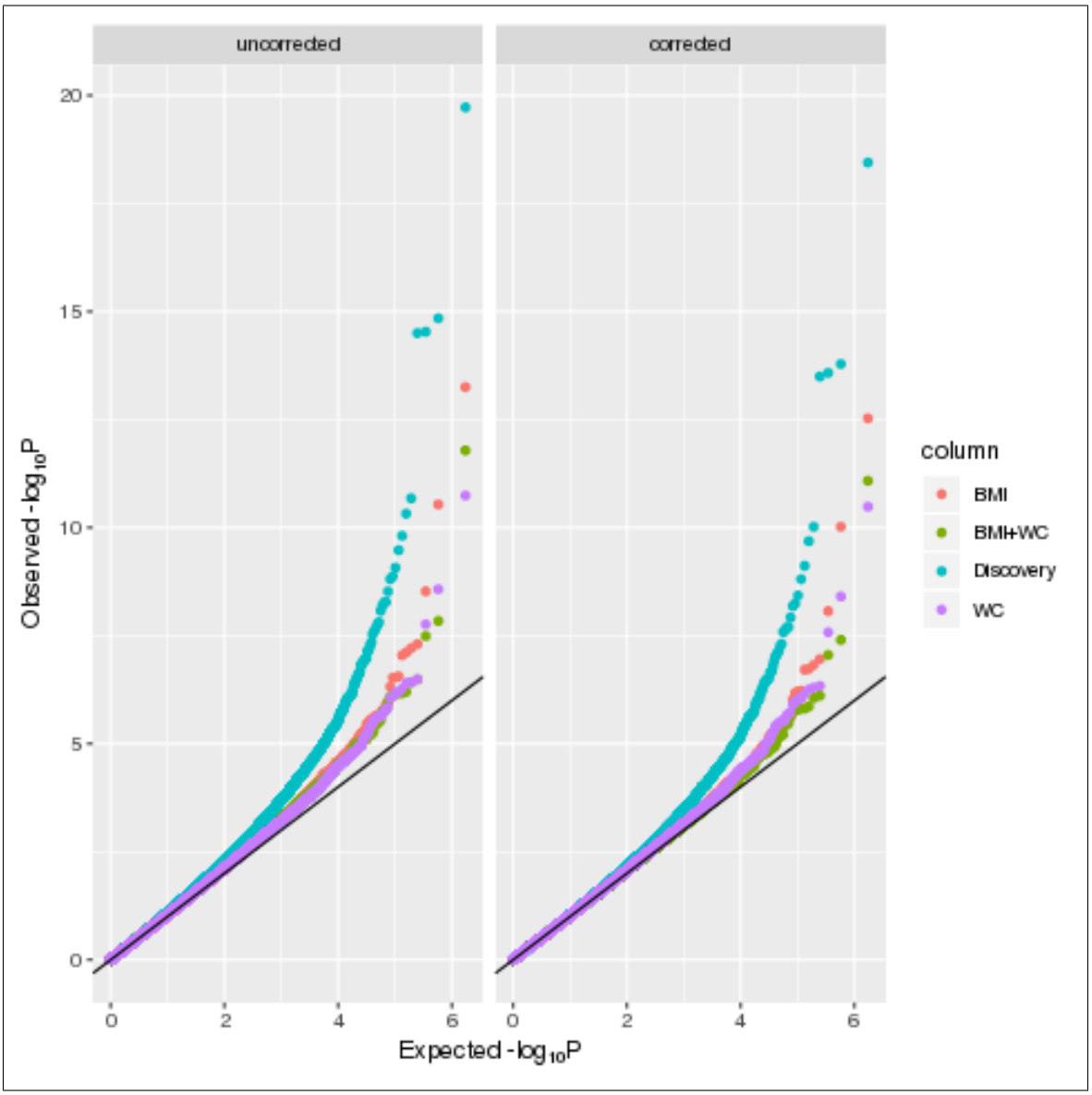


Figure 4.3: Comparison of Quantile-Quantile plots of genome-wide analysis of Triglyceride EWAS including Discovery model and models include diet-related covariates. Estimated inflation (λ) according to the bacon R package (van Iterson *et al.*, 2017) for each model is 1.04 (Discovery), 1.03 (BMI), 1.01 (WC) and 1.03 (BMI+WC).

Table 4.3: Top 23 genome-wide significant probes from the Triglyceride discovery EWAS

Probe ID	Gene Name	CHR	Location	Epic	Effect Size	adj p-value	Previous Association
cg11024682	SREBF1	17	17730094	FALSE	0.01321	7.787e-14	Dekkers *et al.* (2016b); Braun *et al.* (2017); Hedman *et al.* (2017); Sayols-Baixeras *et al.* (2016b); Pfeiffer *et al.* (2015)
cg19693031	TXNIP	1	145441552	FALSE	-0.02258	2.974e-09	Hedman *et al.* (2017); Sayols-Baixeras *et al.* (2016b); Dayeh *et al.* (2016); Pfeiffer *et al.* (2015)
cg00574958	CPT1A	11	68607622	FALSE	-0.005894	5.882e-09	Dekkers *et al.* (2016b); Sayols-Baixeras *et al.* (2016b); Hedman *et al.* (2017); Pfeiffer *et al.* (2015)
cg06500161	ABCG1	21	43656587	FALSE	0.01519	6.269e-09	Hedman *et al.* (2017); Braun *et al.* (2017); Dekkers *et al.* (2016b); Sayols-Baixeras *et al.* (2016b); Pfeiffer *et al.* (2015)
cg16740586	ABCG1	21	43655919	TRUE	0.01508	2.812e-05	
cg27243685	ABCG1	21	43642366	FALSE	0.01105	6.198e-05	Hedman *et al.* (2017); Braun *et al.* (2017); Dekkers *et al.* (2016b); Sayols-Baixeras *et al.* (2016b); Pfeiffer *et al.* (2015)
cg05325763	CPT1A	11	68607719	TRUE	-0.007075	0.0001936	
cg20052079	JARID2	6	15504923	FALSE	-0.01726	0.0004051	
cg18513344	MUC4	3	195531298	FALSE	-0.007236	0.0009992	
cg08309687		21	35320596	FALSE	-0.01785	0.00157	
cg17075888	PDK4	7	95225339	TRUE	-0.01631	0.001803	
cg03173502	JARID2	6	15505345	FALSE	-0.01605	0.003408	
cg18353028	CYP7B1	8	65669513	TRUE	-0.01433	0.00586	
cg21623127	LIX1	5	96432134	TRUE	-0.01456	0.005985	
cg13500852	JARID2	6	15505460	FALSE	-0.01296	0.007171	
cg06723828	PSMD13	11	251223	TRUE	0.009368	0.009346	
cg13027183	JARID2	6	15504872	FALSE	-0.01639	0.01678	
cg00683922	PFKFB2	1	207242569	TRUE	0.01467	0.01811	
cg17058475	CPT1A	11	68607737	FALSE	-0.005529	0.02195	Dekkers *et al.* (2016b); Sayols-Baixeras *et al.* (2016b); Hedman *et al.* (2017); Pfeiffer *et al.* (2015)
cg03062284		2	122994061	TRUE	-0.006741	0.0248	
cg19758958		11	62319222	TRUE	-0.01057	0.02781	
cg06690548	SLC7A11	4	139162808	FALSE	-0.01478	0.03169	
cg07504977		10	102131012	FALSE	0.01385	0.04751	Sayols-Baixeras *et al.* (2016b); Hedman *et al.* (2017)

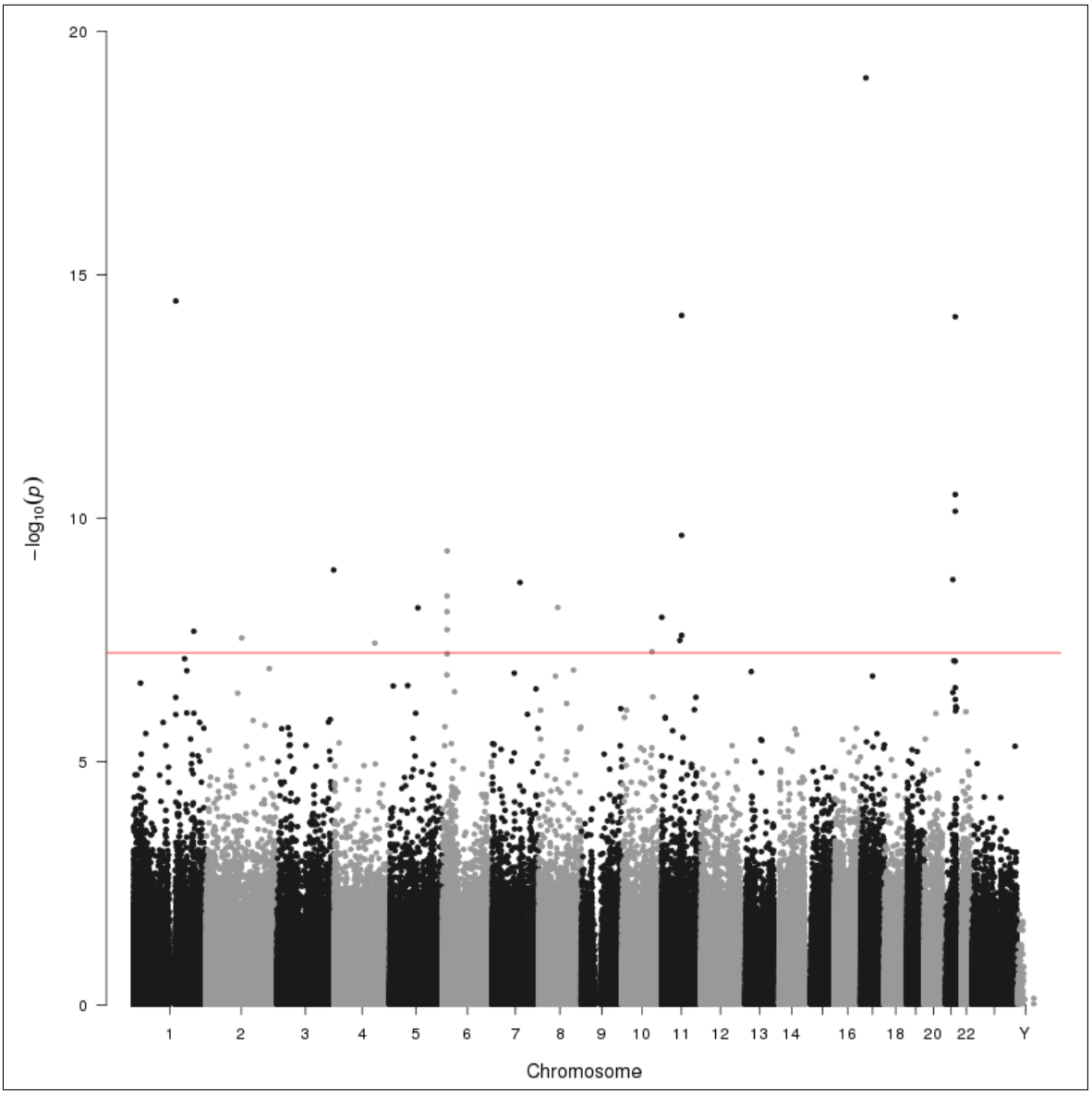


Figure 4.4: Manhattan plot of genome-wide analysis from Triglyceride discovery model. Red line denotes genome-wide significance level equivalent to bonferroni corrected p-value <0.05.

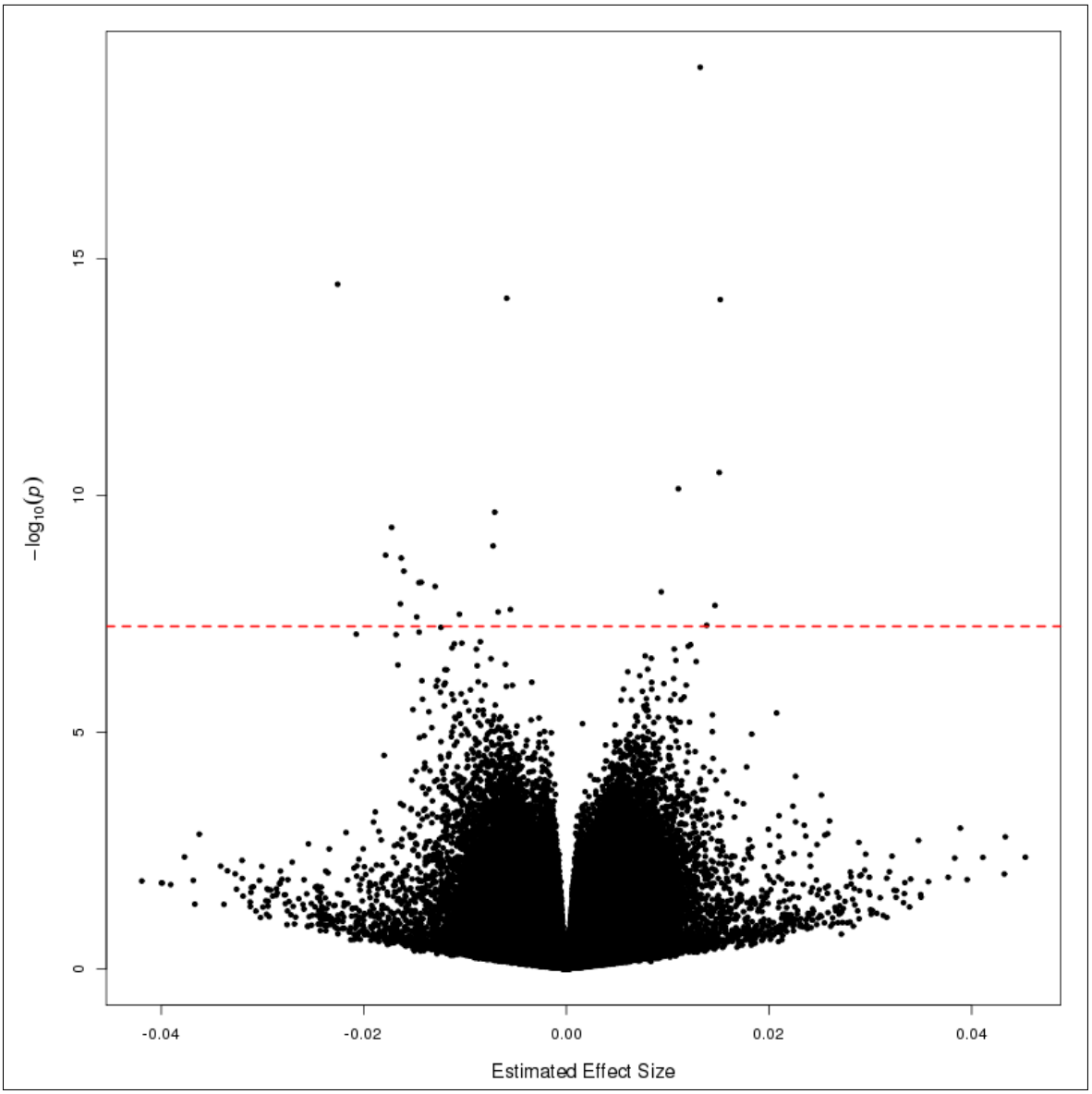


Figure 4.5: Volcano plot of genome-wide analysis from Triglyceride discovery model. Red dashed line denotes genome-wide significance level equivalent to bonferroni corrected p-value <0.05

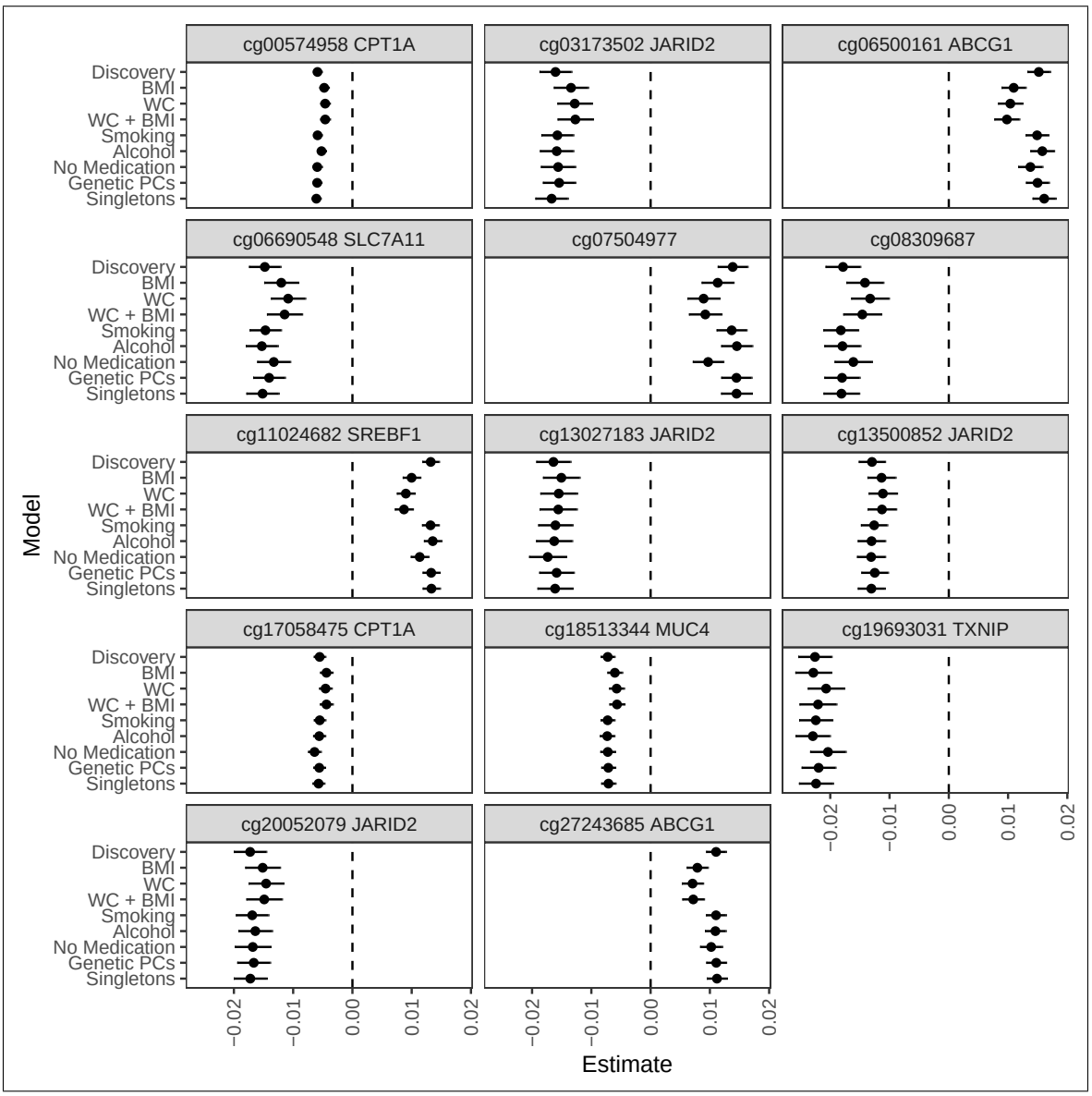


Figure 4.6: Comparison of effect sizes of genome-wide significant probes present on the 450K microarray from Triglyceride models including different covariates. Each point corresponds to the unstandardised effect size for each CpG in a given model, where the lines correspond to the size of the standard error associated with the effect size. Discovery refers to the initial model run which includes age, sex, drug status, cell-type composition estimates and plate number. Annotated labels correspond to the covariate included within the discovery analysis. No medication corresponds to the discovery model being run without any participants using lipid-lowering medication and the singleton model refers to all genetically similar participants being excluded from analysis.

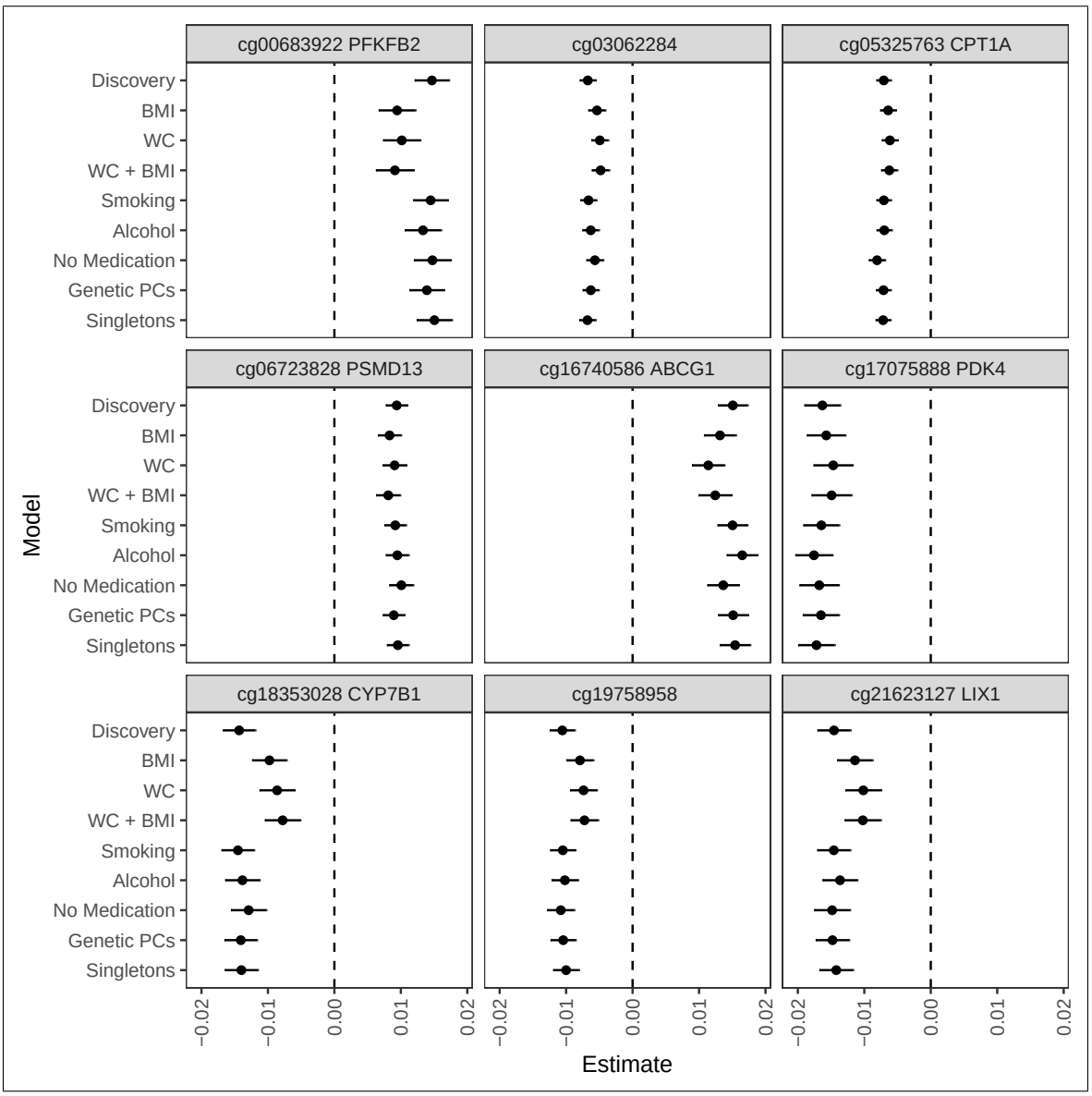


Figure 4.7: Comparison of effect sizes of genome-wide significant probes exclusive to EPIC microarray from Triglyceride models including different covariates. Each point corresponds to the unstandardised effect size for each CpG in a given model, where the lines correspond to the size of the standard error associated with the effect size. Discovery refers to the initial model run which includes age, sex, drug status, cell-type composition estimates and plate number. Annotated labels correspond to the covariate included within the discovery analysis. No medication corresponds to the discovery model being run without any participants using lipid-lowering medication and the singleton model refers to all genetically similar participants being excluded from analysis.

4.3.3 HDL-C

A total of 42 genome-wide significant loci were identified in the initial discovery model between HDL-C and DNA methylation. 23 of these probes are novel because they are only located on the EPIC array and not present on the 450K. Effect sizes ranged from -0.011 to 0.011 (Table 4.4, 4.9 and 4.10). Out of these 42 probes, these annotated to 29 genes, 10 of which did not annotate to any gene but could be distal to nearby gene regions. Similar to the sensitivity analysis of the TG significant loci, the majority of these probes are insensitive to potential confounding, however some probes are affected by the inclusion of diet-related covariates (Figures 4.11 & 4.12).

Quantile-Quantile plots of the discovery model and models including BMI and WC as covariates (Figures 4.8) show that the discovery model is not highly inflated ($\lambda = 1.13$). However, including BMI within the model does reduce this inflation considerably.

Table 4.4: Top 42 genome-wide significant probes from HDL-C discovery EWAS

Probe ID	Gene Name	CHR	Location	Epic	Effect Size	adj p-value	Previous Association
cg06500161	ABCG1	21	43656587	FALSE	-0.01032	1.266e-15	Hedman *et al.* (2017); Braun *et al.* (2017); Dekkers *et al.* (2016b); Sayols-Baixeras *et al.* (2016b); Pfeiffer *et al.* (2015)
cg17901584	DHCR24	1	55353706	FALSE	0.01066	6.903e-10	Braun *et al.* (2017); Dekkers *et al.* (2016b); Hedman *et al.* (2017)
cg00683922	PFKFB2	1	207242569	TRUE	-0.01011	3.191e-06	
cg01676795	POR	7	75586348	FALSE	-0.008554	1.581e-05	Hedman *et al.* (2017)
cg22699725	PFKFB2	1	207242586	TRUE	-0.009005	1.68e-05	
cg16100392	TAAR3	6	132931717	FALSE	-0.008796	0.0001275	
cg02246605		20	39591425	TRUE	-0.007963	0.0003291	
cg00089960	SETD1B	12	122245655	TRUE	-0.005751	0.0005655	
cg20930793	HEATR5A	14	31872220	TRUE	-0.005809	0.0009334	
cg22488164	PLBD1	12	14716910	FALSE	-0.00815	0.001048	Hedman *et al.* (2017)
cg27243685	ABCG1	21	43642366	FALSE	-0.005737	0.001116	Hedman *et al.* (2017); Braun *et al.* (2017); Dekkers *et al.* (2016b); Sayols-Baixeras *et al.* (2016b); Pfeiffer *et al.* (2015)
cg19773170		14	23008246	TRUE	-0.00876	0.001281	
cg08804919		14	100515656	TRUE	-0.005588	0.001343	
cg18353028	CYP7B1	8	65669513	TRUE	0.008293	0.001563	
cg12378285		2	136808729	TRUE	-0.007287	0.002153	
cg15328937	LOC101929452	2	7212089	TRUE	0.005092	0.002891	
cg08319289	IL1RAP	3	190261614	TRUE	-0.00728	0.00408	
cg07426444	MYO9A	15	72209354	TRUE	-0.007329	0.004416	
cg24246165	ACYP2	2	54435033	TRUE	-0.006094	0.005091	
cg16740586	ABCG1	21	43655919	TRUE	-0.007384	0.005167	
cg20301125		5	130975565	TRUE	-0.007711	0.005168	
cg19750657	UFM1	13	38935967	FALSE	-0.00753	0.005653	Hedman *et al.* (2017)
cg00301370	LOC101928162	12	10902496	TRUE	-0.006983	0.005668	
cg02335576	AGBL4	1	49338938	TRUE	-0.005823	0.007602	
cg00949930	CCDC141	2	179916205	TRUE	-0.00562	0.008225	
cg20052079	JARID2	6	15504923	FALSE	0.008867	0.008347	
cg26804423	ICA1	7	8201134	FALSE	-0.005965	0.008404	Hedman *et al.* (2017)
cg00994936	DAZAP1	19	1423902	FALSE	-0.005538	0.01204	
cg09831562	SOX2OT	3	181327125	FALSE	-0.008452	0.01336	
cg00358010		1	221338314	TRUE	0.004536	0.01712	
cg26470501	BCL3	19	45252955	FALSE	0.005793	0.01796	
cg02108045	SOCS2-AS1	12	93962028	TRUE	-0.004015	0.01821	
cg24422344		9	81178462	TRUE	-0.01045	0.03142	
cg03309738		6	13764242	TRUE	0.008042	0.03383	
cg25739715	OSM	22	30663881	FALSE	0.002355	0.03621	Hedman *et al.* (2017)
cg26080684		7	72775853	TRUE	-0.007727	0.03913	
cg10404730	SORL1	11	121330348	TRUE	-0.006776	0.03948	
cg16013680	P4HA3	11	73989529	TRUE	-0.008038	0.04048	
cg18608055	SBNO2	19	1130866	FALSE	0.006276	0.04181	
cg27444020	GAB2	11	78108461	TRUE	-0.00723	0.04688	
cg07375358	ZBTB16	11	114056431	TRUE	-0.007488	0.04815	
cg03062284		2	122994061	TRUE	0.003683	0.04845	

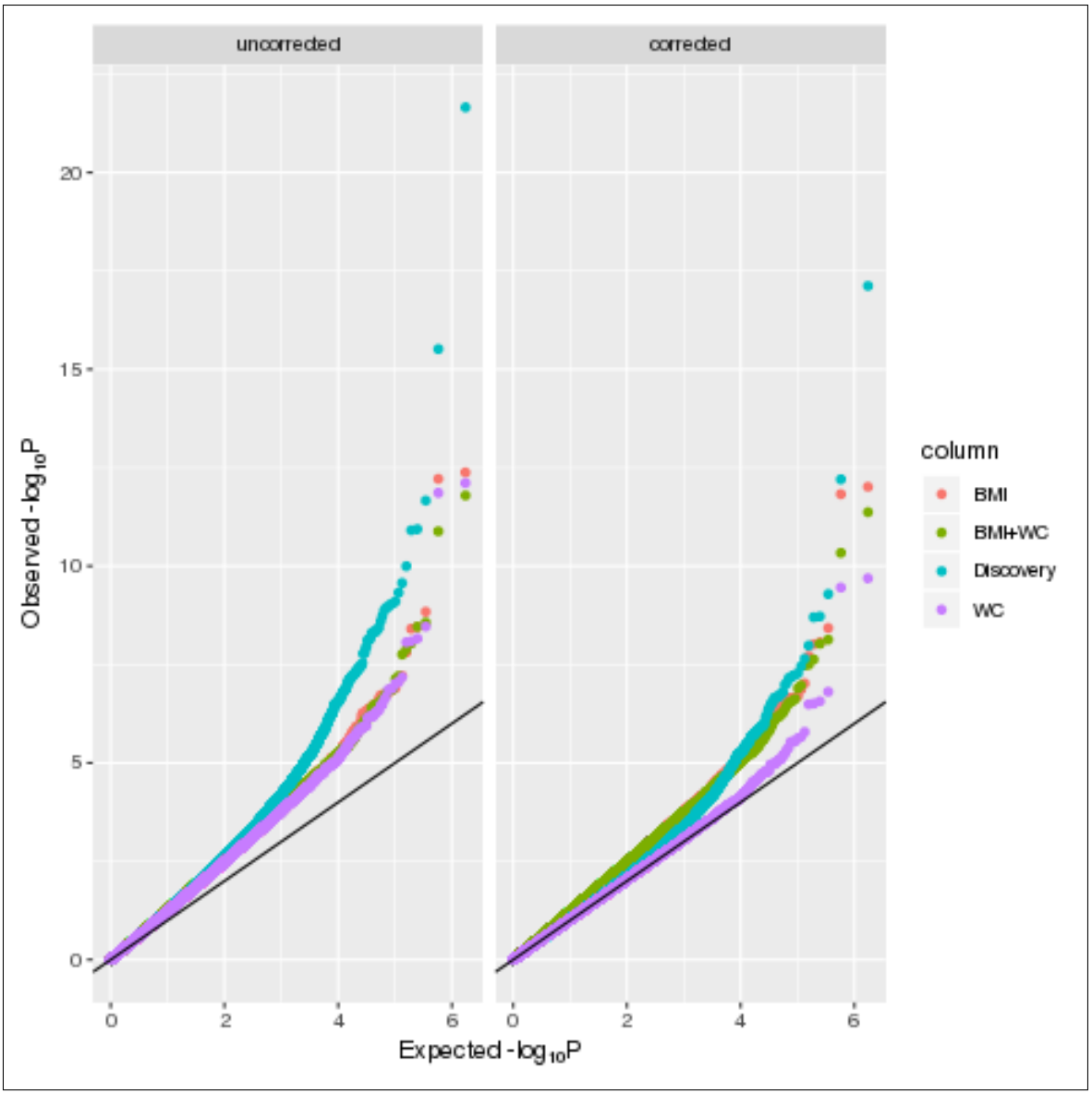


Figure 4.8: Comparison of Quantile-Quantile plots of genome-wide analysis of HDL-C EWAS including Discovery model and models include diet-related covariates. Estimated inflation (λ) according to the bacon R package (van Iterson *et al.*, 2017) for each model is 1.13 (Discovery), 1.02 (BMI), 1.13 (WC) and 1.02 (BMI+WC).

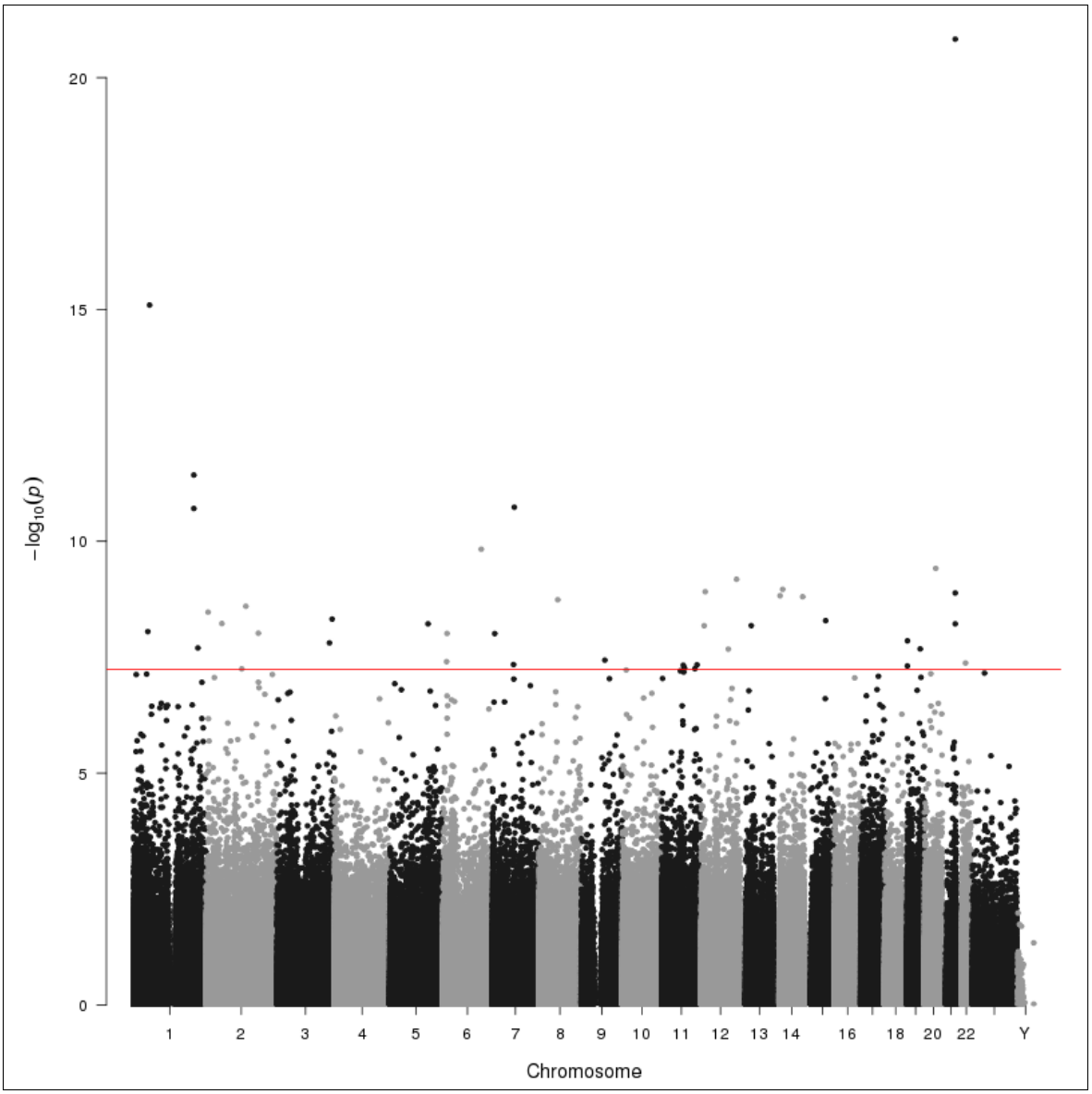


Figure 4.9: Manhattan plot of genome-wide analysis from HDL-C discovery model. Red line denotes genome-wide significance level equivalent to bonferroni corrected p-value <0.05

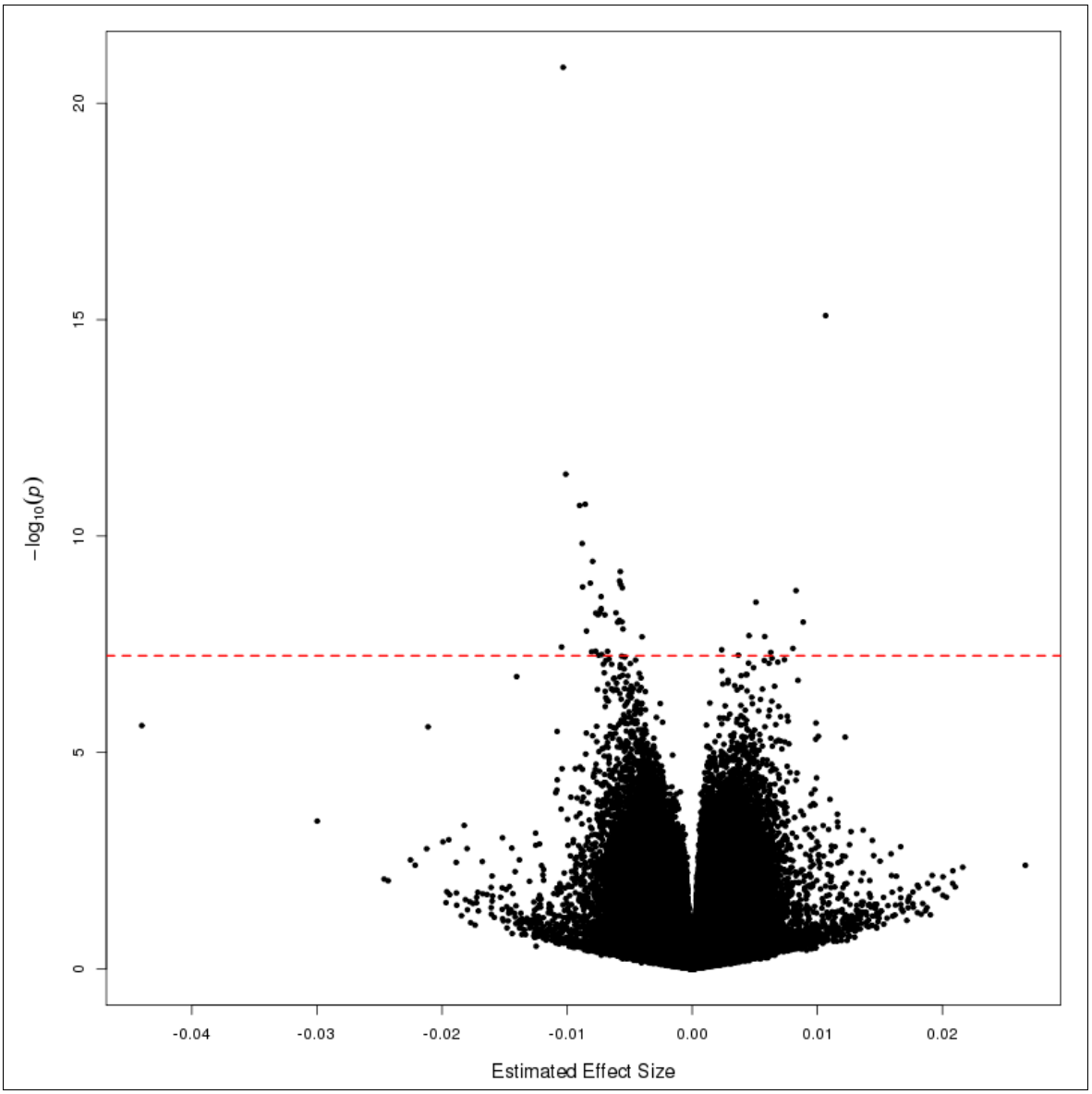


Figure 4.10: Volcano plot of genome-wide analysis from HDL-C discovery model. Red dashed line denotes genome-wide significance level equivalent to bonferroni corrected p-value <0.05

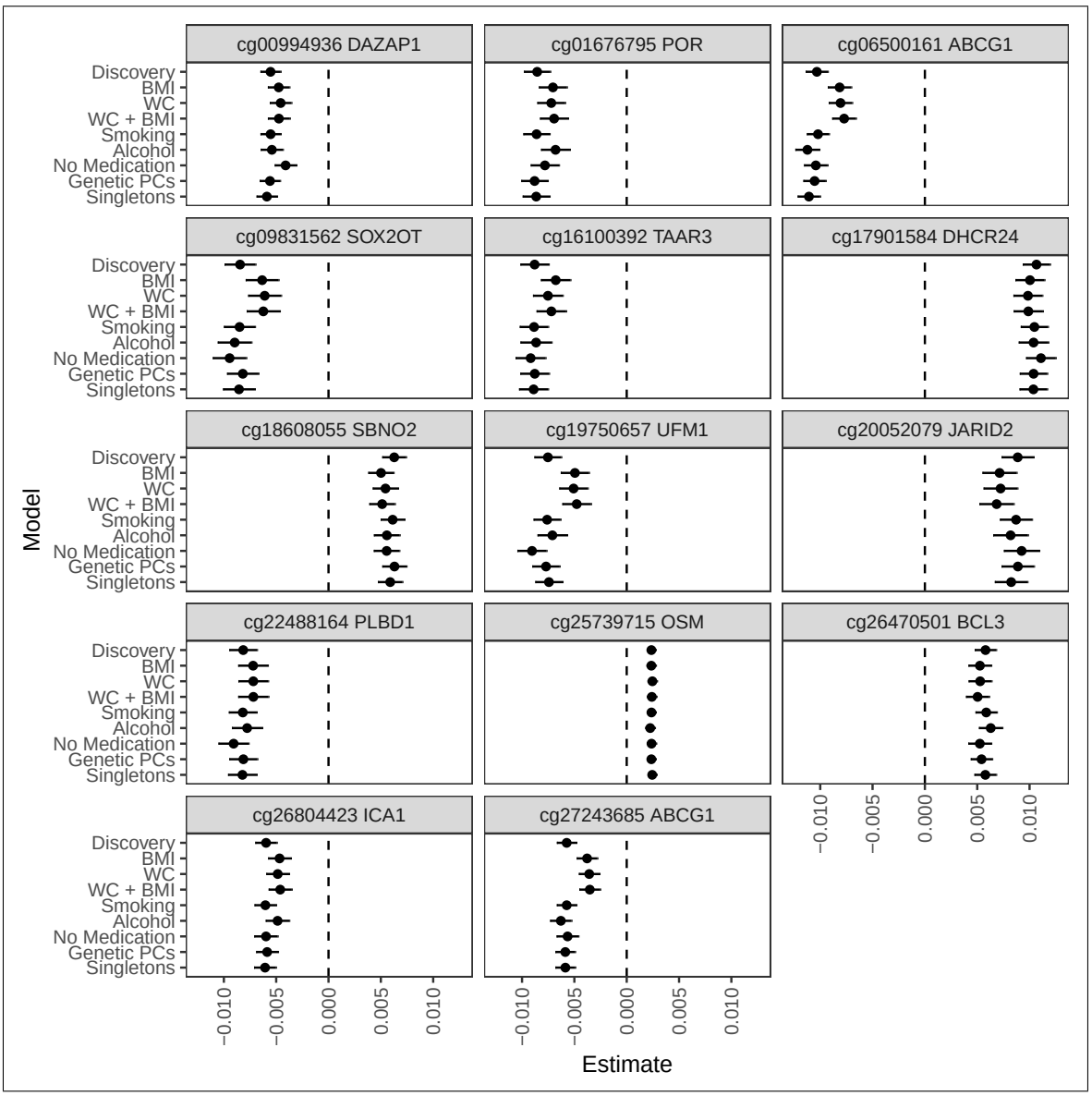


Figure 4.11: Comparison of effect sizes of genome-wide significant probes present on the 450K microarray from HDL-C models including different covariates. Each point corresponds to the unstandardised effect size for each CpG in a given model, where the lines correspond to the size of the standard error associated with the effect size. Discovery refers to the initial model run which includes age, sex, drug status, cell-type composition estimates and plate number. Annotated labels correspond to the covariate included within the discovery analysis. No medication corresponds to the discovery model being run without any participants using lipid-lowering medication and the singleton model refers to all genetically similar participants being excluded from analysis.

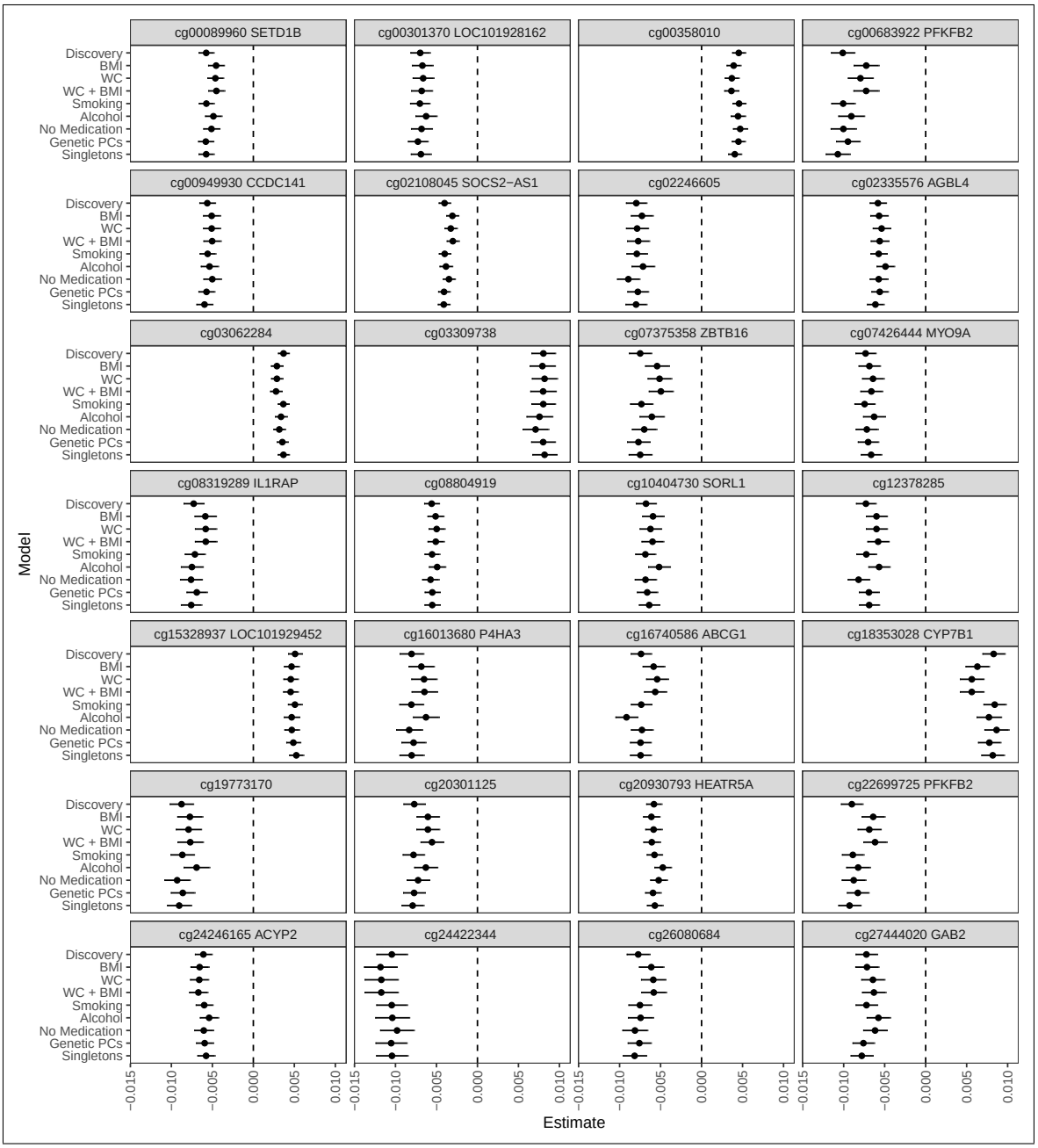


Figure 4.12: Comparison of effect sizes of genome-wide significant probes exclusive to EPIC microarray from HDL-C models including different covariates. Each point corresponds to the unstandardised effect size for each CpG in a given model, where the lines correspond to the size of the standard error associated with the effect size. Discovery refers to the initial model run which includes age, sex, drug status, cell-type composition estimates and plate number. Annotated labels correspond to the covariate included within the discovery analysis. No medication corresponds to the discovery model being run without any participants using lipid-lowering medication and the singleton model refers to all genetically similar participants being excluded from analysis.

4.3.4 Statin-Use

Having examined the various blood-lipid traits, it may be interesting to explore the differential methylation that is associated with statin-use. By using a similar model to the one used to investigate TC, HDL-C and TG however without including a lipid trait in these analyses. Using this approach, I identified 9 genome-wide significant CpG sites that are distinctly novel and annotate to genes that are related to lipid metabolism. Examination of the quantile-quantile plot of the p-values suggests that there is little evidence of test-statistic inflation as there is little deviation from the expected distribution (Figure 4.13).

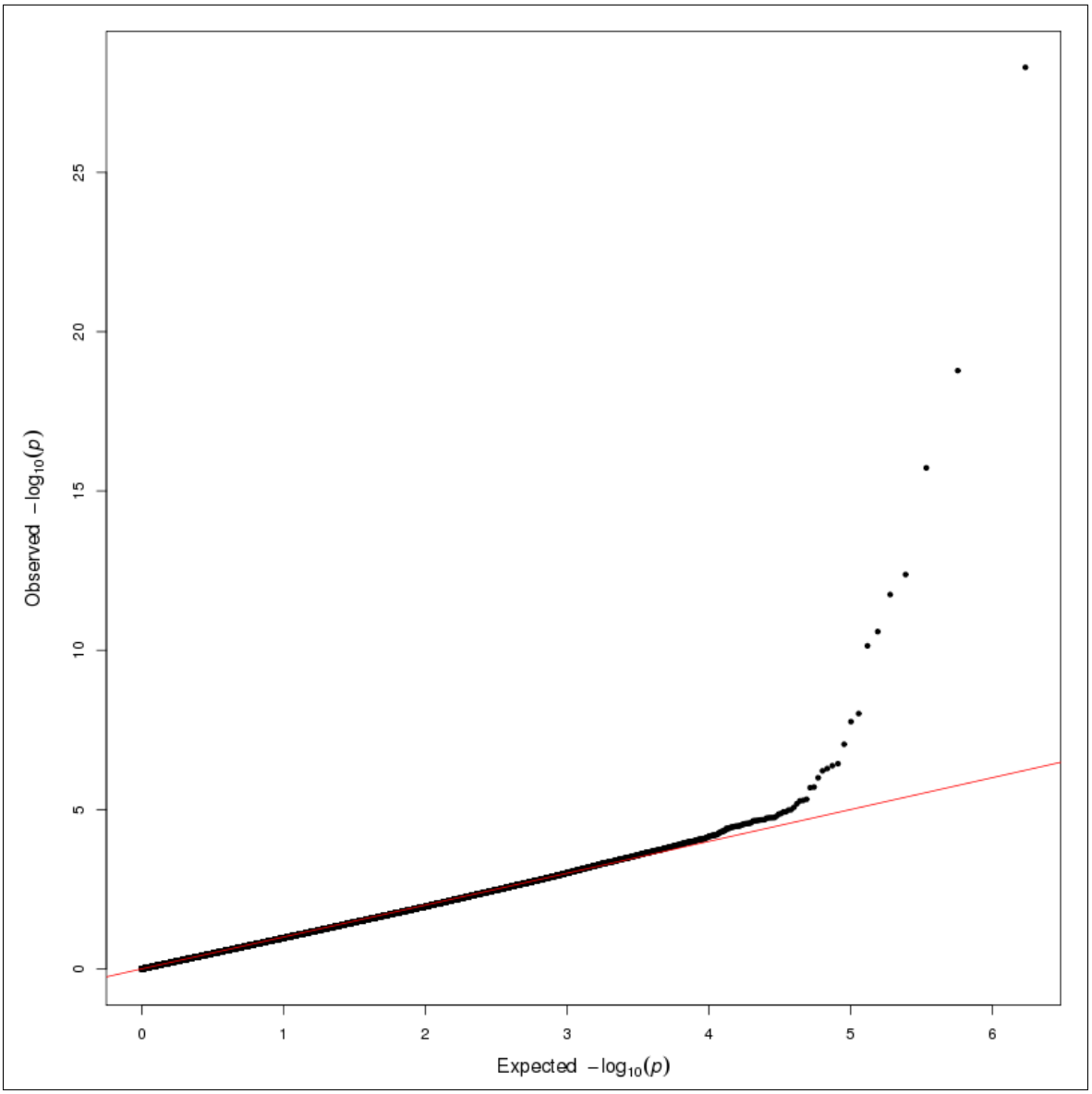


Figure 4.13: Quantile-Quantile plot of genome-wide analysis from Statin-use model

Table 4.5: Top 9 significant loci for Statin-Use EWAS

Probe ID	Gene Name	CHR	Location	Epic	Effect Size	adj p-value	Previous Association
cg17901584	DHCR24	1	55353706	FALSE	-0.03908	4.332e-23	Braun *et al.* (2017); Dekkers *et al.* (2016b); Hedman *et al.* (2017)
cg06500161	ABCG1	21	43656587	FALSE	0.02563	1.431e-13	Hedman *et al.* (2017); Braun *et al.* (2017); Dekkers *et al.* (2016b); Sayols-Baixeras *et al.* (2016b); Dayeh *et al.* (2016); Pfeiffer *et al.* (2015)
cg05119988	SC4MOL	4	166251189	FALSE	-0.02759	1.613e-10	
cg27243685	ABCG1	21	43642366	FALSE	0.01772	3.599e-07	Hedman *et al.* (2017); Braun *et al.* (2017); Dekkers *et al.* (2016b); Sayols-Baixeras *et al.* (2016b); Dayeh *et al.* (2016); Pfeiffer *et al.* (2015)
cg15659943	ABCA1	9	107631656	TRUE	0.01571	1.527e-06	
cg09646062	DHCR24	1	55324150	TRUE	0.0163	2.219e-05	
cg15128785	SREBF2	22	42230879	TRUE	-0.02619	6.21e-05	
cg12403973	SREBF2	22	42230899	TRUE	-0.02665	0.008327	
cg10177197	DHCR24	1	55316481	FALSE	0.01129	0.01494	

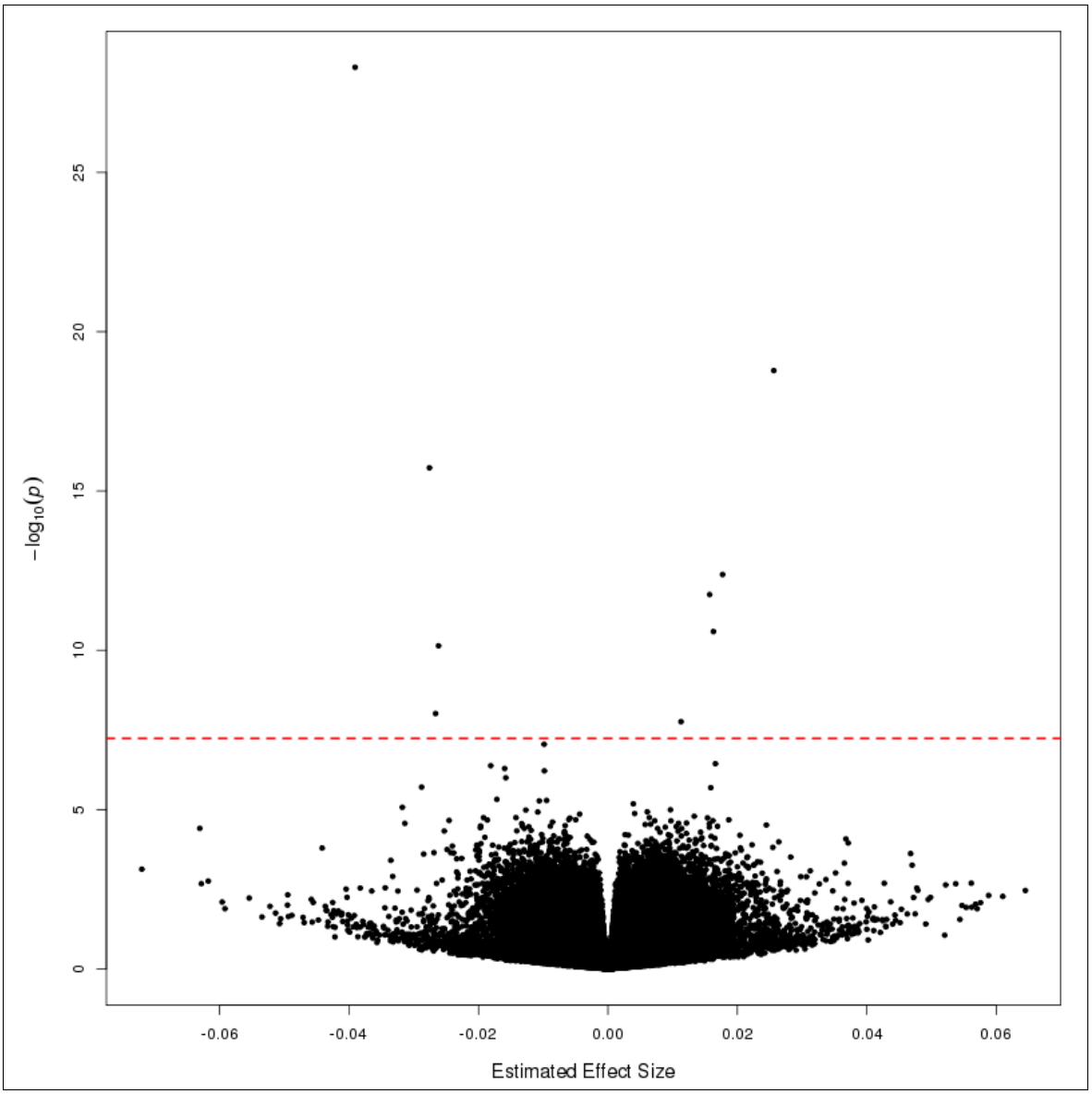


Figure 4.14: Volcano plot of genome-wide analysis from Statin-use model. Red dashed line denotes genome-wide significance level equivalent to bonferroni corrected p-value <0.05

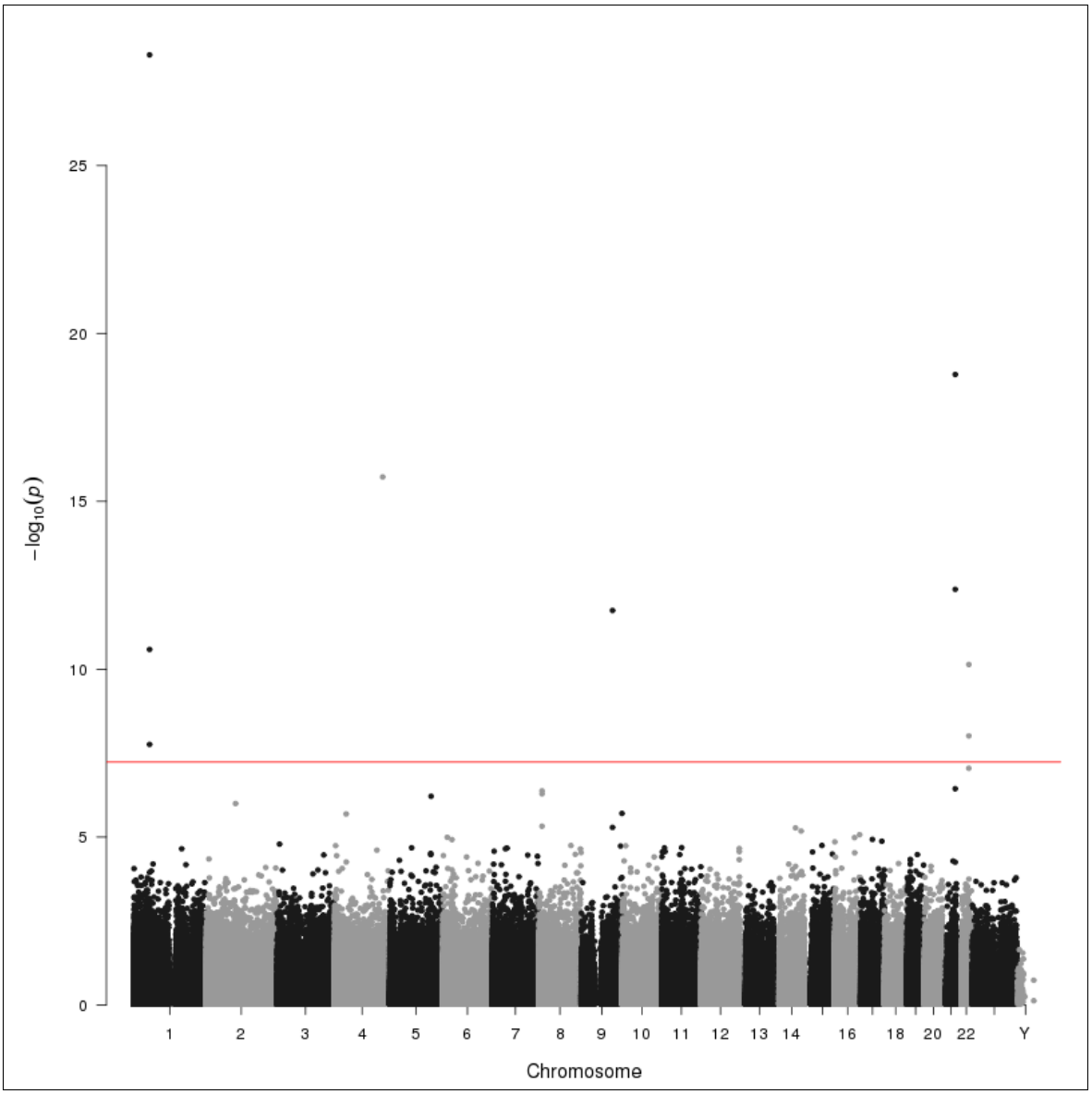


Figure 4.15: Manhattan plot of genome-wide analysis from Statin-use model. Red line denotes genome-wide significance level equivalent to bonferroni corrected p-value <0.05

4.4 Discussion

I present the first EWAS between blood-lipid concentrations and DNA methylation to be performed on the EPIC microarray. This provided a prime opportunity to reproduce the findings from previous studies and explore novel findings that are exclusive to the EPIC microarray. In total, I identify 37 novel CpGs associated with both TG and HDL-C concentrations (9 and 28 CpGs respectively). Many of these novel CpGs are annotated to genes that have been previously reported. The probes that did not appear to annotate to a gene have yet to be fully explored but could be distal to genes or enhancer regions.

A summary of some of the more notable associations identified in these analyses are presented in Table 4.6. I identify numerous genes that directly relate to cholesterol or lipid biology. Many of the identified genes have been reported to be associated with other blood-lipid phenotypes or metabolic traits.

There are two genes that I believe should be discussed in further detail. The first gene is JARID2 of which I identify four CpGs to be associated with both HDL-C and TG concentration. JARID2 encodes a DNA binding molecule which helps form Polycomb repressive complexes which play an important role in stem cell pluripotency (Jones & Wang, 2010). In humans, JARID2 is mostly active during prenatal development which has been identified in an EWAS looking at gestational age (Spiers *et al.*, 2015) but more interesting is that in adult humans JARID2 is almost exclusively expressed in the heart. Mouse models have shown that JARID2 is very important in heart development as JARID2 knockouts and heterozygotes exhibit heart malformations and lethality (Cho *et al.*, 2018). This could suggest that JARID2 could play an important role in the development of the human heart and that alteration of methylation patterns could play a role in the development of CVDs. In embryo's, the DNA methylation state of JARID2 is highly methylated but decreases rapidly as gestational age increases. Here I observe a decrease in methylation with increasing TG concentration and an increase in methylation coupled with increasing HDL-C. What makes this finding so interesting is the fact that it has not been presented in any other studies that look at blood-lipid levels and only identified in a single study that looked at BMI (Wahl *et al.*, 2017). It is not known whether or not these associations are identified in this study because the DNA methylation patterns were assayed on a different platform or that the population from which the

Table 4.6: Summary of results obtained from TC, TG, HDL-C and statin-use EWAS from the Understanding UK Household Dataset

Gene Name	CpG (Association)	Gene Function	Reported by [Association]
ABCA1	cg15659943 (Statin-use)	Cholesterol Transport	Guay *et al.* (2012)[HDL]
ABCG1	cg16740586 (TG) cg06500161 (TG, HDL-C) cg27243685 (TG)	Cholesterol Transport	Pfeiffer *et al.* (2015) [TG, HDL] Hedman *et al.* (2017) [TG, HDL] Braun *et al.* (2017)[TG, HDL] Dekkers *et al.* (2016b)[HDL] Sayols-Baixeras *et al.* (2016b)[TG, HDL]
CPT1A	cg00574958 (TG) cg17058475 (TG) cg05325762 (TG)	Fatty Acid Oxidation	Gagnon *et al.* (2014) Dekkers *et al.* (2016b)[TG] Braun *et al.* (2017) [TG] Hedman *et al.* (2017)[TG] Sayols-Baixeras *et al.* (2016b)[TG] Chasman *et al.* (2009)[GWAS] Kettunen *et al.* (2012)[GWAS]
DHCR24	cg10073091 (TC) cg17901584 (HDL, Statin-use) cg10177197 (Statin-use) cg09646062 (Statin-use)	Cholesterol Biosynthesis	Dekkers *et al.* (2016b)[LDL] Hedman *et al.* (2017) [TC, HDL] Braun *et al.* (2017) [TG, HDL] Demerath *et al.* (2015)[BMI] Kazmi *et al.* (2017)
JARID2	cg20052079 (TG, HDL) cg03173502 (TG) cg13027183 (TG) cg13500852 (TG)	Embryonic Development	Wahl *et al.* (2017)[BMI] Spiers *et al.* (2015)[Gestational Age]
NLRC5	cg07839457 (TC)	Cytokine Response	Hedman *et al.* (2017)[TC] Meeks *et al.* (2017)[Obesity] Zhang *et al.* (2016)[HIV] Zhang *et al.* (2017)[Hepatitis C]
PFKFB2	cg00683922 (HDL, TG) cg22699725 (HDL)	Glycolysis	
SC4MOL	cg05119988 (Statin-Use)	Cholesterol Biosynthesis	
SCD	cg03440556 (TC)	Fatty Acid Biosynthesis	Crujeiras *et al.* (2016)[Insulin resistance] Skuladottir *et al.* (2016)[Sleep-deprivation]
SREBF1	cg11024682 (TG)	Lipid Homeostasis	Pfeiffer *et al.* (2015)[TG] Hedman *et al.* (2017) [TG, HDL] Sayols-Baixeras *et al.* (2016b) [TG, HDL] Dekkers *et al.* (2016b) [TG] (Demerath *et al.*, 2015)[BMI]
SREBF2	cg09978077 (TC) cg15128785 (Statin-Use) cg12403973 (Statin-Use)	Lipid Homeostasis	Hedman *et al.* (2017)[TC] Sayols-Baixeras *et al.* (2016b) [TC]
TXNIP	cg19693031 (TG)	Cellular Redox Signaling	Pfeiffer *et al.* (2015)[TG] Hedman *et al.* (2017) [TG] Sayols-Baixeras *et al.* (2016b)[TG] Dayeh *et al.* (2016)[T2D]

participants come from could be influencing these results. Thus it is possible that there four associations could be driven by genetic variation. Although, the function of JARID2 is mainly implicated in the heart, it may also have a function in blood however this would need further investigation.

The second gene that is interesting in this study is novel and specific to the EPIC array. The PFKFB2 gene encodes a protein that is important in glycolysis. The effect sizes do change slightly with the inclusion of BMI and Waist circumference which suggests that the association between DNA methylation and PFKFB2 may be driven by diet rather than individual blood-lipid concentration.

There are also a number of results that were detected in this statin-use EWAS. This discovery EWAS identified 9 CpGs that annotated to other lipid-related genes. Novel genes include ABCA1 and SC4MOL alongside associations with numerous CpGs within DHCR24 and SREBF2. Of these, the SC4MOL gene and ABCA1 gene have not been identified in previous EWAS but have important roles in lipid metabolism. As this is the first statin-use EWAS to be reported there are no results from previous studies to make comparisons with. Thus it is difficult to have any confidence in these results. The quantile-quantile plot suggests there is no test-statistic inflation which is encouraging alongside many of the top results are annotated to genes with clear lipid-related function. It is entirely possible that the results from this statin-use EWAS could essentially be a low-high cholesterol case-control EWAS as statins are usually prescribed to individuals who have elevated cholesterol levels. Even though those on statins may lower cholesterol, it is possible that there are other factors that could be influencing the methylation patterns thus the changes in DNA methylation patterns may not be related to lipid-lowering medication. This will need careful handling and further investigation as it is likely that the results from this statin-use EWAS could be confounded by other blood-lipid profiles as there is considerable overlap in some of the key findings in all of the EWAS.

Overall none of the results presented in this study appear to be out of the ordinary. Inspection of the quantile-quantile plots for the HDL-C and TG analysis also suggests there is only a small amount of inflation in the discovery analysis. Although the inclusion of BMI and WC does reduce the amount of inflation for both the TG and HDL-C models and the number of genome-wide significant results, there is only a small amount of inflation to begin with. I decided to use the results from the discovery EWAS as

the genome-wide results from the discovery analysis were comparable with previous studies.

The genome-wide significant results from both the HDL-C and TG EWAS appear to be robust to other potential sources of confounding. When these models were performed with other known risk factors of CVDs which could also influence the blood-lipid concentrations most probes showed the same size of effect. In circumstances where the effect size was attenuated in some form, this was almost exclusively caused by the inclusion of diet-related variables (BMI or WC) within the model. The results appear to be robust to genetic confounding as demonstrated in the singletons and models that include the principal components derived from genetic data.

There is evidence which suggests there are small amounts of colinearity between some of the blood-lipid phenotypes that were being investigated. In particular a high correlation between TC and LDL-C concentrations and an anti-correlation between HDL-C and TG concentrations. The models used to obtain these results contained a single blood-lipid phenotype as the variable of interest. This is in agreement with the approach of all the past EWAS that have been performed. However, it is possible that there is the potential that these blood-lipid phenotypes could interact in some way and therefore should be included within the models. Despite the potential of confounding by other blood-lipid measurements the inclusion of BMI and waist circumference for genome-wide significant results only affected a few of the results in a very small manner (Figures 4.11, 4.12, 4.6 & 4.7).

The most substantial limitation of this study is the lack of a replication dataset to reproduce the findings of the discovery EWAS. At this moment in time that are no comparably large datasets that have been assayed on the EPIC array that also have blood-lipid measurements to facilitate such investigation. A comprehensive review of past results shows that the majority of the results obtained here are in agreement with past results. Therefore I am confident that the findings presented here are robust and meaningful despite the lack of reproduction.

Extra care must be taken when interpreting these results as they do not describe any direction of cause.

Rather these results only show that a relationship between DNA methylation and a given blood-lipid phenotype exists. To discern a direction of cause I could employ Mendelian randomisation to establish whether or not the DNA methylation patterns are influenced by genetic variation. Dekkers *et al.* (2016b) implement Mendelian randomisation in their lipid study and had found that the increase in blood-lipid concentrations led to changes in methylation in the loci (cg06500161) that satisfied the requirements for testing. As genotyping data does exist for the Understanding Society data, it is possible to perform Mendelian randomisation and would be a worthwhile investigation to perform in the future. A large number of mQTLs (CpGs that are directly influenced by genetic variation) have been identified using the Understanding Society dataset used in this EWAS (Hannon *et al.*, 2018). Therefore it is possible that the results from this mQTL analyses could be used to identify CpGs that are potentially mediated by genetic variation.

For these analyses I only filter the dataset according to pfilter as described in wateRmelon (Pidsley *et al.*, 2013) and for SNP heterozygotes using pwod. This does not include any of bioinformatically derived probes such as those described by Pidsley *et al.* (2016) which include probes which are known to be cross-reactive or have SNPs that underly the given probe sequence. Moreover, if I were to remove all of the probes described by Pidsley *et al.* (2016) I would remove one of the strongest signals that is presented in this study (cg06500161). According to these probe-lists, cg06500161 has a genetic variant which is inside the body of the probe sequence. Considering that this finding is ubiquitously reported in blood-lipid EWAS, from multiple populations and has been shown not to be influenced by genetic variation I decided to keep this probe in this analysis as it is an important biological replication. While this signal may be robust to technical problems, it is possible that the other probes which are identified in the probe list described by Pidsley *et al.* (2016) could be confounded I could equally be missing out on potentially meaningful findings if I were to remove them from analysis. Instead of worrying about which probes should be tested during an EWAS I believe that further validation of significant hits would uncover the true biological significance of presented findings.

As many as 20 of the findings presented in this study have been previously identified in lipid-related EWAS performed on the 450K microarray. This demonstrates that the EPIC microarray is a reliable tool

and is likely to be useful for EWAS going forward. Functionally the EPIC array offers an incremental improvement over the 450K microarray, but the coverage of the EPIC array still queries approximately 3% of the CpG sites within the human genome. It is likely that the previous studies using the 450K have identified most of the results that would be identified by the EPIC array in the future, but there are now new opportunities to explore the epigenetic landscape of functional genomic elements that have not been as well described.

EWAS offer a mechanistic understanding of the events that lead to or are caused by heightened blood-lipid levels and instances of CVD. However, they do not infer a direction of cause and require alternative study designs or complicated methodology (e.g. Mendelian randomisation). On the contrary, GWAS can infer such a direction of cause but may not offer functional explanations of the genetic variants that are associated with a given trait. Thus belies the strength of examining the epigenome as it is dynamic and can respond to external stimuli such as lifestyle choices and could provide insights into why individuals with the same genetic variations will exhibit different phenotypes.

4.5 Conclusion

In this chapter, I present the first EWAS between DNA methylation and blood-lipid phenotypes using the EPIC microarray. I also describe the first EWAS to explore statin use as an independent trait. In total, I identify many loci that are associated with both lipid metabolism and biosynthesis which have been previously identified in studies that look at a variety of metabolic traits.

Of particular note, I identify two striking results that have yet to be explored in any detail. In both the HDL-C and TG EWAS I identify numerous CpGs within JARID2 and PFKFB2 which are known express almost exclusively in the heart. As these genes are associated with elevated TG and HDL-C concentrations, there is the potential for complex interplay between these two genes and CVDs which needs to be explored further but could be both be interesting candidates for future investigation.

Chapter 5

Large scale analyses using the 450k microarray

To demonstrate the types of analyses that are possible using the bigmelon software I perform two preliminary analyses that both involve a large number of samples. This chapter is subsequently split into two part to describe each analysis.

The first part of this chapter seeks to identify how probes on the 450K microarray behave with respects to various characteristics. The purpose of this is to produce a supplementary probe lists which can be used in conjunction with bioinformatically derived probe-lists such as those present by Zhou *et al.* (2017) and Pidsley *et al.* (2016). These probe lists classify a number of probes on the 450K and EPIC array (respectively) which could be confounded by cross-hybridisation or underlying SNPs in the probe sequences which can affect the accuracy of the DNA methylation measurements of listed probes.

The second part of this chapter examines how tissue-specific DNA methylation patterns correlate with tissue-specific gene expression. This analysis was primarily inspired by the preprint article by Ford *et al.* (2017) where DNA methylation was induced at numerous promoter regions to examine if the induced change in methylation were accompanied by a change in gene expression. In addition, numerous studies

negatively correlate promoter DNA methylation with transcriptional activity (Weber *et al.*, 2007). However, gene-body methylation is positively correlated with gene expression (Yang *et al.*, 2014). As the 450k microarray has been used to assay thousands of samples from a wide variety of tissues, it is possible to use the wealth of data that is publicly available to examine the relationship between gene-region DNA methylation and gene expression in a tissue-specific manner.

5.1 General Introduction

Over the span of a decade, DNA methylation microarrays have undergone two sets of revision. From the 27K (Bibikova *et al.*, 2009) to the extremely popular 450K (Bibikova *et al.*, 2011) and now with the newly released EPIC array (Moran *et al.*, 2016), it is expected that scientists will continue to use these cost-effective platforms and deposit subsequent data to publicly available resources. With this continuous growth of data, it expected that large-scale analyses using thousands of samples from these repositories will eventually be presented. However, there have been very few studies that exceed more than a couple thousand samples or combine data from multiple sources. This is surprising considering the efforts that have been made in improving the techniques and physical compute power to perform such analyses. Exemplary analyses such as Horvath (2013)s 'epigenetic clock' and the collation of the Marmal-aid database by Lowe & Rakyan (2013) were both performed early on during the 450K era, but very few have been performed since.

Horvath (2013) focused his analysis on CpGs that were shared between the 27K and 450K arrays. This enabled the analysis of more than 4,000 samples but consequently missed out on a considerable proportion of the 450K. This approach proved to favourable as Horvath (2013)was able to produce a biologically tangible biomarker ('epigenetic age') that could be derived from DNA methylation patterns which has become extremely popular and has been associated to various adverse conditions including all-cause mortality (Marioni *et al.*, 2015).

Marmal-aid (Lowe & Rakyan, 2013) managed to collect the DNA methylation patterns of more than

14,000 samples from a variety of tissues by storing each sample in native R data format (.Rdata file format), but this approach still required each file to be loaded into memory when requested. Although it was a tremendous resource, it sadly did not take off in popularity and has since been taken offline due to a lack of usage. A distinct advantage to having such a readily accessible resource that contains so many samples is that it made analyses such as those looking at the methylation state between a variety of tissues at given loci considerably easier as it removed the need to search online or collaborate with others to find relevant samples.

The aims of this chapter are to:

1. Characterise some potential weakness of the 450k microarray platform and highlight areas where caution should be taken.
2. Examine the relationships between gene-region DNA methylation patterns and gene expression levels in a tissue-specific manner.

5.1.1 General Methods

5.1.1.1 Datasets

To examine the characteristics of the 450K microarray as described in Aim 1, I first require a large enough number of samples to produce conclusions about how each probe on the 450K performs. The Marmal-Aid dataset is sufficiently large enough to produce such conclusions but is unsuitable for my purposes as it only contains raw and normalised β values. Despite the lack of raw data, it is still useful for some characteristics I wish to explore. Briefly, the Marmal-Aid dataset contains DNA methylation data from 14,573 samples from a variety of tissues obtained from data that was deposited to GEO (up until late 2013) and TCGA. Data from Marmal-aid was downloaded according to the tutorial (currently unavailable) in '.Rdata' format and each sample was converted into a single gds file. This is the same dataset that was used in Chapter 3 to investigate the scalability of the bigmelon R package as it was the largest dataset available at the time.

Due to the Marmal-Aid dataset only containing β values I sought to create a new dataset which contained the information that is provided by the raw idat files, specifically the bead counts and raw signal intensities. More than 60,000 450K microarrays have been deposited to GEO under the platform accession number GPL13534, as of July 2017. This number has since increased to more than 75,000 samples. However due to the time when the data was initially obtained only samples that were submitted before July 2017 were considered in the following analysis.

Combination from so many sources posed one significant problem when collating data from GEO. This problem was caused by inconsistencies between supplied annotations for the data by the data originators. When data is submitted to GEO or other repositories, the curators are encouraged to follow the minimum information about microarray experiment (Brazma *et al.*, 2001, MIAME) guidelines. Despite these guidelines, the amount of information that accompanies the data varies considerably which makes it difficult to combine the data from multiple sources. As my approach was to obtain as many idat files as possible this lead to a wide variety of annotations across datasets.

When constructing the dataset of multiple datasets from GEO, I decided to exclude datasets which had fewer than 50 samples. This decision was made to avoid swamping the dataset with numerous, systematic batch effects. The flowchart presented in Figure 5.1a describes how I arrive at a total n of 15,773 unique samples obtained from 91 different datasets following the deduplication of identical microarray barcodes. At the time this roughly equated to approximately one quarter of all of the 450K microarray data that had been submitted to GEO.

The datasets that had fulfilled the brief criteria (having raw idat files and n $>$ 50) were imported into R using the geotogds function from bigmelon which automatically downloads and imports the raw data and phenotypic annotation into separate gds files. Phenotypic annotations for characteristics such as sex, age and tissue type were derived during data sanitisation using regular expressions and normalised into a consistent format. For example, sex was obtained by searching for "Sex", "Gender", "Male", "Female", "M" or "F" in provided annotations and converted into a binary trait. Missing samples were left as NA or were attempted to be predicted using the predictSex function from wateRmelon. Age was similarly

resolved by searching for relevant keywords (e.g. 'years', 'age'). All ages were converted to years and any ages labelled as prebirth were treated as 0. Missing ages were imputed according to Horvath's Epigenetic clock using the agep function. Lastly, tissue annotations were obtained through regular expressions search for "Tissue" or "Source" and then manually verified with associated publications. Samples with no discernible tissue annotation were treated as NA as there was no reliable way to verify which tissue unannotated samples had come from.

5.1.1.2 Quality Control and Normalisation

The large size of the GEO dataset presented some problems with respects to both normalisation and quality control of the data. Firstly due to the data being comprised of numerous sources and disease applying any normalisation methodology risks removing any small meaningful differences between samples that would otherwise be detected. Lowe & Rakyan (2013) opted to normalise the data based on a quantile subsampling routine which randomly selected 10,000 probes to produce quantiles that were then interpolated back into a full β distribution. This approach was used due to memory constraints that were imposed due to the sheer size of the dataset. Instead, I could have used a quantile-based method such as dasen or a more sensitive approach such as funnorm or ssNoob which can be implemented on a single sample basis (Min *et al.*, 2018; Fortin *et al.*, 2016) I decided to leave the dataset unnormalised because the dataset is large enough that it should be robust to potential confounding. Likewise, quality controlling the dataset presents similar issues. Data-outlier detection tools (outlyx, qual) would not be effective due to the heterogeneous nature of the data. Similarly, control-probe based methods may not be informative as expected as demonstrated in Chapter 2.

5.2 Part 1

During the seven years of the 450Ks extensive use, there have been few attempts to characterise certain characteristics directly relevant to the performance of the microarray. Such characterisation could be useful in circumstances where the idat files are no available or if the sample size is too small to draw a direct

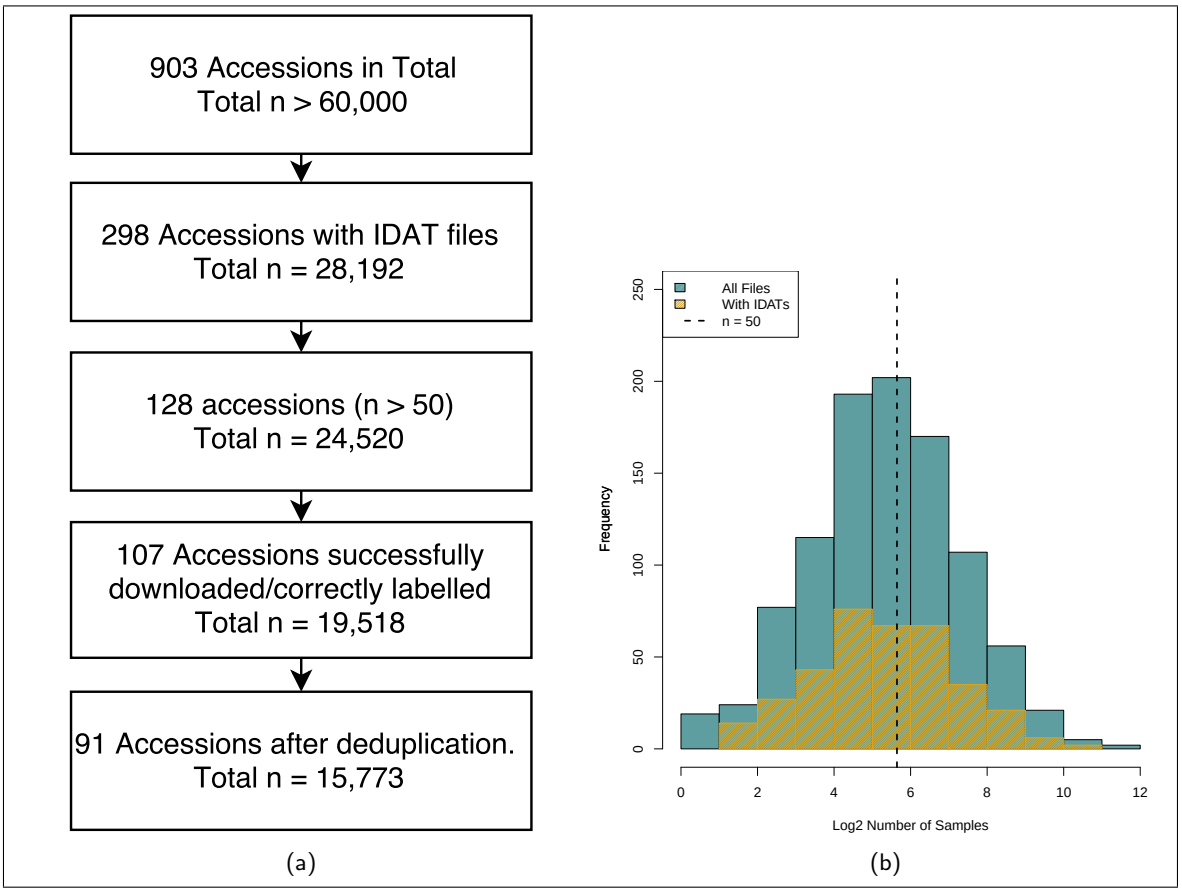


Figure 5.1: Description of process of selection from GEO & Distribtuion of sample size in dataset from GEO. **a)** Flowchart describing the process of selection of GEO accession until July 2017 (under GPL14534) to use in further analysis. **b)** Distribution of sample size for each GEO accession under GPL14534 (July 2017)

conclusion regarding how certain probes are behaving. I feel it is genuinely worthwhile to explore these characteristics with the intention to supplement the widely used probe lists such as those presented by Zhou *et al.* (2017) and Pidsley *et al.* (2016) but with a focus on how they behave in a data-driven context. To do this, I consider more than 15,000 samples from a variety of tissues assayed on the 450K microarray and characterise a variety of properties that are used in quality control pipelines in small datasets, with the intention to extend the guidelines to a larger scale. Providing a supplementary probe list of poorly performing probes will enable scientists to perform some level of quality control on data when the raw data that is usually required to perform such filtering is unavailable.

Characterising the poorly performing probes of the 450K will mostly require exploratory analysis of general trends of certain measures. Such measures that can be used are those that are used to often quality control the data such as pfilter, which uses the detection p-values and beadcounts to identify a set of probes that have 'failed' to generate a reliable signal. Other alternatives also involve identifying probes that exhibiting small minor allele frequencies and are therefore susceptible to genetic confounding if not checked for. This is provided by the pwod function in wateRmelon (also described in Chapter 2), but probes that are in Hardy-Weinberg equilibrium could be detected as I have a large enough sample size. Lastly, I explore how much each probe can vary across the many tissues and samples I have acquired. Considering I have 15,773 samples and access to the Marmal-Aid dataset to reproduce some of the characteristics I suspect that I have enough data to explore these aspects robustly. In addition to this, there is an opportunity to explore the quality control thresholds that are typically used on smaller datasets on a larger scale as could be that the threshold currently used may not be suitable for larger datasets. For example, if using pfilter, a probe in a dataset of 100 samples would only require 5 samples to be considered for removal while if the sample threshold was scaled up to 10,000 samples a staggering 500 samples would be required to flag the very same probe. While the proportion of data is still the same - the threshold required to remove features is much more penalising to a smaller dataset.

5.2.1 Part 1 Methods

To extend Zhou *et al.* (2017)s probe list, I considered the following characteristics:

1. **Bead Counts** – The BeadChip microarrays contain thousands of sequence-specific oligonucleotide coated beads to which DNA can hybridise to. Generally speaking the higher the number of beads that are present on the microarray for a given probe sequence the more reliable (due to higher the overall representation) a signal obtained from that specific probe is. In EWAS it has become common practice to remove any signals that are obtained from probes where the number of beads are below a certain number or a certain number in a proportion of the total number of samples. For example, the pfilter tool removes probes when the beadcount is < 3 in $> 5\%$ of samples. This threshold was arbitrarily defined and may not be appropriate for large datasets.

2. **Detection p-values** – Detection p-values are a measure of the error with respects to the signal obtained from a probe to the background signal. Similar to beadcounts these detection p-values can be used to remove signals based on various thresholds. In most cases, the recommended approach is to simply remove signals (set individual β values to NA) that have a corresponding a detection p-value of ≥ 0.01 or ≥ 0.05, but other methods have been suggested. Lehne *et al.* (2015) suggests excluding based on a detection p-value derived from a low enough p-value such that all signals obtained from the Y-Chromosome from females samples are excluded from analysis. Another alternative is using pfilter where probes are removed if $> 1\%$ of samples have a detection p-value > 0.05. It should be noted that the calculations of detection p-values are intrinsically different between software. Specifically, the method used to derive detection p-values in R using minfi will be different from what is provided by Illumina's GenomeStudio or by other packages such as methylumi. Thus the process of removing low-confidence measurements may vary between analysis.

3. **Probe Variance** – Statistical tests used in EWAS require that there is sufficient variance between groups or per standard deviation of a continuous variable for an effect to be estimated. In rare situations, it is possible that a probe can exhibit little to no variation such that any discernible effect would not be detected, even with thousands of samples. Therefore removing such probes could be beneficial when performing many statistical tests as it can both reduce the multiple testing thresholds and save in compute time as there would be fewer loci being tests. There are a few drawbacks to this approach. Targeting probes with low variance using an arbitrary threshold will select more Type I probes than Type II probes due to Type I probes having a distribution with peaks

closer to 0 and 1 while Type II probes peaks at 0.9 due to background signal. Furthermore, there may be circumstances where highly invariable probes could exhibit wild variation in very specific circumstances (i.e. with rare diseases). As a result, it would be useful to explore the probes which a usually invariable to identify circumstances when low variance probes vary slightly.

4. **Minor Allele Frequencies** – The last characteristic I intend to explore is the common practice of removing probes that exhibit SNP-like distributions. The preexisting probe lists already describe the probes which overlap with public SNPs but do not account for private SNPs and features that can exhibit minor allele frequencies. Thus a sample-wide scan of each probe to find features that display a SNP-like distribution (in the case of β values this is a trimodal distribution with peaks typically at 0.25, 0.50 and 0.75). I use k-means clustering to identify probes that display clusters in Hardy-Weinberg equilibrium. By identifying probes that exhibit SNP-like distributions it could be possible to reduce the amount of genetic confounding in any given analysis.

By identifying a comprehensive set of probes that exhibited abnormalities on a large enough scale, it would not be unreasonable to assume that any probe that is identified in this way could be problematic in both large and small scale analyses. Moreover, when the ability to perform meaningful quality control is limited, it is possible to make use of this analysis to prune the data accordingly as probes identified according to this analysis have been determined to be problematic in some form.

5.2.2 Part 1 Results

Examining the quality of the data within the GEO dataset requires extremely careful and time-consuming decision making to ensure that each sample would have been equally considered. Due to missing annotations, the identification of sex-mismatches or sample-swaps was not possible for the majority of datasets. As a result, the provided annotation for all samples were treated as accurate. When checking the quality of the data using methods that look at the control probes (e.g. bscon) it can be seen that the quality of the DNA varies considerably between datasets (Figure 5.2). Application of a generous threshold of 80% bisulfite conversion shows that the majority of datasets are of good quality and display tight quantile distributions around 90%. As many as 7 datasets (of 91) appeared to have performed poorly according

to this tool. The worst performing dataset according to bscon, GSE79009, assayed DNA obtained from Schwannoma (Agnihotri *et al.*, 2016) where the quality of the DNA could be potentially scarce when compared to the availability of whole blood DNA. Following the conclusions of Chapter 2 of this thesis, I reason that a sample can be of usable quality despite failing control-probe based quality control and therefore still included in downstream analysis.

Similarly, the application of data-driven tools such as pwod or outlyx would also be difficult to apply on the entire GEO dataset. Applying these tools on the entire complement of 15,773 samples would likely yield no meaningful results as these tools perform poorly on heterogeneous data. Application on a per-dataset basis would penalise datasets that contain heterogeneous tissues or have a complex design where specific groups would need to be tested individually. Due to incomplete annotations, it becomes very difficult to appropriate devise quality control pipelines that would be appropriate for each given dataset. Thus the decision to not apply any form of quality control to the GEO dataset was made for the following analyses.

5.2.2.1 Bead Counts

The first characteristic I explore is the overall representation of each CpG probe on the microarray which is determined as the bead count that a probe has on the microarray. When using the default thresholds as outlined by pfilter ($bc < 3, n > 5\%$) a total of 303 probes were identified according to these thresholds. This suggests that a majority of the probes on the 450K are represented reasonably well when considering a large number of samples. When I considered these thresholds on a per dataset basis (Figure 5.3a) the median number of probes identified in all of the datasets is 557 probes. From Figure 5.3a it can be seen that a few datasets have more than 5,000 probes identified as being poorly represented. The reason for this wide variation in the number of poor samples will require further investigations as the number of probes detected by bead counts do not appear to correlate with correlate with quality control metrics and could be solely due to the random distribution of beads that are present on each microarray.

Next, I compared the number of times a probe fails (using the default thresholds) in an increasing proportion of the datasets (Figure 5.3b). Fewer than 750 probes (649) appear to fail in 20% of datasets

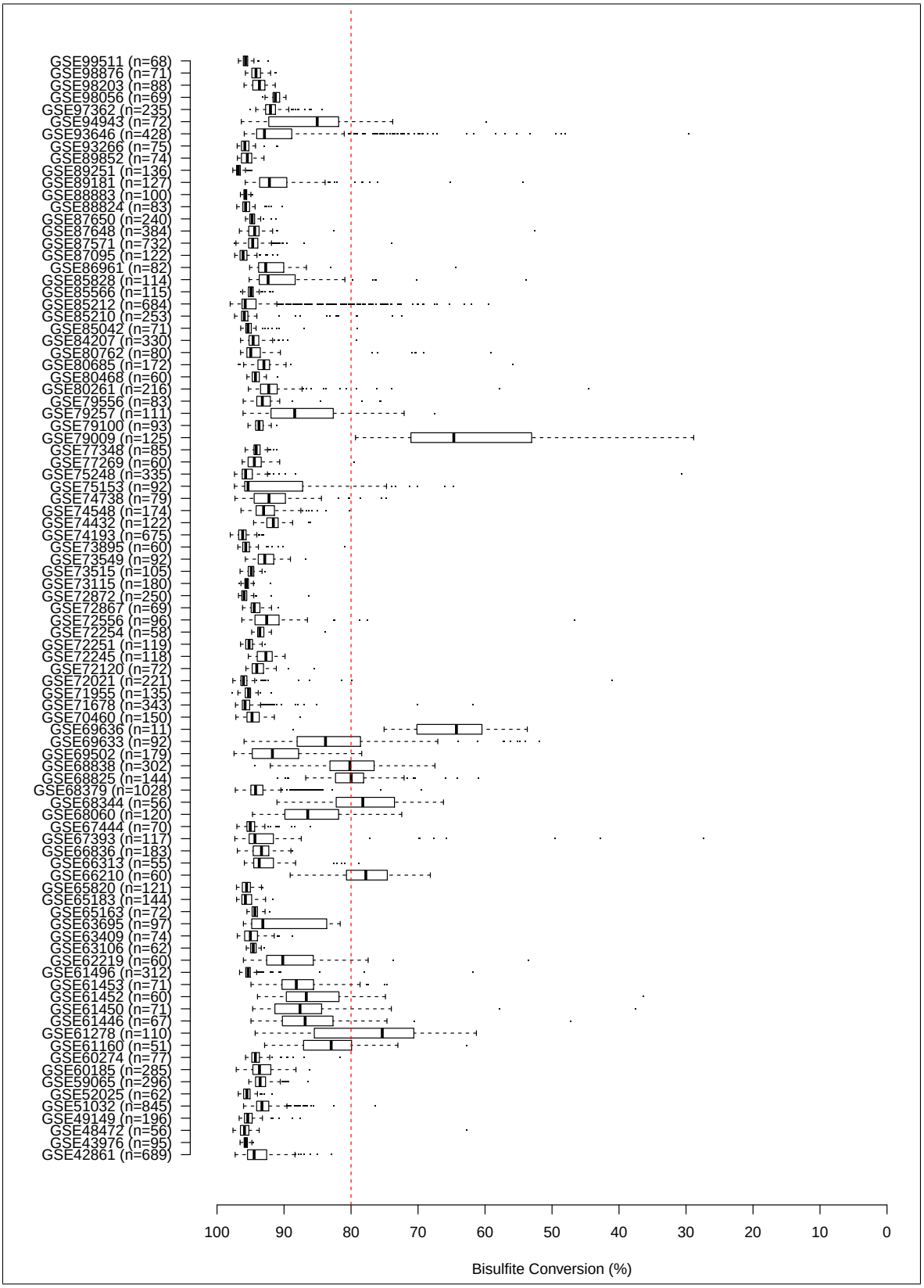
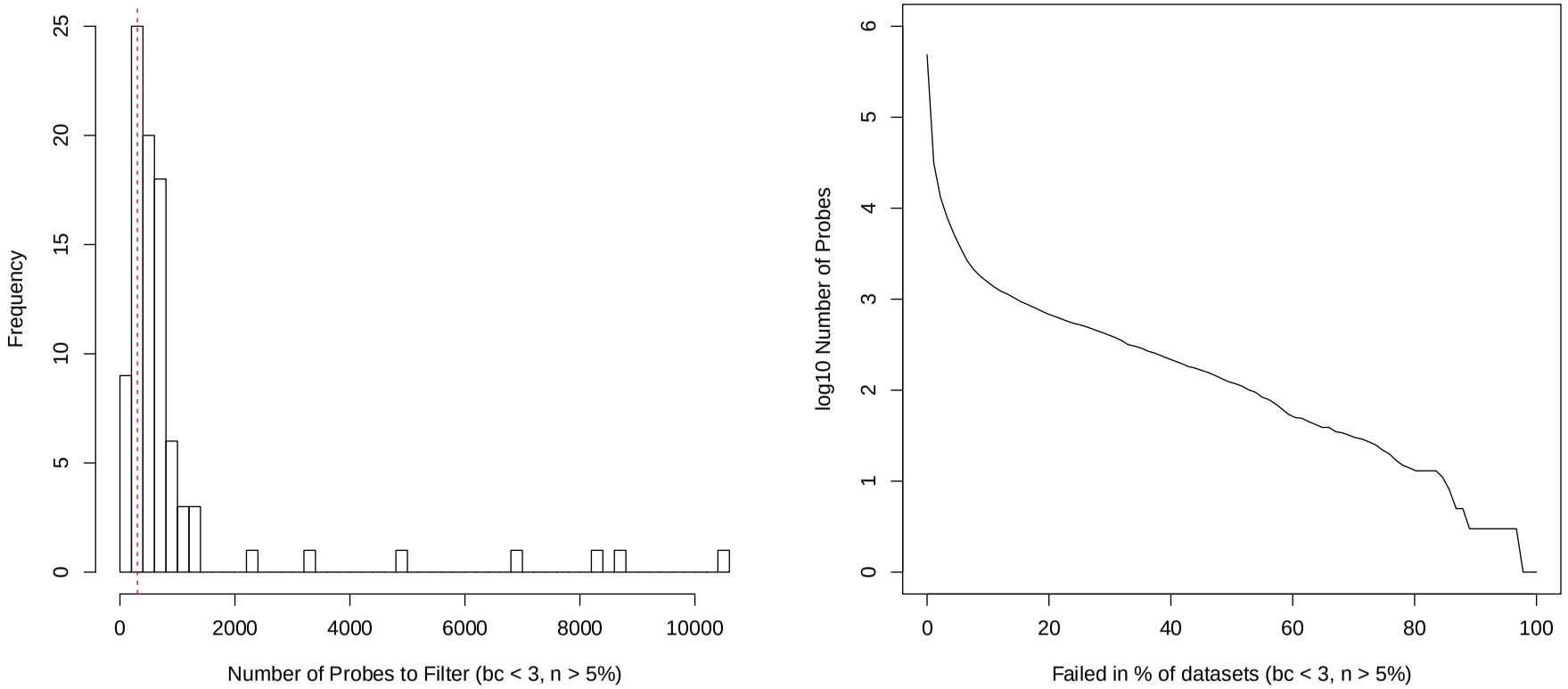


Figure 5.2: Box and whiskers plots of bisulfite conversion values as determined by bscon for 91 datasets obtained from GEO.

which suggests that overall each probe is fairly well represented and that only a small subset of probes could be removed as they are known to fail 20% of the time.

I then explored how variation in the thresholds for this beadcount check affects the number of probes that would be detected in this manner. Figure 5.4 shows box and whisker plots for the number of probes identified in each dataset and the large GEO dataset for a variety of thresholds using cut offs that range from < 2 to < 7 beads in as much as 10% of the samples. A cut-off of n $> 1\%$ appears to be too strict, even when applied to the GEO dataset as a whole as a large number of probes are flagged as a result. If using such a threshold in conjunction with a stricter bead count (e.g. bc < 5) roughly 60,000 probes could be removed from analysis. Relaxed thresholds such as bc < 4 and n $> 10\%$ produce similar numbers of probes to the default parameters and could be a reasonable alternative to the default thresholds as there is only just a significant difference in means ($p = 0.04948$) between these two thresholds.



(a) (b)

Figure 5.3: Characteristics of GEO dataset according to bead counts: **(a)** Histogram of number of probes having a bead count <3 in >5% of samples per dataset. Red dashed line indicates overall number of probes according to the same thresholds globally (not split by dataset). **(b)** Number of times a probe fails in n% of datasets.

5.2.2.2 Detection p-values

The second characteristic I explore to identify problematic probes was using the detection p-values. According to the thresholds defined by pfilter ($p > 0.05, n > 1\%$) a total of 24,141 probes were flagged as having an unreliable signal across all samples. When each dataset was treated separately (Figure 5.5a) the mean number of probes identified was similar (21625). Using an identical approach as used with the beadcounts 14.681 probes failed in 20% of datasets (Figure 5.5b).

Different thresholds in both the detection p-values and the number of samples show similar trends when increasing the threshold. Increasing the detection p-value threshold (while keeping the proportion of samples the same) did not significantly affect the number of probes that were identified as being low quality (Figure 5.6). Increasing the proportion from 1% to 5 or 10% yielded a significant difference in the number of probes detected while an increase from 5% to 10% did not.

Regardless of the thresholds used, it is seen that as many as 10 datasets have more than $>$ 50,000 unreliable signals in most circumstances. Of particular note, one dataset (GSE72556) has as many as 300,000 probes that would be detected by any thresholds.

When combining together the probes that fail both the default parameters of pfilter (beadcounts and detection p-values) on a global scale, 20% of the time, a total of 15,124 probes are found to fail at least one of the two tests (206 probes fail both tests). When considering where each of these probes are located within the genome it can be seen that there is a mostly uniform proportion for each chromosome with the exception of the Y Chromosome which exhibits an 88% failure rate (Figure 5.7). This suggests that any signal generated by probes on the Y chromosome are likely to be confounded by background noise.

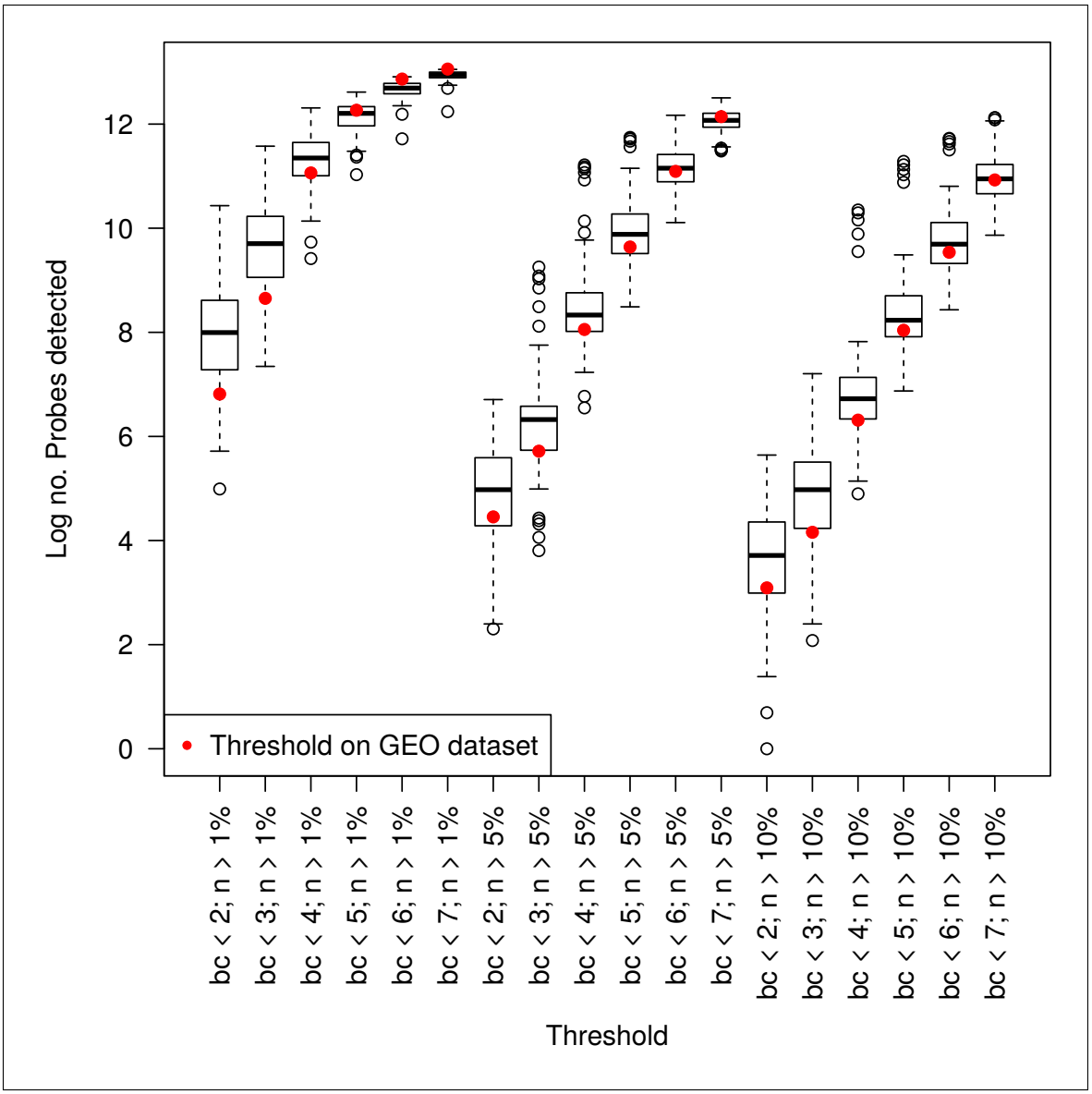


Figure 5.4: Number of probes removed based on a variety of thresholds on beadcounts. Each box and whisker plot corresponds to a set of thresholds applied to each dataset (n=91) in the GEO dataset. Red points correspond to the number of probes identified when thresholds are applied to the full GEO dataset.

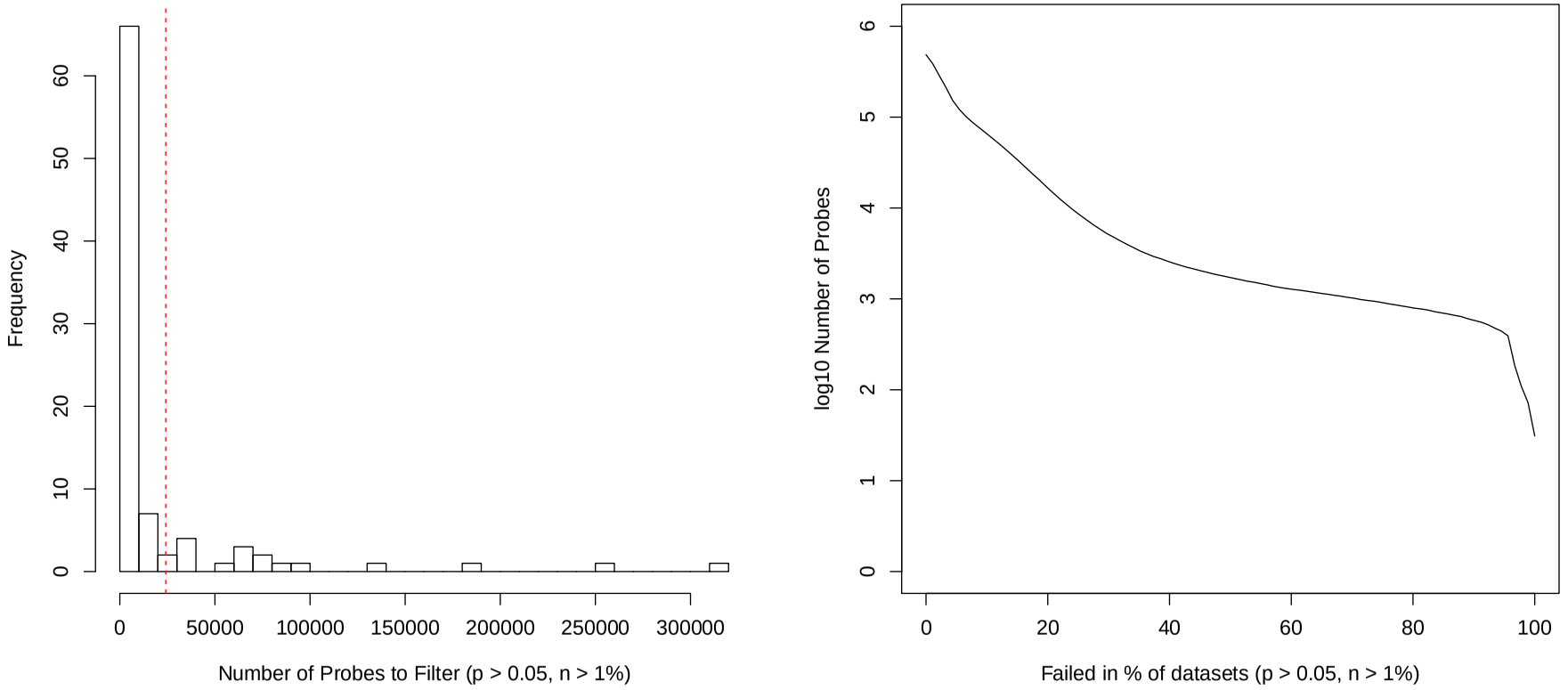


Figure 5.5: Characteristics of GEO dataset according to detection p-values: **(a)** Histogram of number of probes having a detection p-value >0.05 in $>1\%$ of samples per dataset. Red dashed line indicates overall number of probes according to the same thresholds globally (not split by dataset). **(b)** Number of times a probe fails in n% of datasets.

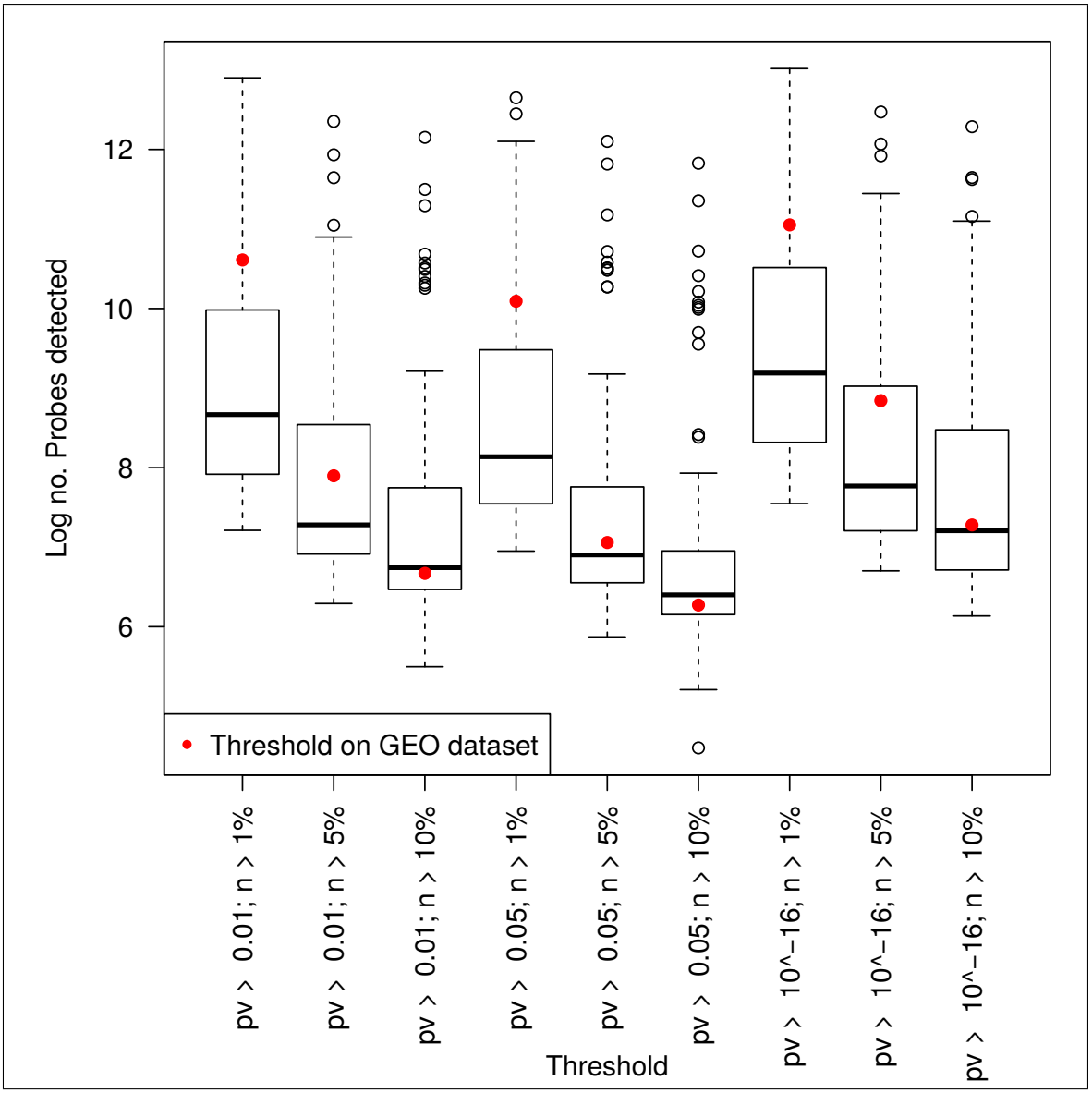


Figure 5.6: Number of probes removed based on a variety of thresholds on detection p-values. Each boxplot corresponds to a set of thresholds applied to each dataset (n=91) in the GEO dataset. Red points correspond to the number of probes identified when thresholds are applied to the full GEO dataset. Detection p-values thresholds were chosen based on two recommended values (0.05 and 0.01) and recommendation by Lehne *et al.* (2015) (10^{-16})

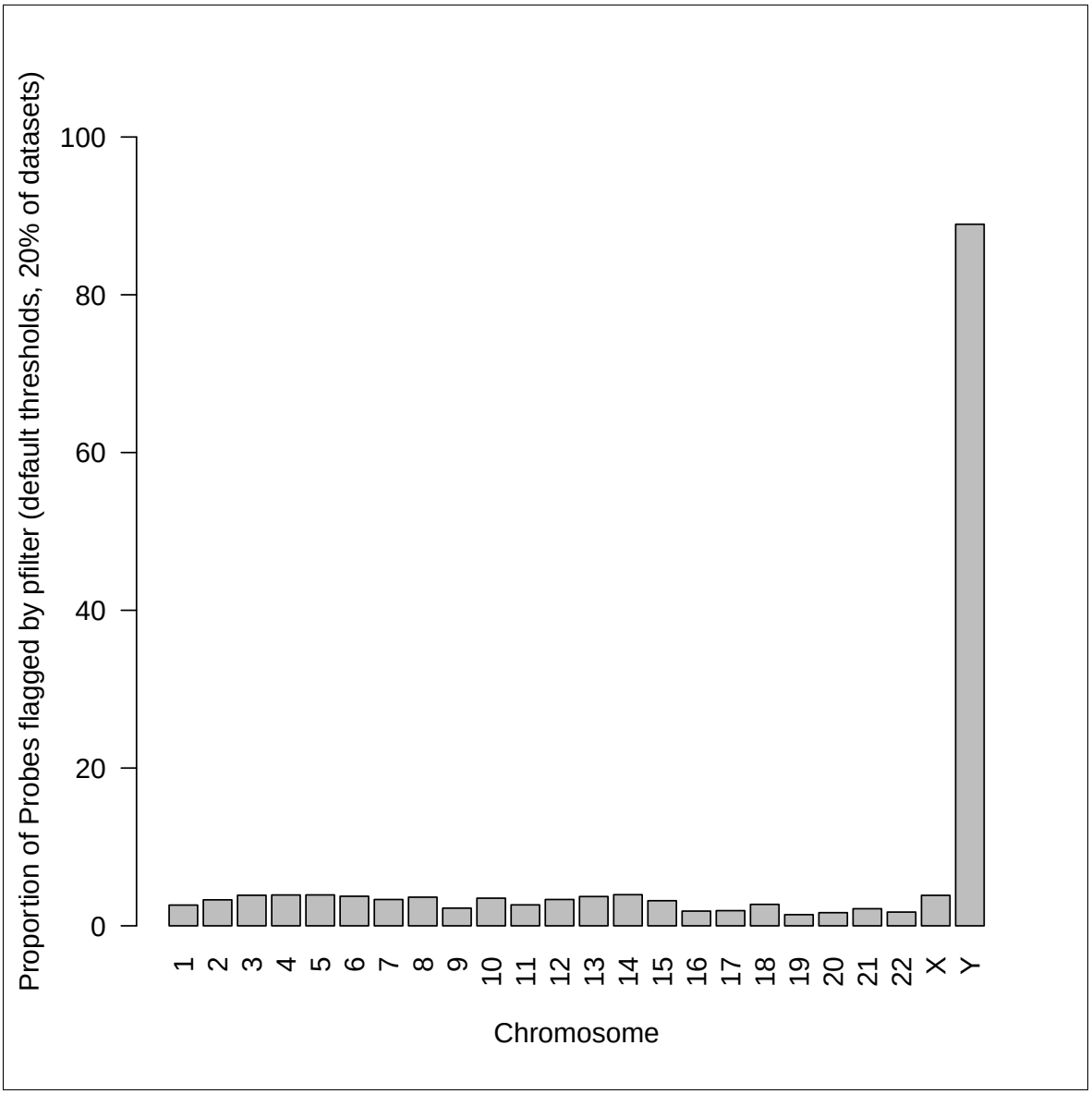


Figure 5.7: Breakdown of probes detected by pfilter separated by Chromosome

5.2.2.3 Sample Variance

A very much unexplored feature of probe filtering is the removal of probes that do not vary at all. While it is expected that a majority of CpGs will vary by some amount, it is uncommon for certain probes to exhibit little to no variation (especially across many samples). Because I have a large number of samples from a variety of tissues it is possible to identify a set of probes which display minuscule amounts of variation even across different tissues and disease. In other words, identifying cases where no variation exists when variation is expected.

There has been no recommendation for what the absolute minimum amount of variance that a probe should exhibit. Therefore, I explored how much each probe can vary. Figure 5.8 shows the distribution of probe standard deviations split across Type I and Type II probes in both the GEO and Marmal-aid datasets. What is clear is that the variation between Type I and Type II probes is different as Type I probes show large peaks at 0.026 and 0.066 for GEO and Marmal-Aid respectively while Type II probes have a more variable distribution with two distinct peaks, one within the range of 0.026-0.066 and the other peak around 0.16.

There are currently no recommendations for what is the absolute minimum amount of variance a probe should exhibit. I decided to explore the bottom 5th percentile of each dataset within the GEO dataset and created an intersection between all of the probes that were identified within the bottom 5th percentile. When considering the probes which consistently exhibit a low variation ($>$ 50% of the time) there are 11,938 (3623 Type I, 8315 Type II) probes that tend to have small variation. Removal of these probes, in particular, will warrant caution as there will be certain circumstances where these probes can exhibit a large variation, thus removing these low variable probes can lead to the possibility of missed associations.

5.2.2.4 Minor Allele Frequencies

Lastly, I examined the presence of minor allele frequencies and private SNPs at the single loci level. The previously described lists have determined numerous CpGs which have known public SNPs that underlie

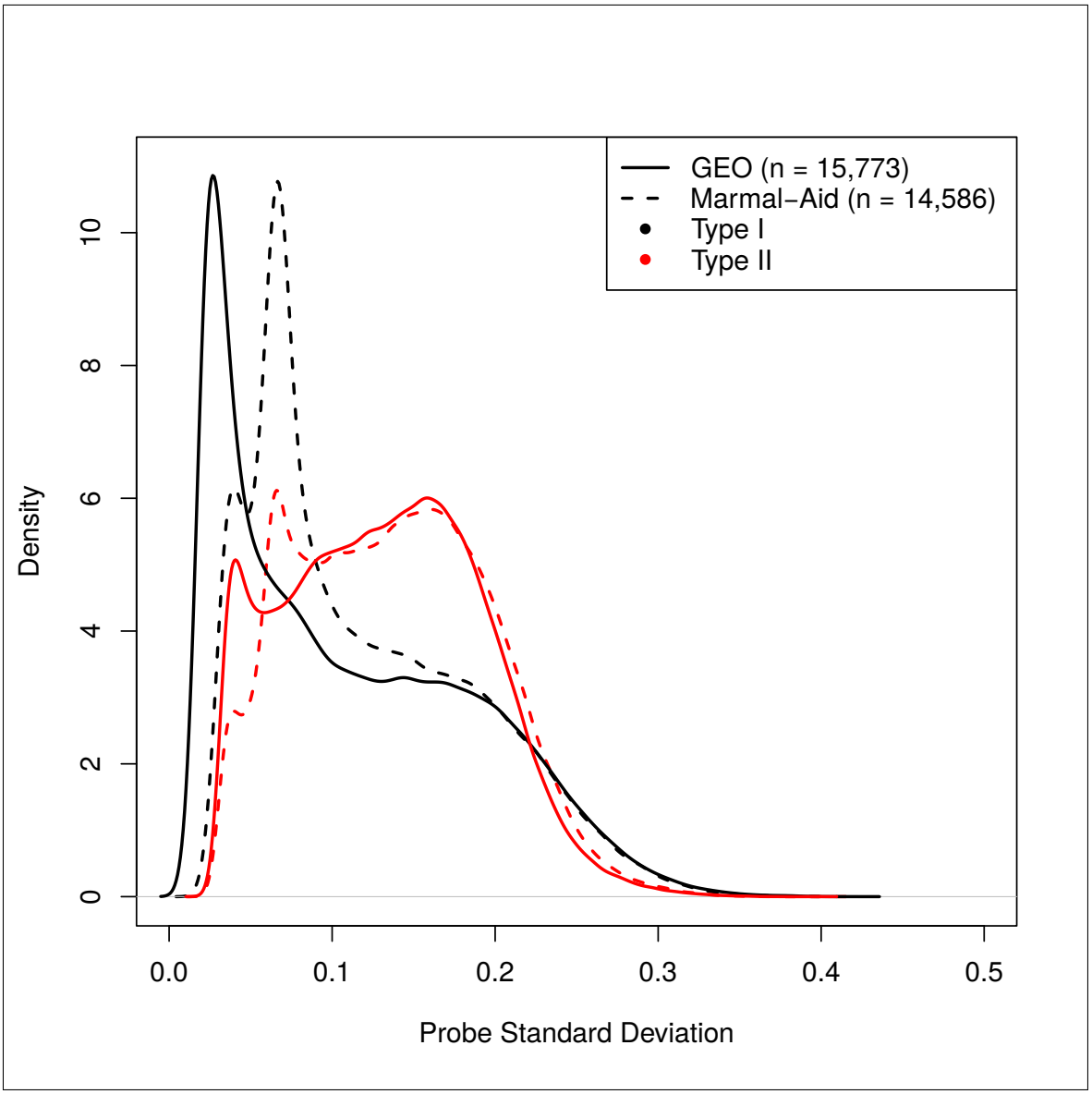


Figure 5.8: Distribution of standard deviations for each probe within the GEO and Marmal-Aid datasets according to probe design.

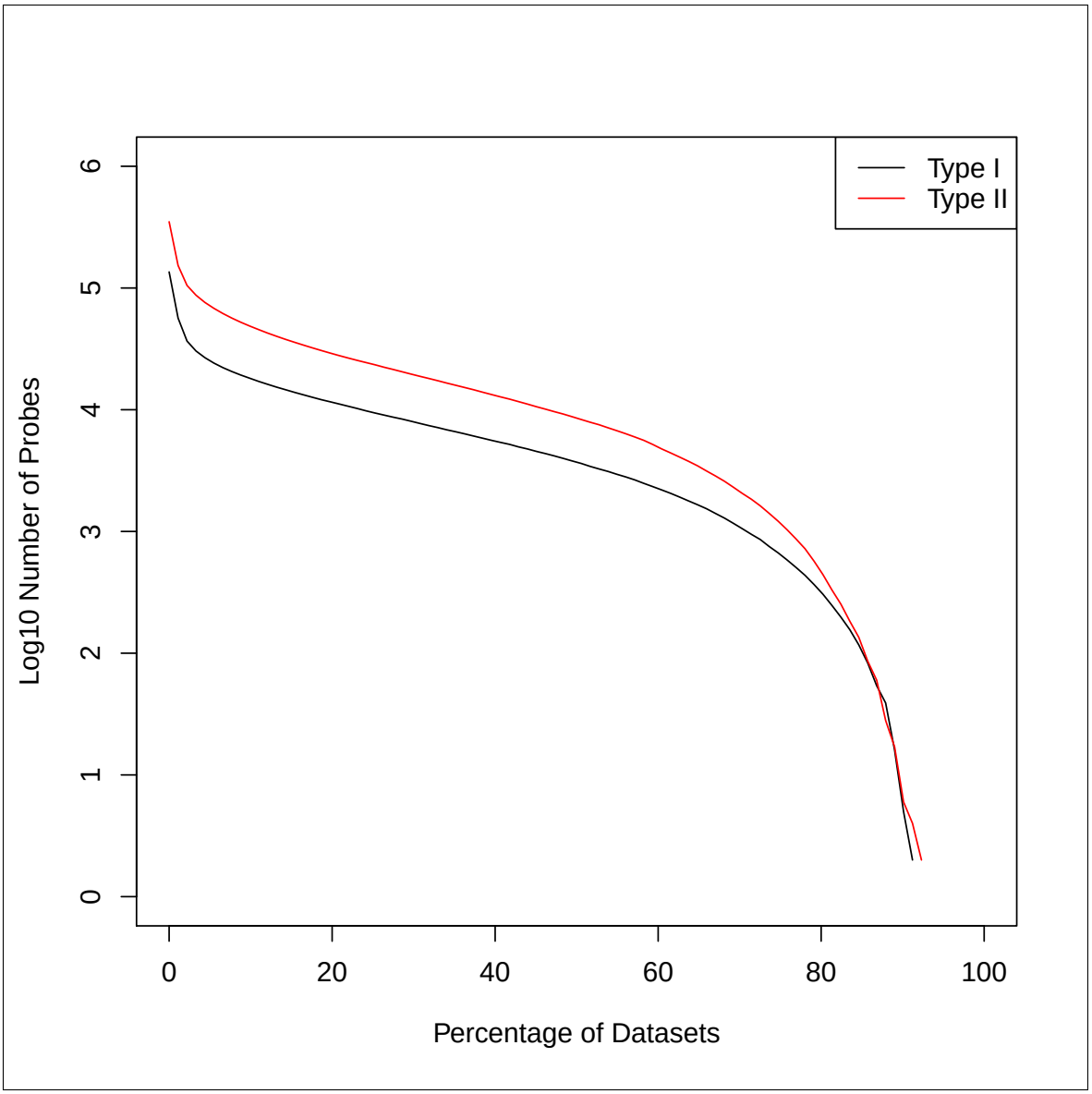


Figure 5.9: Number of probes ranking in bottom 5th percentile for variation in 91 datasets obtained from GEO split by probe design.

the probe sequence. This means that these probes can be confounded by genetic differences and are recommended for removal. However, this does not account for unknown SNPs that could be identified in the remaining set of probes. By using k-means clustering to cluster the raw β values to three separate centres representing low, intermediate and high methylation, it is possible to generate pseudo-allele frequencies for each given locus. After clustering, the allele frequencies were tested if they were in Hardy-Weinberg equilibrium to establish whether or not any of the probes exhibited a SNP-like distribution. This process was applied to both the GEO and Marmal-aid datasets. In total 67892 (GEO) and 49860 (Marmal-aid) probes were in Hardy-Weinberg equilibrium. The intersection of both probe sets revealed a total of 15,004 probes to have SNP-like proportions.

5.2.2.5 Summary

When combined with the preexisting probe lists and other recommendations (such as removing probes from the X and Y chromosomes) a total of 103,128 CpGs can be removed from analysis prior to statistical testing. Table 5.1 shows a breakdown of the number of probes in each category and the properties associated with those features.

5.2.3 Part 1 Discussion

The quality control of DNA methylation microarray data is extremely important in EWAS. On smaller scales, this process is relatively straightforward and many of the tools and methods have been developed with these small datasets in mind. As there is now a large amount of information that is available on public repositories, it is likely that larger analyses will be taking place and therefore it is useful to consider how signals on the 450K microarray will behave during this bigger analyses.

By examining four characteristics of the 450K microarray, I propose an extension to the already popular probe-lists to recommended that up to 103,128 probes that could be removed from analysis as they demonstrate some form of confounding. This is not to say that any analysis that includes these probes

Table 5.1: Breakdown of all 103,128 CpGs identified to be filtered from analysis

Characteristic	CpG Context	Type I	Type II
Bead Count <3, n >5%, >20% datasets	Island	212	89
	Shore	56	70
	Shelf	18	35
	Open Sea	65	104
Detection P >0.05, n >1%, >20% datasets	Island	397	618
	Shore	531	1960
	Shelf	225	2479
	Open Sea	878	7593
< 5th percentile variation in > 50% datasets	Island	3383	5737
	Shore	125	2069
	Shelf	19	60
	Open Sea	96	449
Intersected HW SNPs from GEO and Marmal-Aid	Island	2813	2509
	Shore	651	3455
	Shelf	222	1124
	Open Sea	685	3545
Zhou *et al.* (2017)'s List + Cross Hybridised	Island	8892	8447
	Shore	2711	10589
	Shelf	930	4834
	Open Sea	4544	17218
X, Y and SNP probes	Island	2208	2204
	Shore	481	2497
	Shelf	97	1093
	Open Sea	379	2754
Total Combined	Island	16306	17672
	Shore	4237	19430
	Shelf	1425	8391
	Open Sea	6287	29380
	Total	28255	74873

will be of bad quality but is more of an addition to the existing understanding of how the 450K array performs in preparation for larger analyses. I had set out to describe the characteristics and test a few thresholds to identify a set of probes that perform and behave differently from the bulk of the 450K microarray. I do not recommend that these probes be used instead of quality control but this list can be used in place of quality control if performing such quality control is not possible such as obtaining a preprocessed β matrix from GEO.

When using the default parameters as defined by the pfilter function (beadcount $<$3, n $>$5%; detection p $>$0.05, n $>$1%) to identify probes that fail 20% of the time I identify 15,124 probes to fail one or both of these two thresholds. The probes that fail these thresholds are fairly well distributed across all the chromosomes and divided between Type I and Type II probes evenly. This is with the exception of Type II probes that are located within intergenic regions of the genome. Additionally, this probeset is inclusive of nearly 90% of the CpGs located on the Y Chromosome. The pfilter function was designed more than 6 years ago and the default thresholds were chosen based on past experience with EWAS at the time and was tested on datasets of around 100-200 samples. These thresholds nonetheless appear to perform well on a majority of the datasets that are within GEO with a few exceptions. Alternative threshold such as detection p $>$0.01 or Lehne *et al.* (2015)'s suggestion of 10^{-16} yielded a non-significant difference in the number of probes detected, suggesting that any of these three thresholds are sufficient for use. Most of the probes on the 450K microarray appear to be fairly well represented with only 649 probes consistently having low representations.

I also try to characterise probes by how much they vary across thousands of samples and tissues. What is apparent and expected is that the Type I and Type II probes behave wildly differently. Type I probes are defined by a single peak around a SD of 0.02 while Type II probes have one peak at an SD of 0.03 and a wider peak at an SD of around 0.06. This suggested that using an arbitrary cutoff, e.g. $<$0.02) would be unsuitable as it would unfairly penalise Type I probes as they are known to vary less. Thus I opted to use a different approach. First I split the probes according to design and then ranked the variation of each probe within each dataset (n = 91). I then intersected the bottom 5th percentile of each dataset to identify a set of probes which consistently ($>$50% of the time) had low variation. Using this method, I

identify 11,938 probes which were shown to vary an extremely small amount. However, the treatment of invariable probes is not as clear cut as it is with the other probes on this list. In fact, I would recommend keeping these low variance probes for analysis. As EWAS are growing in size the advantages of having a large number of samples is that there is the a large enough power to identify very small effect sizes - even from probes with a SD of 0.01. While I have included these in the probe-list I would like to note that I do not recommend that these probes be removed when considering a large number of samples as it is possible to identify associations with them. Additionally, it is possible that these probes can vary wildly in certain diseases and tissues that were missed in the 91 datasets I decided to look at.

Lastly, I characterised probes that exhibited a typical SNP-like distribution. Removing probes that have signals that are subjected to genetic confounding is a well known and established technique already. And the probe-lists have already identified public SNPs that are within the probe sequences of the 450K microarray. I set out to identify an additional set of unknown SNPs (or low-frequency alleles) which were otherwise missed. The intersection of the detected probes between GEO and Marmal-aid dataset appeared to be well distributed between both CpG islands and probe designs. Out of the 51,000 SNPs that are listed in Zhou *et al.* (2017)'s list only 1607 of these were identified in my intersection of SNP-like probes.

The detection p-values are a measure of confidence that a signal is presented above the background signal. However, there are numerous ways to calculate this. The most popular way of computing the detection p-values will generate a different set of detection p-values according to the official method (Illumina's GenomeStudio). As a result, the software that was used to read in the raw data will affect the detection p-values for each probe. Although this is a small matter, the detection p-values used in this analysis were derived from the detection p algorithm described in methylumi. If I were to go back and read all of the idat files using minfi I would have received a completely different set of results. Thus the application of pfilter (with the detection p thresholds) could have identified a completely different probe-set to the one presented here.

This analysis does not answer whether or not that removing the probes identified is actually useful for

analysis. In the future, I would like to implement the extension I provide in this chapter to a number of EWAS. It is entirely possible to combine the results of this study with the multiple EWAS that were performed in Chapter 2. This would allow me to explore whether or not the application of probe-filtering does lead to improved results.

In this part of this chapter, I extend the widely used probe lists that researchers use extensively in EWAS with a new set of features that are more reflective of the overall quality of the probes in question. The intention behind this is to provide alternative methods of quality control where there is no opportunity to perform the routine quality control steps as defined by individual software packages.

Considering the 450K is approaching the end of its lifespan where the last few remaining datasets are being produced and submitted to online repositories the wealth of data provided by these 450K arrays will enable us to generate more assumptions about these DNA methylation microarrays behave. As the use of meta-analyses and large scale data analyses taken precedence the need for generalised, reproducible quality control will become increasingly more needed. This work describes the first steps towards achieving a thoughtful starting point towards the quality control of large numbers of data.

5.3 Part 2

Recent efforts in the field of epigenetics have been focused on understanding the role of DNA methylation and disease. The concept that DNA methylation is involved in transcriptional regulation has been well-observed (Jones, 2012) and has been shown to play important roles in gene silencing mechanisms such as X-inactivation and genomic imprinting (Jaenisch & Bird, 2003). Despite this, there is little evidence to suggest whether or not stochastic DNA methylation has a causal role in gene silencing or is left as a result of gene silencing. Early studies had identified correlations between promoter hypomethylation with gene expression and conversely, gene body hypermethylation has been correlated with gene expression as well (Jones, 2012).

A recent study by Ford *et al.* (2017) set out to ascertain whether induction of DNA methylation at promoter regions is accompanied with transcriptional repression. As such Ford *et al.* (2017) sought to assess changes of DNA methylation alongside changes in gene expression and chromatin state by inducing the methylation state in numerous promoters. To do this Ford *et al.* (2017) used engineered cell-lines derived from MCF-7 cell lines which upon doxycycline (dox) treatment, expresses Zinc Finger (ZF) Domain proteins fused with DMNT3A proteins (ZF-D3A) that induced the methylation of thousands of promoter regions to which these ZF domains can bind to. Originally designed to target an 18 bp GC-rich sequence within the SOX2 promoter, these ZF fusion proteins were also found to non-specifically bind to other GC-rich sequences such as those found in CpG islands. Ford *et al.* (2017) measured the methylation state, mRNA counts and chromatin state of the MCF-7 cell lines at three conditions: MCF-7 control, ZF-D3A +dox and ZF-D3A dox-withdrawn (where dox treatment was used but then withdrawn after DNA methylation was induced). What Ford *et al.* (2017) had found was there was not enough evidence to support the idea that induced DNA methylation alone would not be sufficient as a gene silencing mechanisms. They had also found that active chromatin (H3K4me3) could exist simultaneously alongside DNA methylation and that following the removal of dox the induced methylation patterns were quickly removed.

However, the data from Ford *et al.* (2017)s study was recently reanalysed by Korthauer & Irizarry (2018) who had found contradictory results with the original conclusions of Ford *et al.* (2017)s study. Instead, Korthauer & Irizarry (2018) found that the induction of methylation at a variety of promoter regions did in fact have a gene-silencing effect. These contradictory results were found due to changes in how the data were analysed. Of note, Korthauer & Irizarry (2018) used statistical inference to identify induced DMRs instead of arbitrary cut-offs. Korthauer & Irizarry (2018) tested the relationship between DNA methylation with mRNA and H3K4me3 using fold-change values instead of the absolute count data. This allowed for small differences between regions of low count to have the potential to be significant rather than only considering large differences between regions that have high counts. These differences in analysis were enough to establish that as much as 80% of forcibly methylated promoter CpGs were accompanied with a decrease in gene expression and a decrease in H3K4me3.

Despite the contrary results of the initial analysis, citeFord2017s study caught my attention as it demon-

strated that role DNA methylation plays as an epigenetic mechanism still has room for investigation. As I had previously obtained the DNA methylation patterns of more than 15,773 samples from a variety of tissues, it should be possible to explore if gene-region specific DNA methylation patterns correlate with gene expression.

5.3.1 Part 2 Methods

5.3.1.1 General Trends gene region Methylation

The 450K microarray interrogates the methylation state of numerous CpGs that annotated to roughly 20,000 genes. These probes can be categorised based on the area of where they are located within the gene. These are the TSS1500 (1,500 bp away from transcription start site), TSS200, 5' UTR, 1st Exon, Gene Body and 3' UTR. Specific gene region methylation was estimated for each gene as the mean β value of all CpGs that were annotated to a given gene region. CpGs that were annotated to multiple gene regions or genes were also included in the estimate of all genes or regions such probes annotated to. Neighbouring CpGs that did not correspond to any gene regions were not included in these calculations. The gene region methylation were obtained from 6464 samples across 17 tissues from the GEO dataset that is described in Section 5.1.1.1.

5.3.1.2 Tissue specific Gene Expression

Relevant gene expression values for the 17 tissues for 11,334 genes were obtained from the InterMine web resource (Smith *et al.*, 2012; Kalderimis *et al.*, 2014). These gene expression values are stored as a moderated T-statistic which corresponded to the ratio of the $\log_2$Fold change to its standard error. Where a negative refers to the down-regulation of a gene in a specific tissue and a positive value corresponds to up-regulation.

In summary: moderated gene expression values and the average DNA methylation state for TSS1500, TSS200, 5'UTR, 1st Exon, Gene Body and 3' UTR were obtained from 6464 450K microarrays for a total

of 11,334 genes and 17 different tissues. The relationship between gene expression and DNA methylation was explored using Pearson correlation both for each tissue separately and all tissues combined.

5.3.2 Part 2 Results

The Pearson correlations between moderated gene expression T-statistics obtained from InterMine and the average DNA methylation across 6 different gene regions are described in Table 5.2. Overall there appears to be little to no correlation with these gene expression values with most of the tissues or with all of the tissues combined.

Figure 5.10 shows the overall distribution of the DNA methylation patterns per gene region across all tissues. Low methylation in promoters and intermediate to high methylation within the gene body and 3' UTRs suggest the methylation of these regions are consistent with previous observations.

Approximately two-thirds of the correlations are negative which provides support for the previously seen negative correlations with gene expression. And all correlations when combining all 17 tissues together yielded small negative correlations and thus a decrease of gene expression is observed with increasing DNA methylation across all gene regions although the correlation is small.

5.3.3 Part 2 Discussion

The results presented in this analysis are extremely preliminary and are intended to demonstrate the types of analysis that could be performed when combining large amounts of data. Here I take thousands of DNA methylation patterns that had been deposited to GEO and combined them with tissue-specific gene expression data taken from the InterMine web resource. I report that there is a small negative correlation with increasing DNA methylation across all gene regions with gene expression in more than 10,000 genes across 17 different tissues, which is consistent with previous observations (Jones, 2012). Although each tissue separately can demonstrate different directions of correlations this can be explored in the future.

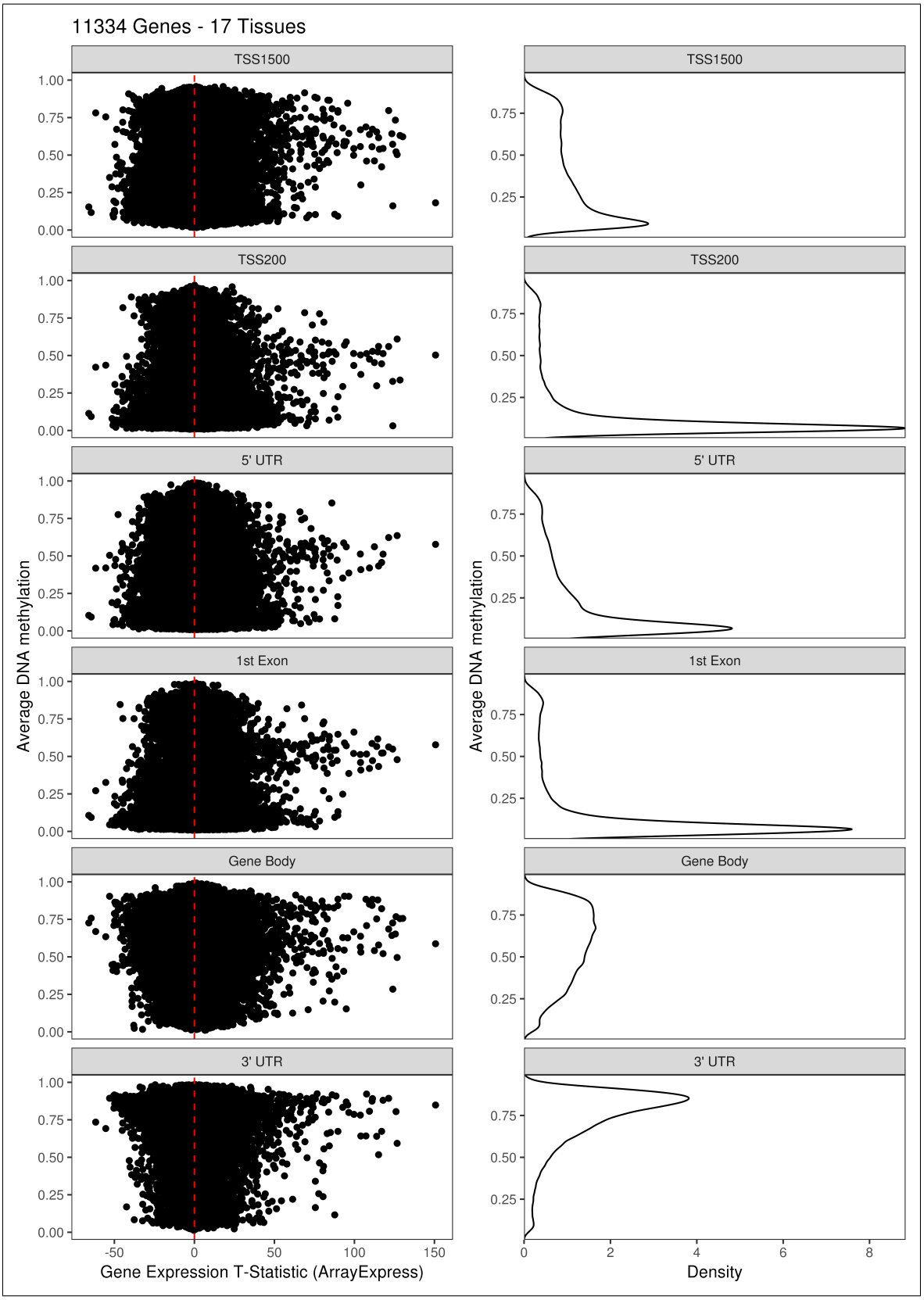


Figure 5.10: Average DNA methylation per genomic region vs Tissue Specific Gene Expression obtained from intermine for 17 different tissues.

Table 5.2: Pearsons Correlation between moderated T-statistics for Gene Expression and average DNA methylation of Genomic Regions by Tissue

Tissue	TSS1500	TSS200	5'UTR	1st Exon	Gene Body	3' UTR
Adipose	0.0838*	0.0938*	0.0983*	0.097*	0.0466*	-0.0147
Bone Marrow	-0.0555*	-0.0795*	-0.106*	-0.0783*	-0.0645*	-0.0176
Bulk Brain	-0.074*	-0.0935*	-0.0748*	-0.095*	-0.0205	-0.00139
Cervix	-0.149*	-0.181*	-0.204*	-0.182*	-0.157*	0.0217
Colon	-0.117*	-0.166*	-0.19*	-0.163*	-0.113*	0.0423*
Cord Blood	0.0625*	0.0624*	0.0226*	0.0675	-0.047*	-0.0285
Dorsolateral Prefrontal Cortex	0.139*	0.181*	0.176*	0.186*	0.127*	-0.0746*
Esophagus	-0.0402*	-0.0887*	-0.0748*	-0.0983*	-0.0571*	-0.0111
Kidney	0.0203	-0.00489	-0.023	0.0191	-0.0479*	-0.0373
Liver	0.0537*	0.0211	0.0162*	0.0415	-3.61e-05	-0.00727
Lung	-0.200*	-0.287*	-0.283*	-0.287*	-0.143*	0.0903*
Lymph node	0.104*	0.127*	0.121*	0.129*	0.0649*	-0.0323
Placenta	-0.138*	-0.174*	-0.170*	-0.170*	-0.0531*	0.120*
Prostate	-0.0791*	-0.116*	-0.134*	-0.104*	-0.0908*	-0.0367*
Thyroid	-0.0568*	-0.0811*	-0.0823*	-0.0766*	-0.0605*	-0.0241
Tongue	0.0869*	0.0967*	0.0853*	0.0905*	0.0325	0.00532
Whole Blood	-0.0121	-0.0511*	-0.0701*	-0.0478*	-0.0609	4.56e-05
All Tissues	-0.0299*	-0.0565*	-0.0654*	-0.0518*	-0.0425*	-0.000451

* $p < 0.0001$

The lack of correlations between DNA methylation and gene expression does not mean that DNA methylation does not play a role in gene expression as there have been well studied examples of this. Rather these results demonstrate that the data that is available from 450K microarrays are not just limited to epigenome-wide association studies and can be used in alternative analyses.

Many of the correlations between gene region methylation and gene expression were significant, negative and small in size. These correlations were seen across various tissues and in general across all tissues combined. However, nearly all of the correlations between 3' UTR methylation and gene expression were not significant with the exception of four tissues. This can suggest that 3' UTR methylation may not be indicative of transcriptional activity. More interestingly, gene body methylation was found to be negatively correlated with gene expression across numerous tissues which is in disagreement with previous studies (Yang *et al.*, 2014).

The largest (negative) correlations that are observed are all from lung tissue. This could be due to the fact that approximately half of the lung samples used in this analysis were sourced from cancerous tissues. Therefore it is possible that the relatively large correlations observed in lung tissues may be caused by global demethylation events which are a known characteristic of cancerous tissues (Feinberg, 2007). This

would need to be examined as it is also possible that batch effects that were not accounted for could be confounding these observations. As no normalisation was applied to the data it is possible that there could be some unmediated effects that are also influencing these results.

What is also interesting is that some tissues presented a positive correlation with gene expression as DNA methylation increased. Tissues such as adipose have positive correlations between DNA methylation and gene expression across the majority of gene-regions. It is not known why these tissues specifically are positively correlated with their respective gene expression. This would be an interesting avenue for future work, firstly to verify that this positive correlation exists and to understand why and how this can come about.

Despite these results not being entirely translatable to Ford *et al.* (2017)'s study - the analyses are somewhat related. Initially, Ford *et al.* (2017) provided evidence that suggested that induced DNA methylation of promoter regions was not accompanied by a gene silencing response. These results were reanalysed and shown that induced DNA methylation did in fact lead to gene silencing. Here I present on a genome-wide scale using thousands of publicly available data that there are little (or no) correlations between DNA methylation and gene expression. More importantly, I have demonstrated how it is possible to efficiently handle and manipulate large numbers of samples to generate such conclusions as a proof of concept for future work. And as a result - highly encourage others to reproduce these results.

Although this study was brief there are numerous limitations I would like to address:

The gene expression values that were obtained from interMine were based on moderated test statistics which were based on comparisons of given tissues to a specific set of other tissues. Therefore the gene-expression values that were obtained for a given tissue (e.g. Lung) may not be representative of a true comparison between it and another tissue. Furthermore, the gene expression values used in this study are not indicative of any gene silencing mechanisms and are more or less related to the up or down-regulation of a given gene in a tissue. This analysis could be drastically improved if data from other -omics-based

experiments were included such as chromatin state, similar to the approach that Ford *et al.* (2017) used in their study.

For this analysis, I had only considered the CpGs that were annotated to genes according to the manifest that was provided by Illumina. As a direct result, there is plenty of room to extend these analyses to include enhancer regions and gene-flanking CpGs or even unannotated CpG Islands. There is a strong likelihood that the method used to estimate the average gene region methylation could be skewed by the overall number of CpGs that are present in each region. In addition, it is possible that simply the 'average methylation score' is not the best metric to consider gene-region wide methylation as the methylation pattern can be highly variable across the length of a given region (such as a gene body). Therefore alternative methods of computing the gene-region methylation should be considered. Although in this preliminary analysis it is unlikely this would have had much of an impact on the results but in the future I think the CpG density per gene region should be factored into the analysis.

I also use the gene-region annotation for each CpG to determine the functional role of a CpG. However the location of a CpG within a gene may not be a good indication of a given CpGs functional capacity. It is possible that CpG island context is a more informative medium to identify whether or not DNA methylation of a CpG could play a role in gene regulation as it has been demonstrated that CpG island methylation is important in gene regulation (Irizarry *et al.*, 2009). Therefore a natural extension for this analysis would be to perform this analysis using CpG island context instead of gene-region annotation. It is possible that categorising CpGs according to their location within a CpG island may correlate better with gene expression. Furthermore, Ford *et al.* (2017)'s study had essentially selected CpGs that were located within CpG islands. As Ford *et al.* (2017) induced methylation of CpGs using ZF-D3A hybrid proteins which bind with high specificity to CpG dense sections of DNA. It is likely that only CpGs located within or near CpG islands were used for their analysis. My analysis looked at all CpGs that annotated to genes on the 450K which included CpGs that may not be nearby CpG islands.

I use data from a large number of sources. This lead to the data being very noisy and heterogeneous. In addition to this, I decided early on in the analysis to leave the data unnormalised as a majority of

the normalisation methods are known to drastically attenuate global differences in methylation such as those observed between different tissues. This lack of normalisation is potentially an issue as there is likely going to be extremely large variation at almost every loci. As I was looking at DNA methylation patterns in a tissue specific manner it is possible that I may be able to each tissue within the GEO dataset separately before carrying this analysis out. Additionally, the gene-region methylation calculation combined the signals from both Type I and Type II probes in certain cases. These probe designs differ considerably which is why methods that adjust the probe distributions such that they are comparable with each other are often used (e.g. dasen or BMIQ), as a result the gene region methylation signals used could have been quite inaccurate. In the future I would like to explore this analysis by implementing some form of normalisation (either across all samples or in a tissue-wise manner) which will also handle these differences in probe designs as it could vastly improve the results present in this analysis.

Some of the tissues within the data-set are subject to numerous diseases and cancer which could be influencing the DNA methylation patterns in some form. In this study, I opt to treat all non-cancerous tissues as healthy but this doesn't rule out other diseases which can affect DNA methylation patterns. Specifically, some of these tissues had a large proportion of samples obtained from cancerous sources. These were: brain (83%), Lung (50%), Prostate (95%), Tongue (100%) and Thyroid (50%). I expect this will need further investigation but it is likely that these tissues in particular could have been subjected to a small amount of confounding as the gene expression data used for these tissues would likely have been sourced from healthy tissues. As a result the DNA methylation patterns and gene expression data for these mostly cancerous tissues may not be entirely comparable.

This analysis was only performed with a small selection of the data on GEO that had raw idat files associated with it. As there are still an excess of 50,000 samples that are still available on GEO (that are in a variety of processed states) it could be possible to go back through GEO and collect data from more tissues and more samples. There will however be issues with missing probes and differing preprocessing methodologies but I do not expect this to have much of an issue as you will be reaching sample numbers that would have the capacity to produce robust results.

5.4 Conclusion

In this chapter I demonstrate two preliminary analyses that make use of more than 15,000 samples. Although the analyses performed here are quite limited it is important to consider that the scale of these analyses is larger than most studies to date. I show that the bigmelon software can easily handle tens of thousands of samples without imposing too many technical limitations such as high memory. These analyses were performed using data that fulfilled a brief criteria (having idat files, $n > 50$), it is likely that using a more strict criteria or focusing analysis on a specific tissue would likely produce an easier analysis with fewer potential confounding.

Firstly, using the 15,773 samples I investigated the characteristics of the probes that are present on the 450K microarray. Considerable effort has been made in the past to identify problematic probes which could produce unreliable signals. However, these efforts did not consider the sample-wide quality control metrics that I explored here. By looking at these metrics I extend these probe-lists to 103,128 probes which fail a variety of criteria. I believe that these additional probes could be useful for analysis where the raw idats are unavailable for quality control.

The second piece of analysis I used the large dataset I created and combined it with data that has been collected from another public resource (interMine). Using both these data, I investigated the correlation between DNA methylation and gene expression. From this preliminary study it may appear that there is no correlation between DNA methylation and gene expression. However it is important to remember that the DNA methylation has been shown to regulate gene expression in various circumstance. Although I perform a genome-wide approach for this analysis it is likely that there are still facets that are missing and limitations that perhaps could dealt with using a better designed experiment.

The publicly available data that is on GEO is a useful resource as it contains more than 70,000 450ks and a growing number of EPIC array data. I have created a tool (geo2gds) which downloads and parses DNA methylation data (only GEO accessions with raw idat files) from GEO into a gds format file. This can then be combined with other datasets which makes data collation easy. It is possible that this tool

can be converted to work on accessions that only contain processed βs but disparities between processing methodologies and probe-filtering can limit the analysis of such data.

The geo2gds function also attempts to download the phenotypic annotations which accompany the GEO accessions but due to inconsistencies between annotations, the combination of multiple datasets using this method requires manual verification and sanitisation of annotations to produce something meaningful. The MIAME guidelines are suggested for GEO submissions but not all datasets submitted to GEO contain the necessary information to reproduce the analysis that had been performed. In the future, I believe the submission guidelines for DNA methylation data (and other omics data) should be updated to include the sex, age and tissue sources of samples, in addition to the raw files being made available. This would allow for data to be readily combined and make large scale analyses easier to perform. I foresee that such changes would make constructing large datasets from multiple sources quite straightforward. However, I acknowledge that certain data protection and privacy issues will not allow for so much information to be made publicly available.

Previous large scale analyses such as those presented by Karlsson Linnér *et al.* (2017) and Horvath (2013) are dwarfed in comparison to these analyses and any technical limitations that they had faced could be remedied through the use of the infrastructure I have provided (Gorrie-Stone *et al.*, 2018, Chapter 3). It is my hope that others will go on to use the tools that I have created to produce large and interesting analysis that push the limits of the software and uncover biologically important discoveries.

Chapter 6

General Discussion

The DNA methylation microarrays have been invaluable for EWAS as a whole. As a single platform, they have proven themselves to be an incredibly cost-effective platform to assay thousands of samples. As the number of samples submitted to GEO alone approaches 100,000, there is truly a fantastic wealth of data that is now available to analyse. There is no doubt that this data will become an invaluable resource for future work.

As these microarray platforms have recently been extended, it is expected that a similar number of samples will be assayed on the EPIC array in the future. Both sets of microarray provide thousands of gigabytes of information all from a rich diversity of disease and tissue types. Although there has yet to be a standard for this type of large scale analysis I am confident that practical guidelines and well thought out methods will be established in the future. Despite these microarrays being selective in the field of epigenetics they investigate, I imagine that they can and will be combined with other omics data to comprehensively investigate a variety of interesting and impactful biological questions.

It is apparent that datasets are becoming larger and the physical size of computing resources could become a potential limitation for both EWAS and multi-omics studies going forward. As part of this thesis, I developed the bigmelon R package which reduces the memory requirements of the R programming language and EWAS to more manageable levels. I extensively tested this software on datasets comprised

of more than 15,000 samples and was able to perform analyses that would otherwise require hundreds of gigabytes of memory to perform in a timely manner. Despite the software being specifically tailored to microarray data, I imagine that repurposing or developing a new platform that is specifically designed to combine microarray data and sequencing data from multiple omics experiments is possible and could enable complex bioinformatical analysis without the need of large compute resources. I demonstrate that data stored on the hard-disk using optimised accessory routines is a viable alternative to using cloud-computing or computing stacks specifically built for big data. Software such as bigmelon was designed so that it integrates relatively effortlessly with the other popular software used for the analysis of DNA methylation data within R. For this reason, I believe it is possible to apply this approach of analysis to other branches of -omics data.

As the cost of whole-genome sequencing is becoming cheaper with every passing year it is not difficult to imagine that sequencing-based techniques are likely going to replace the use of microarrays when performing genome-wide analysis. It is almost guaranteed that the size of data would balloon and therefore make EWAS inaccessible. Therefore it is highly important that there are equal amounts of focus and attention placed into the development of efficient and accessible software to facilitate truly genome-wide analyses.

It needs to be considered whether or not the bisulfite treatment and subsequent assaying of DNA is an appropriate measurement of CpG methylation. As bisulfite treatment is insensitive to alternative forms for cytosine modifications, such as hydroxymethylation, it may be unsuitable for the precise estimation of DNA methylation patterns. While it has been generally accepted that the contribution of other DNA modification does not have a large impact on the robustness of findings but it can make the interpretation or identification of underlying mechanisms of disease difficult to discern. Adjustments to methodologies such as oxidative bisulfite treatment can be performed in conjunction with normal bisulfite treatment to identify a more accurate estimation of DNA methylation and other modifications but these considerations have not been widely adopted. Herein lies another problem, because splitting the signal between 5mC and 5hmC essentially requires two experiments both the cost and the size of the resultant data are essentially doubled which could potentially limit the number of samples that are being analysed. Therefore the need

for highly optimised methods such as those provided in bigmelon could be integral moving forward.

During the course of this PhD I examined over 100 different DNA methylation microarray datasets that were obtained from a variety of sources and quality. Given that I have spent a considerable amount of time looking at this data and how to analyse it I have some insights that I would like to discuss.

Data, when submitted to public repositories, should be submitted with the raw idat files. This may seem obvious but nearly three quarters of the datasets that were deposited to GEO did not contain the raw files and contained some form of processed β or intensities where assessing the quality is difficult. Depositing the raw data will prove favourable for many analyses and will not limit submitted data to reproduction analyses. In addition to the deposition of raw idat files, I also recommend that data should be supplied with information that goes beyond the key model that is performed in studies supported by the submitted data. Frequently, data is supplied with very little information detailing key variables including the age and sex of the participant and even which tissue the sample was sourced from. These basic annotations are extremely useful for generalised analyses such as the one described in Chapter 5 but can become difficult to handle if data is unconventionally annotated. I understand that ethical considerations and data privacy issues will inevitably make providing the raw data and even some key variables such as Age, Sex and Tissue Type difficult. From my understanding, it is the MIAME guidelines which are most commonly adopted (if any are adopted at all) when data is submitted. I think that extending these guidelines purposefully towards 'reasonable information' or even 'extensive information' (e.g. including sex, age, tissue type and other information to satisfactorily reproduce an analysis) would allow for better opportunities for both reproduction and innovative analyses.

I have explored the majority of the available quality control methods that are available for EWAS. From my analyses I ultimately recommend a comprehensive approach to quality control is likely to be the more favourable for any EWAS. I stress, much like those before me, that the need for reproducible research is paramount to the success of a study and that provide clear and concise details about what steps were involved during the analysis of data is highly valued in any EWAS. By comprehensive quality control I refer to quality control that includes both control-probe and data-driven based methods. For control-

probe based methods majority of the tools are quite useful, although the thresholds may need adjusting. The interactive GUIs that are provided by both MethylAid and shinyMethyl are very informative and fit nicely into the minfi workflow. The data-driven methods specifically refer to the tools (outlyx, qual, pwod) described in Chapter 2 all of which can be used as part of any work-flow as they only require a β matrix to run and are otherwise insensitive to the useful data structures in which data are stored. Quality control is a subjective task and will depend on the data that is being analysed however providing a short description of how the data was quality controlled will be valuable in being able to verify and reproduce research independently.

There is also an excellent variety of normalisation methods that can be applied to DNA methylation data. Generally there is no consensus on which method is the "best" for both 450K and EPIC microarray data. This can be disorientating for users who are unfamiliar with the slight differences between each method and may result in an inappropriate method being applied. I would recommend using some variety of quantile normalisation as it is consistent and simple to understand. The dasen method was seen to be highly effective for 450K microarrays as it accounts for the difference in probe designs and slide positioning. However it has yet to be determined whether or not the slide positioning effects that are prominent on the 450K array are in fact present on the EPIC array at all. Alternative preprocessing methods such as background correction and adjusting for dye bias should be considered if one suspects that these issues could be problematic for downstream analysis.

I explore the idea of expanding the probe-lists that are commonly used in EWAS by examining certain characteristics of the 450K which are often used for low-level data processing. Current probe filtering focuses on the removal of probes that could have unreliable signals by virtue of how the probe is designed. This has quickly become a widely accepted form of quality control as it provides a consistent list of features that are readily accessible. This has been considered as a thoughtful and conservative way to reduce false positive findings. However, this type of probe-filtering is arbitrary and it may remove highly significant findings that are robust to these technical defects. It could therefore be more appropriate to consult probe-lists after performing discovery analysis to identify and verify probes that do have an unreliable signal. For a specific example: one of the most robust signals that is associated with increased blood-

lipid levels (cg06500161) has a SNP within the underlying probe sequence (See Chapter 4). Surely the fact that this CpG is so strongly associated with a human trait does not warrant it's removal from analysis.

There are multiple ways one can design an EWAS and perform statistical analyses. Analyses of DNA methylation data done by either detecting differentially methylated positions (DMP) and differentially methylated region (DMR). Each of these have their own advantages. Identifying DMPs is straightforward and analogous to how GWAS are performed. Testing each available CpG enables the possibility to identify loci which have functional significance such as those located within transcription factor binding sites. Identification of DMRs potentially reduce the number of false positive findings caused by technical artefacts and reduce multiple testing thresholds but can be limited by both the coverage and the density of CpGs for each region as certain regions will have more loci available to estimate the overall DNA methylation pattern. However due to the considerable expansion of probes that comes with the EPIC microarray this problem could be overcome depending whether or not the coverage per region is more balanced. Realistically either of these approaches are suitable for EWAS but I would steer towards DMP approaches as there are numerous ways to handle false positive results, such as quality control.

Further to this, the type of model is also important in what results are obtained from analysis. Majority of the models that are used in EWAS are sensible as we have many years of GWAS to draw experience on. Due to the dynamic nature of the methylome and the often heterogeneous nature of methylomes obtained from common tissues, there has been a movement to include various extraneous variables within the statistical model. Some of these are reasonable such as cell-type composition estimations as not accounting for these will have drastic downstream effects. There are pipelines that suggest spending the time to identify surrogate variables or including principal components within models stating that such an approach will vastly improve the results. Ultimately, a well designed study and a thoughtful model should already contain the key variables to robustly produce meaningful results.

There are hundreds of different decisions that can be made while performing an EWAS and it is extremely important that any decisions made are communicated clearly. Doing so will enable the ease of reproduction and the longevity of the data in the future.

Finally it is possible that all of the tools and methods that were developed for the 450K and EPIC microarray could be inappropriate for extensively large datasets (n >10,000). It is likely that the current methods would probably be sufficient and I imagine that these could be used naively and produce results. However, I would feel much better if at some point in the future these methods were revised for large datasets but this would require a considerable amount of time and effort.

As we are entering the late stages of EWAS that focus solely on a single epigenetic factors it is expected that more studies will begin to combine different experiments to produce multi-omics based EWAS. Through combinations of methylomics, transcriptomics and other -omics related experiments the potential to truly disseminate the functional and mechanistic role the epigenome has upon gene regulation in a wide variety of human traits and disease. Such analyses will likely involve complicated methodologies and software to consistently collate together data from multiple experiments or to facilitate large-scale meta analyses of results from multiple studies. This will elevate EWAS to an extremely effective analysis technique which will be the foundation of identifying and developing novel therapeutic interventions.

Bibliography

References that are used in Chapter 3 are not included within this bibiliography because it was published and contains its own bibiliography.

Agnihotri S, Jalali S, Wilson MR, *et al.* (2016) The genomic landscape of schwannoma. *Nature Genetics*, **48**, 1339–1348.

Al Muftah WA, Al-Shafai M, Zaghlool SB, *et al.* (2016) Epigenetic associations of type 2 diabetes and BMI in an Arab population. *Clinical Epigenetics*, **8**, 13.

Arking DE, Chakravarti A (2009) Understanding cardiovascular disease through the lens of genome-wide association studies. *Trends in Genetics*, **25**, 387–394.

Arner P, Sinha I, Thorell A, Rydén M, Dahlman-Wright K, Dahlman I (2015) The epigenetic signature of subcutaneous fat cells is linked to altered expression of genes implicated in lipid metabolism in obese women. *Clinical Epigenetics*, **7**, 93.

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA (2014) Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

Aslibekyan S, Demerath EW, Mendelson M, *et al.* (2015) Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. *Obesity*, **23**, 1493–1501.

Assenov Y, Müller F, Lutsik P, Walter J, Lengauer T, Bock C (2014) Comprehensive analysis of DNA methylation data with RnBeads. *Nature methods*, **11**, 1138–40.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Benzeval M, Davillas A, Kumari M, Lynn P (2014) Understanding Society: The UK Household Longitudinal Study Biomarker User Guide and Glossary. URL https://www.understandingsociety.ac.uk/sites/default/files/downloads/legacy/7251-UnderstandingSociety-Biomarker-UserGuide-2014-1.pdf.

Bernstein BE, Meissner A, Lander ES (2007) The Mammalian Epigenome. *Cell*, **128**, 669–681.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, **28**, 1045–8.

Bestor TH (1992) Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. *The EMBO journal*, **11**, 2611–7.

Bestor TH (2000) The DNA methyltransferases of mammals. *Human molecular genetics*, **9**, 2395–2402.

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

Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, Gunderson KL (2009) Genome-wide DNA methylation profiling using Infinium assay. *Epigenomics*, **1**, 177–200.

Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes and Development*, **16**, 6–21.

Bird AP (1987) CpG islands as gene markers in the vertebrate nucleus. *Trends in Genetics*, **3**, 342–347.

Bonder MJ, Kasela S, Kals M, *et al.* (2014) Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC genomics*, **15**, 860.

Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, Balasubramanian S (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, **336**, 934–937.

Booth MJ, Ost TWB, Beraldi D, Bell NM, Branco MR, Reik W, Balasubramanian S (2013) Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nature protocols*, **8**, 1841–1851.

Branco MR, Ficz G, Reik W (2011) Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature Reviews Genetics*, **13**, 7–13.

Braun KVE, Dhana K, de Vries PS, *et al.* (2017) Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. *Clinical epigenetics*, **9**, 15.

Brazma A, Hingamp P, Quackenbush J, *et al.* (2001) Minimum information about a microarray experiment (MIAME)toward standards for microarray data. *Nature Genetics*, **29**, 365–371.

Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *American Journal of Human Genetics*, **88**, 450–457.

Chambers JC, Loh M, Lehne B, *et al.* (2015) Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: A nested case-control study. *The Lancet Diabetes and Endocrinology*, **3**, 526–534.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2018) shiny: Web Application Framework for R. URL `https://cran.r-project.org/package=shiny`.

Chanock SJ, Manolio T, Boehnke M, *et al.* (2007) Replicating genotypephenotype associations. *Nature*, **447**, 655–660.

Chasman DI, Paré G, Mora S, *et al.* (2009) Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS genetics*, **5**, e1000730.

Cho E, Mysliwiec MR, Carlson CD, Ansari A, Schwartz RJ, Lee Y (2018) Cardiac-specific developmental and epigenetic functions of Jarid2 during embryonic development. *Journal of Biological Chemistry*, **293**, 11659–11673.

Choufani S, Cytrynbaum C, Chung BH, *et al.* (2015) NSD1 mutations generate a genome-wide DNA methylation signature. *Nature Communications*, **6**, 10207.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, **4**, 265–270.

Costa FF (2008) Non-coding RNAs, epigenetics and complexity. *Gene*, **410**, 9–17.

Crujeiras AB, Diaz-Lagares A, Moreno-Navarrete JM, *et al.* (2016) Genome-wide DNA methylation pattern in visceral adipose tissue differentiates insulin-resistant from insulin-sensitive obese subjects. *Translational Research*, **178**, 13–24.e5.

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

Dayeh T, Tuomi T, Almgren P, *et al.* (2016) DNA methylation of loci within ABCG1 and PHOSPHO1 in blood DNA is associated with future type 2 diabetes risk. *Epigenetics*, **11**, 482–488.

Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F (2011) Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, **3**, 771–784.

Dekkers KF, Slagboom PE, Jukema JW, Heijmans BT (2016a) The multifaceted interplay between lipids and epigenetics. *Current Opinion in Lipidology*, **27**, 288–294.

Dekkers KF, van Iterson M, Slieker RC, *et al.* (2016b) Blood lipids influence DNA methylation in circulating cells. *Genome Biology*, **17**, 138.

Demerath EW, Guan W, Grove ML, *et al.* (2015) Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Human Molecular Genetics*, **24**, 4464–4479.

Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Devlin B, Roeder K, Wasserman L (2001) Genomic Control, a New Approach to Genetic-Based Association Studies. *Theoretical Population Biology*, **60**, 155–166.

Du P, Zhang X, Huang CC, *et al.* (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, Gehrke C (1982) Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic acids research*, **10**, 2709–21.

Feil R, Fraga MF (2012) Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics*, **13**, 97–109.

Feinberg AP (2007) Phenotypic plasticity and the epigenetics of human disease. *Nature*, **447**, 433–440.

Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, **52**, 1694–1711.

Florath I, Butterbach K, Heiss J, Bewerunge-Hudler M, Zhang Y, Schöttker B, Brenner H (2016) Type 2 diabetes and leucocyte DNA methylation: an epigenome-wide association study in over 1,500 older adults. *Diabetologia*, **59**, 130–138.

Ford EE, Grimmer MR, Stolzenburg S, *et al.* (2017) Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation. *bioRxiv*, p. 170506.

Fortin JP, Fertig E, Hansen K (2014a) shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. *F1000Research*, **3**, 175.

Fortin JP, Labbe A, Lemire M, *et al.* (2014b) Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*, **15**, 503.

Fortin JP, Triche TJ, Hansen KD (2016) Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*, **30**, btw691.

Frazier-Wood AC, Aslibekyan S, Absher DM, *et al.* (2014) Methylation at CPT1A locus is associated with lipoprotein subfraction profiles. *Journal of lipid research*, **55**, 1324–30.

Friedewald WT, Levy RI, Fredrickson DS (1972) Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical chemistry*, **18**, 499–502.

Frommer M, McDonald LE, Millar DS, *et al.* (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 1827–31.

Gagnon F, Aïssi D, Carrié A, Morange PE, Trégouët DA (2014) Robust validation of methylation levels association at CPT1A locus with lipid plasma levels. *Journal of lipid research*, **55**, 1189–1191.

Gagnon-Bartsch J (2018) *ruv: Detect and Remove Unwanted Variation using Negative Controls*, URL `https://cran.r-project.org/package=ruv`.

Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *Journal of molecular biology*, **196**, 261–82.

Gentleman RC, Carey VJ, Bates DM, *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**, R80.

Gorrie-Stone TJ, Smart MC, Saffari A, *et al.* (2018) Bigmelon: tools for analysing large DNA methylation datasets. *Bioinformatics*, p. 10.1093/bioinformatics/bty713.

Greally JM (2018) A user's guide to the ambiguous word 'epigenetics'. *Nature Reviews Molecular Cell Biology*, **19**, 207–208.

Guay SP, Brisson D, Munger J, Lamarche B, Gaudet D, Bouchard L (2012) ABCA1 gene promoter DNA methylation is associated with HDL particle profile and coronary artery disease in familial hypercholesterolemia. *Epigenetics*, **7**, 464–472.

Guo J, Su Y, Zhong C, Ming Gl, Song H (2011) Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. *Cell*, **145**, 423–434.

Haig D (2004) The (dual) origin of epigenetics. In: *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 69, pp. 67–70. URL `http://www.ncbi.nlm.nih.gov/pubmed/16117634http://symposium.cshlp.org/cgi/doi/10.1101/sqb.2004.69.67`.

Hannon E, Gorrie-Stone TJ, Smart MC, *et al.* (2018) Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. *American journal of human genetics*, **103**, 654–665.

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

Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, **405**, 486–489.

Hedman ÅK, Mendelson MM, Marioni RE, *et al.* (2017) Epigenetic Patterns in Blood Associated With Lipid Traits Predict Incident Coronary Heart Disease Events and Are Enriched for Results From Genome-Wide Association Studies. *Circulation: Cardiovascular Genetics*, **10**, e001487.

Heiss JA, Just AC (2018) Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. *Clinical epigenetics*, **10**, 73.

Herman JG, Baylin SB (2003) Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine*, **349**, 2042–2054.

Hindorff La, Sethupathy P, Junkins Ha, Ramos EM, Mehta JP, Collins FS, Manolio Ta (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9362–7.

Horvath S (2013) DNA methylation age of human tissues and cell types. *Genome Biology*, **14**, R115.

Houseman EA, Accomando WP, Koestler DC, *et al.* (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, **13**, 86.

Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE*, **5**, e8888.

Illingworth RS, Bird AP (2009) CpG islands - 'A rough guide'. *FEBS Letters*, **583**, 1713–1720.

Inoshita M, Numata S, Tajima A, *et al.* (2015) Sex differences of leukocytes DNA methylation adjusted for estimated cellular proportions. *Biology of Sex Differences*, **6**, 11.

Irizarry RA, Ladd-Acosta C, Wen B, *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*, **41**, 178–86.

Irvin MR, Zhi D, Joehanes R, *et al.* (2014) Epigenome-Wide Association Study of Fasting Blood Lipids in the Genetics of Lipid-Lowering Drugs and Diet Network Study. *Circulation*, **130**, 565–572.

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

Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, **41**, 200–209.

Johansson Å, Enroth S, Gyllensten U (2013) Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLoS ONE*, **8**, e67378.

Johnson AD, O'Donnell CJ (2009) An open access database of genome-wide association results. *BMC medical genetics*, **10**, 6.

Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Jones A, Wang H (2010) Polycomb repressive complex 2 in embryonic stem cells: An overview. *Protein and Cell*, **1**, 1056–1062.

Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, **13**, 484–492.

Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nature reviews. Genetics*, **3**, 415–28.

Kalderimis A, Lyne R, Butano D, *et al.* (2014) InterMine: Extensive web services for modern biology. *Nucleic Acids Research*, **42**, W468–W472.

Karlsson Linnér R, Marioni RE, Rietveld CA, *et al.* (2017) An epigenome-wide association study meta-analysis of educational attainment. *Molecular Psychiatry*, **22**, 1680–1690.

Kazmi N, Gaunt TR, Relton C, Micali N (2017) Maternal eating disorders affect offspring cord blood DNA methylation: A prospective study. *Clinical Epigenetics*, **9**, 120.

Kessler T, Vilne B, Schunkert H (2016) The impact of genome-wide association studies on the pathophysiology and therapy of cardiovascular disease. *EMBO Mol Med*, **8**, 688–701.

Kettunen J, Tukiainen T, Sarin AP, *et al.* (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genetics*, **44**, 269–276.

Klose RJ, Bird AP (2006) Genomic DNA methylation: The mark and its mediators. *Trends in Biochemical Sciences*, **31**, 89–97.

Knies G (2015) The UK Household Longitudinal Study Waves 1-7 User Guide. URL `https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/mainstage/user-guides/mainstage-user-guide.pdf`.

Kok DEG, Dhonukshe-Rutten RAM, Lute C, *et al.* (2015) The effects of long-term daily folic acid and vitamin B12 supplementation on genome-wide DNA methylation in elderly subjects. *Clinical Epigenetics*, **7**, 121.

Korthauer K, Irizarry RA (2018) Genome-wide repressive capacity of promoter DNA methylation is revealed through epigenomic manipulation. *bioRxiv*, p. 381145.

Kulkarni H, Kos MZ, Neary J, *et al.* (2015) Novel epigenetic determinants of type 2 diabetes in Mexican-American families. *Human Molecular Genetics*, **24**, 5330–5344.

Kundaje A, Meuleman W, Ernst J, *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Laurent L, Wong E, Li G, *et al.* (2010) Dynamic changes in the human methylome during differentiation. *Genome Research*, **20**, 320–331.

Lee KWK, Pausova Z (2013) Cigarette smoking and DNA methylation. *Frontiers in Genetics*, **4**, 132.

Leek JT, Johnson WE, Parker HS, *et al.* (2017) *sva: Surrogate Variable Analysis*, URL `http://bioconductor.org/packages/sva/`.

Lehne B, Drong AW, Loh M, *et al.* (2015) A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biology*, **16**, 37.

Li E, Beard C, Jaenisch R (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.

Li M, Zou D, Li Z, *et al.* (2019) EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Research*, **47**, D983–D988.

Lister R, Pelizzola M, Dowen RH, *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–22.

Liu Y, Aryee MJ, Padyukov L, *et al.* (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, **31**, 142–147.

Lowe R, Rakyan VK (2013) Marmal-aid–a database for Infinium HumanMethylation450. *BMC bioinformatics*, **14**, 359.

Maksimovic J, Gordon L, Oshlack A (2012) SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome biology*, **13**, R44.

Mamtani M, Kulkarni H, Dyer TD, *et al.* (2016) Genome- and epigenome-wide association study of hypertriglyceridemic waist in Mexican American families. *Clinical Epigenetics*, **8**, 1–14.

Marioni RE, Shah S, McRae AF, *et al.* (2015) DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biology*, **16**, 25.

Meeks KA, Henneman P, Venema A, *et al.* (2017) An epigenome-wide association study in whole blood of measures of adiposity among Ghanaians: The RODAM study. *Clinical Epigenetics*, **9**, 103.

Mendelson MM, Marioni RE, Joehanes R, *et al.* (2017) Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLoS Medicine*, **14**, e1002215.

Mendis S, Puska P, Norrving B (2011) *Global Atlas on cardiovascular disease prevention and control Joint Publication of the World Health Organization the World Heart*. Geneva: World Health Organization, 1–166 pp., URL `https://www.who.int/cardiovascular{_}diseases/publications/atlas{_}cvd/en/`.

Michels KB, Binder AM, Dedeurwaerder S, *et al.* (2013) Recommendations for the design and analysis of epigenome-wide association studies. *Nature methods*, **10**, 949–55.

Mill J, Heijmans BT (2013) From promises to practical strategies in epigenetic epidemiology. *Nature reviews. Genetics*, **14**, 585–94.

Min JL, Hemani G, Davey Smith G, Relton C, Suderman M (2018) Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics*, **34**, 3983–3989.

Mittelstraß K, Waldenberger M (2018) DNA methylation in human lipid metabolism and related diseases. *Current Opinion in Lipidology*, **29**, 116–124.

Moran S, Arribas C, Esteller M (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, **8**, 389–399.

Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S (2014) ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics*, **30**, 428–430.

Petersen AK, Zeilinger S, Kastenmüller G, *et al.* (2014) Epigenetics meets metabolomics: An epigenome-wide association study with blood serum metabolic traits. *Human Molecular Genetics*, **23**, 534–545.

Pfeiffer L, Wahl S, Pilling LC, *et al.* (2015) DNA Methylation of Lipid-Related Genes Affects Blood Lipid Levels. *Circulation: Cardiovascular Genetics*, **8**, 334 – 342.

Phipson B, Maksimovic J, Oshlack A (2015) MissMethyl: An R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*, **32**, 286–288.

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC genomics*, **14**, 293.

Pidsley R, Zotenko E, Peters TTJ, *et al.* (2016) Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, **17**, 208.

Portales-Casamar E, Lussier AA, Jones MJ, *et al.* (2016) DNA methylation signature of human fetal alcohol spectrum disorder. *Epigenetics & Chromatin*, **9**, 25.

Portela A, Esteller M (2010) Epigenetic modifications and human disease. *Nature biotechnology*, **28**, 1057–1068.

R Core Team (2017) R: A Language and Environment for Statistical Computing. URL `https://www.r-project.org/`.

Rakyan VK, Down Ta, Balding DJ, Beck S (2011) Epigenome-wide association studies for common human diseases. *Nature reviews. Genetics*, **12**, 529–41.

Razin A, Cedar H (1994) DNA methylation and genomic imprinting. *Cell*, **77**, 473–476.

Razin A, Riggs AD (1980) DNA Methylation and Gene Function. *Science*, **210**, 604–610.

Reik W, Dean W, Walter J (2001) Epigenetic Reprogramming in Mammalian Development. *Science*, **293**, 1089–1093.

Relton CL, Davey Smith G (2012) Two-step epigenetic mendelian randomization: A strategy for establishing the causal role of epigenetic processes in pathways to disease. *International Journal of Epidemiology*, **41**, 161–176.

Riggs AD (1975) X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, **14**, 9–25.

Rushton MD, Reynard LN, Barter MJ, Refaie R, Rankin KS, Young DA, Loughlin J (2014) Characterization of the cartilage DNA methylome in knee and hip osteoarthritis. *Arthritis and Rheumatology*, **66**, 2450–2460.

Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 1412–1417.

Sayols-Baixeras S, Irvin MR, Elosua R, Arnett DK, Aslibekyan SW (2016a) Epigenetics of Lipid Phenotypes. *Current Cardiovascular Risk Reports*, **10**, 31.

Sayols-Baixeras S, Subirana I, Lluis-Ganella C, *et al.* (2016b) Identification and validation of seven new loci showing differential DNA methylation related to serum lipid profile: an epigenome-wide approach. The REGICOR study. *Human Molecular Genetics*, **25**, 4556–4565.

Sender R, Fuchs S, Milo R (2016) Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, **14**, e1002533.

Singer-Sam J, Riggs AD (2012) X chromosome inactivation and DNA methylation. In: *DNA Methylation*, pp. 358–384. Birkhäuser Basel, Basel, URL `http://link.springer.com/10.1007/978-3-0348-9118-9{_}16`.

Skuladottir GV, Nilsson EK, Mwinyi J, Schiöth HB (2016) One-night sleep deprivation induces changes in the DNA methylation and serum activity indices of stearoyl-CoA desaturase in young healthy men. *Lipids in Health and Disease*, **15**, 137.

Smith RN, Aleksic J, Butano D, *et al.* (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, **28**, 3163–3165.

Spiers H, Hannon E, Schalkwyk LC, *et al.* (2015) Methylomic trajectories across human fetal brain development. *Genome Research*, **25**, 338–352.

Tan Q, Frost M, Heijmans BT, von Bornemann Hjelmborg J, Tobi EW, Christensen K, Christiansen L (2014) Epigenetic signature of birth weight discordance in adult twins. *BMC Genomics*, **15**, 1062.

Teschendorff AE, Breeze CE, Zheng SC, Beck S (2017) A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC bioinformatics*, **18**, 105.

Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S (2013) A

beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, **29**, 189–196.

Teschendorff AE, Relton CL (2018) Statistical and integrative system-level analysis of DNA methylation data. *Nature Reviews Genetics*, **19**, 129–147.

Tobi EW, Lumey LH, Talens RP, *et al.* (2009) DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Human Molecular Genetics*, **18**, 4046–4053.

Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD (2013) Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research*, **41**, e90–e90.

Tserel L, Kolde R, Limbach M, *et al.* (2015) Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. *Scientific Reports*, **5**, 13107.

Van Iterson M, Cats D, Hop P, Heijmans BT (2018) OmicsPrint: Detection of data linkage errors in multiple omics studies. *Bioinformatics*, **34**, 2142–2143.

van Iterson M, Tobi EW, Slieker RC, den Hollander W, Luijk R, Slagboom PE, Heijmans BT (2014) MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics*, **30**, 3435–3437.

van Iterson M, van Zwet EW, Heijmans BT (2017) Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology*, **18**, 19.

Varley KE, Gertz J, Bowling KM, *et al.* (2013) Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Research*, **23**, 555–567.

Voorman A, Lumley T, McKnight B, Rice K (2011) Behavior of QQ-plots and Genomic Control in studies of gene-environment interaction. *PLoS ONE*, **6**.

Waddington CH (1942) The epigenotype. *Endeavours*, **1**, 18–20.

Waddington CH (1957) *The Strategy of the Genes*. Routledge, URL `https://www.taylorfrancis.com/books/9781315765471`.

Wahl S, Drong A, Lehne B, *et al.* (2017) Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, **541**, 81–86.

Wang J, Zhao Q (2015) *cate: High Dimensional Factor Analysis and Confounder Adjusted Testing and Estimation*, URL `https://cran.r-project.org/package=cate`.

Wang Y, Leung FCC (2004) An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics*, **20**, 1170–1177.

Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, **39**, 457–466.

Welter D, MacArthur J, Morales J, *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, **42**, D1001–D1006.

Willer CJ, Schmidt EM, Sengupta S, *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, **45**, 1274–1283.

Wilson LE, Harlid S, Xu Z, Sandler DP, Taylor JA (2017) An epigenome-wide study of body mass index and DNA methylation in blood using participants from the Sister Study cohort. *International Journal of Obesity*, **41**, 194–199.

Wu TP, Wang T, Seetin MG, *et al.* (2016) DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature*, **532**, 329–333.

Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G (2014) Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer cell*, **26**, 577–90.

Zhang X, Hu Y, Justice AC, *et al.* (2017) DNA methylation signatures of illicit drug injection and hepatitis C are associated with HIV frailty. *Nature Communications*, **8**, 2243.

Zhang X, Justice AC, Hu Y, *et al.* (2016) Epigenome-wide differential DNA methylation between HIV-infected and uninfected individuals. *Epigenetics*, **11**, 750–760.

Zhou W, Laird PW, Shen H (2017) Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic acids research*, **45**, e22.

Zhou W, Triche TJ, Laird PW, Shen H (2018) SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic acids research*, **46**, e123.

Ziller MJ, Gu H, Müller F, *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–81.

Appendices

Appendix A

Chapter 3 - Supplementary Materials

3

Writing `bigmelon`-ised Functions

Tyler J. Gorrie-Stone

Preface

This is an introduction to implementing functions for bigmelon. Given enough memory, it's straightforward to extract the complete intensity or beta metrix from a .gds file and work on that. In order to make a function that works memory-efficiently on large datasets though, you need to think about how the function works, and what subsets you should extract to work on sequentially or in parallel.

We will demonstrate using `bumphunter::bumphunter` as an example and the six 450k samples from the minfiData package as example data.

We will assume you have also read the package vignettes.

```
library(bigmelon)
library(parallel)       # optional, for parallel processing examples
library(microbenchmark) # optional, to demonstrate some code performance
library(bumphunter)
library(minfiData)
  # make a gdsfile
  bd <- system.file('extdata', package='minfiData')
  gfile <- iadd2(bd, gds = 'melon.gds')
  closefn.gds(gfile)
  # open the file again, allowing forking (important for multicore processing)
  gfile <- openfn.gds('melon.gds', allow.fork = T)
```

Accessing Data, loops and `apply`

Preprocessing steps such as quantile normalisation, tend to involve processing an array at a time, ie looping over columns or `apply` on margin=2. Analyses are more often probewise, ie looping over rows or `apply` on margin=1. In either case it is important to analyse what the function has to keep from these operations and whether that has to be kept in a memory-efficient form. This generally comes down to whether it is a column or row summary (ie output is one or a few rows/columns) or the same shape as the input. There is an `apply.gdsn` function that optionally keeps the output as a gdsfile node.

Because of the overhead of file access, it's also worth considering combining several operations into a pass over the matrix instead of making several pass

Accessing Data

Within `bigmelon` we provide user friendly `[` functions to enable users to directly access data similar to that of an `expressionSet` object like a `MethylSet` object. This is described, in the vignette. These are particularly useful for interactive use if you are inclined to look at certain regions.

```
# Pulling out the first row of the data-set
betas(gfile)[1, 1:4, name = TRUE]
```

```
## 5723646052/5723646052_R02C02 5723646052/5723646052_R04C01
##                    0.4143280                    0.3733613
## 5723646052/5723646052_R05C02 5723646053/5723646053_R04C02
##                    0.2125911                    0.1893959
```

```
# Alternative
gfile[1, 1:4, node = 'betas', name = FALSE]
```

```
## [1] 0.4143280 0.3733613 0.2125911 0.1893959
```

There is little difference between the two above examples. You can access all data as you normally would using logical, character or numerical indexing. A key distinction is the `name` argument will provide the dimnames of the resultant vector/matrix. In the first example we are using a familiar function `betas` on `gfile` and then indexing. While in the second example we are calling the `gfile` object directly and adding an additional argument within the `[` function to call a specific node that we are interested in selecting. This is particularly useful for calling data that does not have a standard name or a function associated with it.

Alternatively you can use `readex.gdsn`, which `[` calls, directly. In most cases this is faster but requires a list of indices and does not provide dimnames

```
node <- index.gdsn(gfile, 'betas') # target specific node of interest
readex.gdsn(node = node, sel = list(1, 1:4))
```

```
## [1] 0.4143280 0.3733613 0.2125911 0.1893959
```

Lastly, and most importantly, accessing data by column is considerably faster than accessing data by row! So in terms of performing analysis if you can restructure the code to handle columns instead of rows the time spent accessing data is greatly reduced.

Looping Examples

Now that we know how to access data we can begin with some loops. Here we will compare a few ways that a for-loop can be done - and evaluate the caveats of each, and then compare it with the `apply`-like functions in `gdsfmt`. In these examples we will emulate `colSums` for a gds object.

```
# Example 1 using `[`
sums1 <- function(gfile){
  sums <- vector('numeric', length(colnames(gfile)))
  for(i in seq_along(colnames(gfile))){
    sums[i] <-sum(betas(gfile)[,i], na.rm = TRUE)
  }
  sums
}
microbenchmark(sums1(gfile), times = 10)
```

```
## Unit: seconds
##          expr      min       lq     mean   median       uq      max neval
##  sums1(gfile) 1.423583 1.424809 1.464565 1.469261 1.494947 1.501467    10
```

```
# Example 2 using readex.gdsn
sums2 <- function(gfile){
  sums <- vector('numeric', length(colnames(gfile)))
  for(i in seq_along(colnames(gfile))){
    sums[i] <- sum(readex.gdsn(index.gdsn(gfile, 'betas'), sel = list(NULL, i)), na.rm = TRUE)
  }
  sums
}

microbenchmark(sums2(gfile), times = 10)
```

```
## Unit: milliseconds
##          expr      min       lq    mean   median       uq      max neval
```

```
##  sums2(gfile) 29.23284 29.44034 30.4932 29.83899 30.17611 36.85755    10
```

The time difference between `[` and using `readex.gdsn` is noticeable, while it is relatively small here it could be a problem in larger data-sets.

Alternatively the same result can be achieved with `sapply`

```
sums3 <- function(gfile){
  sums <- sapply(seq_along(colnames(gfile)), function(i, gfile){
    sum(readex.gdsn(index.gdsn(gfile, 'betas'), sel = list(NULL, i)), na.rm = TRUE)
  }, gfile = gfile)
  sums
}
microbenchmark(sums3, times = 10)
```

```
## Unit: nanoseconds
##   expr min lq  mean median uq  max neval
##  sums3  70 70 279.9   70.5 71 2165    10
```

apply-like functions

The `apply.gdsn` is **usually** faster than any for-loop in R and has added benefits that it can store the output directly into a gds node should you prefer it, usually this allows for cleaner code and looks nicer in my opinion. I highly recommend reading the manual pages for `apply.gdsn`! Here we are also able to compute the `colSums` of the small matrix in rapid time.

```
sums4 <- function(gfile){
  sums <- apply.gdsn(node = index.gdsn(gfile, 'betas'),
    margin = 2, # colSums
    FUN = sum,
    selection = NULL,
    # Otherwise selection can be a list akin to readex.gdsn
    as.is = "double",
    # Can be "list", "none", "character", "logical", "gdsnode"
    na.rm = TRUE # Other arg for sum!
  )
  sums
}

microbenchmark(sums4, times = 10)
```

```
## Unit: nanoseconds
##   expr min lq  mean median uq max neval
##  sums4  70 70 168.3     71 71 978    10
```

This is comparible with the other examples, however the benefit of using `apply.gdsn` lies in its ability to process data by rows. (see below).

```
sums5 <- function(gfile){
  sums <- apply.gdsn(node = index.gdsn(gfile, 'betas'),
    margin = 1, # rowSums
    FUN = sum,
    selection = NULL,
    # Otherwise selection can be a list akin to readex.gdsn
    as.is = "double",
    # Can be "list", "none", "character", "logical", "gdsnode"
    na.rm = TRUE # Other arg for sum!
```

```
  )
  sums
}

microbenchmark(sums5, times = 10)
```

```
## Unit: nanoseconds
##   expr min lq mean median uq  max neval
##  sums5  70 70  252     71 71 1816    10
```

As we can see, `apply.gdsn` wastes little time in computing the `rowSums` of a matrix. If we were to do this with a for loop, we would be here for a very long time.

Depending on what you are using `apply.gdsn` for it is usually possible to parallelise it by replacing it with `clusterApply.gdsn` or writing a `mclapply` function. **n.b.** `clusterApply.gdsn` has problems when being used within functions (and in this Rmarkdown document and cannot be demonstrated but below is an example of how to use it). Also doing things in parallel will use more memory.

Another distinction of `clusterApply.gdsn` is that is cannot write data to a gds file, if you ever wish to do this; you will need to use a for loop or `apply.gdsn`.

```
  cl <- makeCluster(2)
  sums <- clusterApply.gdsn(cl = cl,
     gds.fn = gfile[[1]],
     # gfile[[1]] is the absolute path of the gdsfile
     node.name = "betas",
     margin = 1,
     FUN = sum,
     selection = NULL,
     as.is = 'double'
   )
stopCluster(cl)
```

```
mcsums <- function(gfile){
  sums <- mclapply(seq_along(colnames(gfile)), FUN = function(i, gfile){
     sum(readex.gdsn(
        index.gdsn(gfile, 'betas'),
        sel = list(NULL, i)),
        na.rm = TRUE
     )
   }, gfile = gfile, mc.cores = 2)
   sums
}
microbenchmark(mcsums(gfile), times = 10)
```

```
## Unit: milliseconds
##           expr      min       lq     mean   median       uq      max neval
##  mcsums(gfile) 86.84666 102.3932 115.1241 109.1796 130.3954 148.4915    10
```

There's a balance between the added hassle of parallelising methods and the speedup that it produces. For most analyses to date we have managed without it.

A Note

Some operations may not be parallelisable. These include copying data from one gds-object to another, doing operations that require the results of a previous iteration and writing data to a gds file. In general

terms, always try to do things using `apply.gdsn` or `clusterApply.gdsn`. If you cannot move onto a for-loop, sapply, mclapply etc., if you are attempting to iterate over rows consider chunking the matrix into 1000 by p matrix and load and process each chunk into RAM instead of loading the individual row (loading a small chunk (up to 1000 rows) is just as fast loading a single row). If all else fails, bite the bullet and load the entire matrix into memory or find a heuristic approach to the problem.

An example of optimisation: bumphunter

We continue to implement bigmelon methods for popular EWAS realated functions, but can't anticipate all users' needs. It is often possible to do this yourself without rewriting very much of the code.

We will be optimising the `bumphunter` function. The code is quite long and there is a lot to go through but I will try to describe some of my thought process behind the optimisation.

Review the Code:

The bumphunter function is an interesting function and heavily used in the realms of EWAS. What distinguishes it from other functions is that it computes its own test statistics and provides two cross-validation methods for the tools. As a result it involves a lot of matrix arithmetic on the entire dataset. I of course am not the original author of the code and the coments and documentation can be a huge help in breaking the problem down.

After some careful review I was able to narrow down that the optimisation of bumphunter could be acheived by rewriting two parts of the function. Specifically these parts required the entire dataset. With the current test data it is a 485577 x 6 matrix) but in the intended use case it could be multiple Gb.

To begin optimisation we can start by making a direct copy of the code and strip out any of the preliminary checks that consider data sanity (e.g `stopifnot(is.matrix(data))`).

I made the following changes to the start of the code:

```
n <- objdesp.gdsn(mat)$dim[1] # new
p <- objdesp.gdsn(mat)$dim[2] # new
#if (!is.matrix(mat))
#  stop("'mat' must be a matrix.")
if (p != nrow(design))
  stop("Number of columns of 'mat' must  match number of rows of 'design'")
```

Maybe I should have used better names than n and p, but it's useful to have the dimensions stored at the beginning and to check they make sense.

Step 1: 'Permutation'

Moving onto the first piece of code we will need to optimise: This is the `.getEstimate` function - which as the name implies computes the beta estimates for the model you intend to run. The code is very fast and it is a shame that we need to break it down into something slower to make it memory efficient.

Here is what I came up with (Changes I have made I have indicated with a `##` the line above:

```
.getEstimate2 <- function(mat, design, coef, B = NULL, permutations = NULL, full = FALSE){
  ##
  p <- objdesp.gdsn(mat)$dim[2]
  ##
  n <- objdesp.gdsn(mat)$dim[1]
  v <- design[, coef]
```

```
A <- design[, -coef, drop = FALSE]
qa <- qr(A)
S <- diag(nrow(A)) - tcrossprod(qr.Q(qa)) # ncol * ncol matrix, "small"
vv <- if(is.null(B)){
  matrix(v, ncol = 1)
} else {
  if (is.null(permutations)) {
    replicate(B, sample(v))
  } else {
    apply(permutations, 2, function(i) v[i])
  }
}
sv <- S %*% vv
vsv <- diag(crossprod(vv, sv))
#b <- (mat %*% crossprod(S, vv))/vsv
#  if (!is.matrix(b))
#      b <- matrix(b, ncol = 1)
if(full){
  # sy <- mat %*% S
  df.residual <- p - qa$rank - 1
  if(is.null(B)){
  ## New Chunk
    o <- apply.gdsn(node = mat, margin = 1, as.is = 'list',
       FUN = function(x, S, vv, vsv, sv, df.residual){
         sy <- x %*% S
         b <- (x %*% crossprod(S, vv))/vsv
         tcross <- tcrossprod(b, sv)
         sigma <- sum((sy - tcross)^2)/df.residual
         list('B'=b, 'sigma'= sigma)
       }, S = S, vv = vv, vsv = vsv, sv = sv, df.residual = df.residual
    )
  } else {
    o <- apply.gdsn(node=mat, margin=1, as.is = 'list',
       FUN = function(x, S, vv, vsv, sv, B, df.residual){
         tmp <- sy <- x %*% S
         sigma <- b <- (x %*% crossprod(S, vv))/vsv
         for(j in seq_len(B)){
           tmp <- tcrossprod(b[,j], sv[,j])
           sigma[j] <- sum((sy-tmp)^2)
         }
         sigma <- sqrt(sigma/df.residual)
         list('B'= b, 'sigma'=sigma)
       }, S = S, vv = vv, vsv = vsv, sv = sv, df.residual = df.residual, B = B
    )
  }
  coef  <- if(is.null(B)) sapply(o, '[[', 'B') else t(sapply(o, '[[', 'B'))
  sigma <- if(is.null(B)) sapply(o, '[[', 'sigma') else t(sapply(o,
     '[[', 'sigma'))
  out <- list(coef = coef, # n * B big
      sigma = sigma, # n * B big
      stdev.unscaled = sqrt(1/vsv),
      df.residual = df.residual)
  if(is.null(B)) out$stdev <- as.numeric(out$stdev)
```

```
  } else {
    out <- apply.gdsn(node=mat, margin = 1, as.is = 'list',
       FUN = function(x, S, vv, vsv){
         b <- (x %*% crossprod(S, vv))/vsv
       }, S = S, vv = vv, vsv = vsv
    )
    out <- do.call(rbind, out)
  }
  ## End new Chunk
  return(out)
}
```

There is a lot to take in but we are certain the new function works!

```
mat <- betas(gfile)
design <- model.matrix(~c(1,1,1,2,2,2))
head(bumphunter:::.getEstimate(mat = mat[,] , design = design, coef = 2, B=NULL, full = F))
```

```
##                    [,1]
## cg00000029 -0.089415463
## cg00000108 -0.014796262
## cg00000109 -0.008447514
## cg00000165  0.182835398
## cg00000236  0.007545778
## cg00000289 -0.048587910
```

```
head(.getEstimate2(mat = mat, design = design, coef = 2, B=NULL, full = F))
```

```
##              [,1]
## [1,] -0.089415463
## [2,] -0.014796262
## [3,] -0.008447514
## [4,]  0.182835398
## [5,]  0.007545778
## [6,] -0.048587910
```

We must remember to at some point relabel the dimnames, this can be usually be done at the end.

There is a lot to unpack here. So we will begin at the top and work down:

The code is remarkably different from the original code (the parts that have been commented out). Most notably I have moved the two large cross products (`b <- (mat %*% crossprod(S, vv))/vsv` and `sy <- mat %*% S`) within `apply.gdsn`, and modified the structures of the code so that they will compute the crossproduct of a single row.

Taking a closer look at one of the `apply.gdsn`'s being used here...

```
... # Rest of code above
      o <- apply.gdsn(node=mat, margin=1, as.is = 'list',
          FUN = function(x, S, vv, vsv, sv, B, df.residual){
            tmp <- sy <- x %*% S
            sigma <- b <- (x %*% crossprod(S, vv))/vsv
            for(j in seq_len(B)){
              tmp <- tcrossprod(b[,j], sv[,j])
              sigma[j] <- sum((sy-tmp)^2)
            }
            sigma <- sqrt(sigma/df.residual)
            list('B'= b, 'sigma'=sigma)
```

```
          },
          S = S, vv = vv, vsv = vsv, sv = sv,
          df.residual = df.residual, B = B
      )
    coef <- if(is.null(B)) sapply(o, '[[', 'B') else t(sapply(o, '[[', 'B'))
    sigma <- if(is.null(B)) sapply(o, '[[', 'sigma') else t(sapply(o, '[[', 'sigma'))
    out <- list(coef = coef, # n * B big
        sigma = sigma, # n * B big
        stdev.unscaled = sqrt(1/vsv),
        df.residual = df.residual)
... # Rest of code
```

The structure is somewhat similar to a regular `apply` or `lapply` but with a few differences. The `as.is` specifies the output format, this can can numeric, character, a list or a gds file (which we will see later).

Here we can see that for each row of `mat` we compute using the `%*%` and then convert the output into the correct format at the end of the apply. Since the output of `.getEstimate` is at minimum 2 matricies of length n, and B columns. This is fairly small in terms of memory usage, so we are comfortable with keeping this in memory. If we suspect that we would have a `B > 1000` then we may want to consider storing the output into a gds file, and thus we would need to change the code to store a large matrix. We provide the `apply.gdsn` with static elements of the function such as `S`, `vv`, etc. so that we do not have to continuously recalculate them as this can eat into computation time, expecially when these variables are quite small.

Step 2: Bootstrapping

After computing the estimates, we need to do the null boot-strapping or null permutations. We already optimised the permutation step by updating `.getEstimate` so we can look towards the boot-strapping part of bumphunter. The bootstrapping section makes use of the `foreach` package to do some multicore processing if specified to, but we will initially do the analysis on a single core.

Here is what I came up with:

```
if (nullMethod == "bootstrap"){
  message("[bumphunterEngine] Performing ", B, " bootstraps.")
  qr.X <- qr(design)
  ##rescale residuals
  h <- diag(tcrossprod(qr.Q( qr(design))))
  ##create the null model to which we add bootstrap resids
  design0 <- design[,-coef,drop=FALSE]
  qr.X0 <- qr(design0)
  ##
  boots <- createfn.gds('bs.gds', allow.duplicate = TRUE)
  res <- add.gdsn(node = boots, name='resids', val = NULL, storage = 'float64',
                  valdim = c(p,0))
  null <- add.gdsn(node = boots, name='null', val = NULL, storage = 'float64',
                  valdim = c(p,0))
  apply.gdsn(node = mat, margin = 1, as.is = 'gdsnode', target.node = list(x=res, y=null),
             FUN = function(x, s1, s2, n1){
               res <- t(s1 %*% x)/s2
               null <- t(n1 %*% x)
               list(x=res, y=null)
              }, s1 = t(diag(nrow(design)) - tcrossprod(qr.Q(qr.X))),
                 s2 = sqrt(1-h), n1 = tcrossprod(qr.Q(qr.X0))
            )
```

```
  ##Now do the boostraps
  chunksize <- ceiling(B/workers)
  bootIndexes<-replicate(B, sample(1:p,replace=TRUE),simplify=TRUE)
#    tmp <- foreach(bootstraps = iter(bootIndexes, by = "column", chunksize = chunksize),
#     .combine = "cbind", .packages = "bumphunter") %dorng% {
#         apply(bootstraps, 2, function(bootIndex){
#             ##create a null model
#             matstar <- null+resids[,bootindex]
##              ##compute the null beta estimate
#            nullbetas <- backsolve(qr.R(qr.X),crossprod(qr.Q(qr.X),t(matstar)))[coef,]
#           if (useWeights){
#                 ##compute sigma
#                  sigma <- rowSums(t(tcrossprod( diag(nrow(design)) -
#                           tcrossprod(qr.Q(qr.X)),  matstar))^2)
#                   sigma <-
#                        sqrt(sigma/(nrow(design)-qr.X$rank))
#                     outList <- list(coef=nullbetas,sigma=sigma)
#                  } else {
#                       outList <- nullbetas
#                   }
#                    return(outList)
#                  })
#                           }

## replace the foreach...
  tmp <- lapply(seq_len(ncol(bootIndexes)),
    FUN = function(x, resids, null, s1,s2,s3,s4,s5,coef, useWeights){
      outList <- apply.gdsn(list(x=resids, y=null), margin=c(2,2), as.is='list',
        FUN = function(X, j, s1, s2, s3, s4, s5, useWeights, coef){
          # create null model
          matstar <- X$y + X$x[j]
          # compute estimate
          nullbetas <- backsolve(s1, crossprod(s2, matstar))[coef]
          if(useWeights) {
            # compute sigma
            sigma <- sqrt(sum((s4%*%matstar)^2)/s5)
            outList <- list(coef = nullbetas, sigma = sigma)
          } else {
            outList <- nullbetas
          }
          return(outList)
        }, j = bootIndexes[,x],
           s1 = s1,
           s2 = s2,
           s3 = s3,
           s4 = s4,
           s5 = s5,
           useWeights = useWeights,
           coef = coef)
         if(useWeights) return(list(coef = sapply(outList, '[[', 'coef'),
                                    sigma = sapply(outList, '[[', 'sigma')))
         else return(unlist(outList))
       }, resids = index.gdsn(boots, 'resids'),
```

```
          null = index.gdsn(boots, 'null'),
          s1 = qr.R(qr.X),
          s2 = qr.Q(qr.X),
          useWeights = useWeights,
          coef = coef,
          s3 = tcrossprod(qr.Q(qr.X)),
          s4 = t(diag(nrow(design))-tcrossprod(qr.Q(qr.X))),
          s5 =  (nrow(design) - qr.X$rank)
  )
## Done
if (useWeights && smooth) { # Here...
  bootRawBeta <- do.call(Map, c(cbind, tmp))$coef # or sapply(tmp, '[[' ,'coef')
  weights <- do.call(Map, c(cbind, tmp))$sigma
  } else {
  ##
  bootRawBeta <- sapply(tmp, '[[' ,'coef')
  weights <- NULL
  }
  NullBeta<-bootRawBeta
  rm(tmp)
  rm(bootRawBeta)
  ##
  closefn.gds(boots)
  unlink(boots[[1]])
}
```

In summary: I replace the `foreach` with an lapply and optimised the bootstraps with a funky `apply.gdsn`. So there is alot to go though.

Once more I will go through some interesting features:

```
      boots <- createfn.gds('bs.gds', allow.duplicate = TRUE)
      res <- add.gdsn(node = boots, name='resids', val = NULL,
         storage = 'float64',
         valdim = c(p,0)
      )
      null <- add.gdsn(node = boots, name='null', val = NULL,
         storage = 'float64',
         valdim = c(p,0)
      )
      apply.gdsn(node = mat, margin = 1, as.is = 'gdsnode',
         target.node = list(x=res, y=null),
         FUN = function(x, s1, s2, n1){
             res <- t(s1 %*% x)/s2
             null <- t(n1 %*% x)
             list(x=res, y=null)
         },
         s1 = t(diag(nrow(design)) - tcrossprod(qr.Q(qr.X))),
         s2 = sqrt(1-h), n1 = tcrossprod(qr.Q(qr.X0))
      )
```

In this chunk we create a new gds file to store some values in, since the output of these `%*%` is going to generate a matrix the same shape as our input. We use `apply.gdsn` with `as.is = 'gdsnode'` and add `target.node = list(x=res, y=null)`.

What is handy is we can label where each of the data goes in the list output to avoid confusion. Inside the

`apply.gdsn` we compute both the scaled residuals and the null model estimate in a row-wise manner (this is done in two memory intensive steps in bumphunter: `resids <- t(tcrossprod( diag(nrow(design)) - tcrossprod(qr.Q(qr.X)), mat))` and `null <- t(tcrossprod(tcrossprod(qr.Q(qr.X0)), mat))`) but in bigmelon we take it nice and slow. Again like above we provide non-trivial computations (again do not take up much memory) as arguments to the `apply.gdsn` to avoid having to compute the same thing hundreds of thousands of times.

The next chunk is where things get interesting...

```
tmp <- lapply(seq_len(ncol(bootIndexes)),
  FUN = function(x, resids, null, s1,s2,s3,s4,s5,coef, useWeights)
    outList <- apply.gdsn(list(x=resids, y=null), margin=c(2,2), as.is='list',
      FUN = function(X, j, s1, s2, s3, s4, s5, useWeights, coef){
        # create null model
        matstar <- X$y + X$x[j]
        # compute estimate
        nullbetas <- backsolve(s1, crossprod(s2, matstar))[coef]
        if(useWeights) {
          # compute sigma
          sigma <- sqrt(sum((s4%*%matstar)^2)/s5)
          outList <- list(coef = nullbetas, sigma = sigma)
        } else {
          outList <- nullbetas
        }
        return(outList)
      }, j = bootIndexes[,x],
         s1 = s1,
         s2 = s2,
         s3 = s3,
         s4 = s4,
         s5 = s5,
         useWeights = useWeights,
         coef = coef)
       if(useWeights) return(list(coef = sapply(outList, '[[', 'coef'),
                                  sigma = sapply(outList, '[[', 'sigma')))
       else return(unlist(outList))
    }, resids = index.gdsn(boots, 'resids'),
       null = index.gdsn(boots, 'null'),
       s1 = qr.R(qr.X),
       s2 = qr.Q(qr.X),
       useWeights = useWeights,
       coef = coef,
       s3 = tcrossprod(qr.Q(qr.X)),
       s4 = t(diag(nrow(design))-tcrossprod(qr.Q(qr.X))),
       s5 =  (nrow(design) - qr.X$rank)
)
```

We remove the foreach in the original function and replace it with an `lapply` to iterate of the bootstraps. Then within each bootstrap we call `apply.gdsn` and compute the null model and get the beta estimates. Similar to `.getEstimate2` we supply non-trivial computations to the functions as stored variables to avoid computing them many times.

I would like to draw interest to this line: `apply.gdsn(list(x=resids, y=null), margin=c(2,2),...` as it demonstrates one of the most impressive functionalities of `apply.gdsn`, similar to `mapply` where you give multiple variables to be looped over in a matrix we can likely specify more than one object to the first argument in `apply.gdsn` in this case we supply two `gdsn.class` nodes within a list under the names `x` and

`y`. This translates to the ability to call either of the two objects within the applied function by calling `X$x` and `X$y` respectively. This makes it impressively easy to write memory efficient functions that require more than one large matrix and is one of the biggest advantages to using `apply.gdsn` over other ways of looping.

The rest of the code remains unchanged from the original function.

Testing the finished product

In most scenarios you will be able to write some code that is able to reproduce the results you want without using a large amount of memory - this would be in particularly useful if you intend on doing analysis that is not possible on the large scale. The trade of is a considerable amount of speed though.

```
set.seed(1)
pos <- sample(1:100000, 485577, rep=T)
set.seed(2)
chr <- sample(as.character(1:22), 485577, rep = T)

out <- bumphunterEngine(betas(gfile)[,], design=model.matrix(~c(1,1,1,2,2,2)),
                        chr=chr, pos=pos, nullMethod = 'bootstrap',
                        B = 3, coef = 2, verbose = T, pickCutoff = T)
```

```
## [bumphunterEngine] Using a single core (backend: doSEQ, version: 1.4.4).

## [bumphunterEngine] Computing coefficients.

## [bumphunterEngine] Performing 3 bootstraps.

## Loading required package: rngtools

## Loading required package: pkgmaker

## Loading required package: registry

##
## Attaching package: 'pkgmaker'

## The following object is masked from 'package:S4Vectors':
##
##     new2

## The following object is masked from 'package:base':
##
##     isNamespaceLoaded

## [bumphunterEngine] Computing marginal bootstrap p-values.

## [bumphunterEngine] cutoff: 0.181

## [bumphunterEngine] Finding regions.

## Warning in regionFinder(x = beta, chr = chr, pos = pos, cluster =
## cluster, : NAs found and removed. ind changed.

## [bumphunterEngine] Found 20838 bumps.

## [bumphunterEngine] Computing regions for each bootstrap.

## Warning in FUN(newX[, i], ...): NAs found and removed. ind changed.

## Warning in FUN(newX[, i], ...): NAs found and removed. ind changed.

## Warning in FUN(newX[, i], ...): NAs found and removed. ind changed.
```

```
## [bumphunterEngine] Estimating p-values and FWER.
```

```
out <- bumphunterEngine.gdsn(betas(gfile), design=model.matrix(~c(1,1,1,2,2,2)),
                             chr=chr, pos=pos, nullMethod = 'bootstrap',
                             B = 3, coef = 2, verbose = T, pickCutoff = T)
```

```
## [bumphunterEngine] Using a single core (backend: doSEQ, version: 1.4.4).

## [bumphunterEngine] Computing coefficients.

## [bumphunterEngine] Performing 3 bootstraps.

## [bumphunterEngine] Computing marginal bootstrap p-values.

## [bumphunterEngine] cutoff: 0.204

## [bumphunterEngine] Finding regions.

## Warning in regionFinder(x = beta, chr = chr, pos = pos, cluster =
## cluster, : NAs found and removed. ind changed.

## [bumphunterEngine] Found 14867 bumps.

## [bumphunterEngine] Computing regions for each bootstrap.

## Warning in FUN(newX[, i], ...): NAs found and removed. ind changed.

## Warning in FUN(newX[, i], ...): NAs found and removed. ind changed.

## Warning in FUN(newX[, i], ...): NAs found and removed. ind changed.

## [bumphunterEngine] Estimating p-values and FWER.
```

Here in this example data-set there is very little difference. However when testing this very function in a dataset of 1,200 EPIC arrays, the bigmelon version I have written uses very little memory (it depends on the number of bootstraps you want to do) the original bumphunter used in excess for 40Gb of memory, however this function takes a considerably longer time than just simply extracting the values and feeding it to the original function. It is likely that the function could be further optimised (through parallelisation) but in the interest of keeping this short we will leave it here.

```
# Closing and deleting gds file for this example
closefn.gds(gfile)
unlink('melon.gds')
```

```
sessionInfo()
```

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.5 LTS
##
## Matrix products: default
## BLAS: /usr/lib/libblas/libblas.so.3.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.0
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
```

```
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] doRNG_1.6.6
##  [2] rngtools_1.2.4
##  [3] pkgmaker_0.22
##  [4] registry_0.5
##  [5] minfiData_0.24.0
##  [6] IlluminaHumanMethylation450kmanifest_0.4.0
##  [7] microbenchmark_1.4-3
##  [8] bigmelon_1.5.7
##  [9] gdsfmt_1.14.1
## [10] wateRmelon_1.25.1
## [11] illuminaio_0.20.0
## [12] IlluminaHumanMethylation450kanno.ilmn12.hg19_0.6.0
## [13] ROC_1.54.0
## [14] lumi_2.30.0
## [15] methylumi_2.24.1
## [16] minfi_1.24.0
## [17] bumphunter_1.20.0
## [18] locfit_1.5-9.1
## [19] iterators_1.0.9
## [20] foreach_1.4.4
## [21] Biostrings_2.46.0
## [22] XVector_0.18.0
## [23] SummarizedExperiment_1.8.1
## [24] DelayedArray_0.4.1
## [25] FDb.InfiniumMethylation.hg19_2.2.0
## [26] org.Hs.eg.db_3.5.0
## [27] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
## [28] GenomicFeatures_1.30.0
## [29] AnnotationDbi_1.40.0
## [30] GenomicRanges_1.30.1
## [31] GenomeInfoDb_1.14.0
## [32] IRanges_2.12.0
## [33] S4Vectors_0.16.0
## [34] ggplot2_2.2.1
## [35] reshape2_1.4.3
## [36] scales_0.5.0
## [37] matrixStats_0.53.0
## [38] limma_3.34.8
## [39] Biobase_2.38.0
## [40] BiocGenerics_0.24.0
##
## loaded via a namespace (and not attached):
##  [1] TH.data_1.0-8           colorspace_1.3-2
##  [3] siggenes_1.52.0         mclust_5.4
##  [5] rprojroot_1.3-2         base64_2.0
##  [7] affyio_1.48.0           bit64_0.9-7
##  [9] mvtnorm_1.0-6           xml2_1.2.0
## [11] codetools_0.2-15        splines_3.4.3
```

```
## [13] knitr_1.18                Rsamtools_1.30.0
## [15] annotate_1.56.1           readr_1.1.1
## [17] compiler_3.4.3            httr_1.3.1
## [19] backports_1.1.2           assertthat_0.2.0
## [21] Matrix_1.2-12             lazyeval_0.2.1
## [23] htmltools_0.3.6           prettyunits_1.0.2
## [25] tools_3.4.3               bindrcpp_0.2
## [27] gtable_0.2.0              glue_1.2.0
## [29] GenomeInfoDbData_1.0.0    affy_1.56.0
## [31] dplyr_0.7.4               Rcpp_0.12.15
## [33] multtest_2.34.0           preprocessCore_1.40.0
## [35] nlme_3.1-131              rtracklayer_1.38.2
## [37] stringr_1.2.0             XML_3.98-1.9
## [39] beanplot_1.2              nleqslv_3.3.1
## [41] zoo_1.8-1                 zlibbioc_1.24.0
## [43] MASS_7.3-48               BiocInstaller_1.28.0
## [45] hms_0.4.0                 sandwich_2.4-0
## [47] GEOquery_2.46.14          RColorBrewer_1.1-2
## [49] yaml_2.1.16               memoise_1.1.0
## [51] biomaRt_2.34.1            reshape_0.8.7
## [53] stringi_1.1.6             RSQLite_2.0
## [55] genefilter_1.60.0         RMySQL_0.10.13
## [57] BiocParallel_1.12.0       rlang_0.1.6
## [59] pkgconfig_2.0.1           bitops_1.0-6
## [61] nor1mix_1.2-3             evaluate_0.10.1
## [63] lattice_0.20-35           purrr_0.2.4
## [65] bindr_0.1                 GenomicAlignments_1.14.1
## [67] bit_1.1-12                plyr_1.8.4
## [69] magrittr_1.5              R6_2.2.2
## [71] multcomp_1.4-8            DBI_0.7
## [73] pillar_1.1.0              mgcv_1.8-23
## [75] survival_2.41-3           RCurl_1.95-4.10
## [77] tibble_1.4.2              KernSmooth_2.23-15
## [79] rmarkdown_1.8             progress_1.1.2
## [81] grid_3.4.3                data.table_1.10.4-3
## [83] blob_1.1.0                digest_0.6.14
## [85] xtable_1.8-2              tidyr_0.8.0
## [87] openssl_0.9.9             munsell_0.4.3
## [89] quadprog_1.5-5
```