# Methods for estimating changes in DNA methylation in the presence of cell type heterogeneity

Kevin McGregor

Master of Science

Department of Epidemiology, Biostatistics, and Occupational Health

McGill University

Montréal, Québec

2015-04-14

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

## ACKNOWLEDGEMENTS

I would first like to express my gratitude towards my co-supervisors Dr. Aurélie Labbe and Dr. Celia Greenwood. The completion of this thesis would not have been possible without their hard work and sound advice. They have been great mentors and have truly succeeded in making me feel at home at McGill. I also thank the other professors and staff in the Epi/Biostat department for creating a very student-friendly environment.

I would like to thank Dr. Marie Hudson for the opportunity to work on the unique data set which is featured in this thesis as well as her colleagues Dr. Tomi Pastinen, Dr. Sasha Bernatsky, and Dr. Ines Colmegna. I would also like to acknowledge this work was funded by the CIHR Catalyst Grant: Environments, Genes, and Chronic Disease as well as the Lady Davis Clinical Research Pilot Project Grant.

I also acknowledge my Master's funding as coming from the grant CIHR MOP-300545, with co-principal investigators Dr. Celia Greenwood, Dr. Aurélie Labbe, and co-investigators Dr. Antonio Ciampi, Dr. David Stephens, Dr. Tomi Pastinen, and Dr. Guillaume Bourque.

I would like to thank Dr. Kathleen Klein Oros, Dr. Stepan Grinek, Gregory Voisin, and Dr. Vince Forgetta for their advice and technical support. They were always there to answer my questions and share relevant code, and their expertise has undoubtedly saved me many hours of work. I thank Dr. James Wagner for sharing the code for his adjustment method as well as for pointing me towards helpful literature. I thank Maxime Turgeon and Sahir Bhatnagar for their friendship and support. I

also thank Maxime for helping me (i.e. doing most of it himself) translate my abstract into French.

Finally, I thank my friends and family for always being there for me.

## ABSTRACT

DNA methylation occurring at a cytosine-guanine (CpG) site blocks binding to the DNA and hence can influence gene function and regulation. Therefore, it is often valuable to investigate which methylation sites are associated with diseases or other phenotypes of interest. Though a large proportion of CpG sites in mammals are methylated, methylation signatures differ notably between cell types. Consequently, when measuring methylation levels on whole blood or other types of tissues involving multiple cell types, it can be difficult to distinguish the changes associated with a phenotype of interest from those occurring as a result of varying proportions of different cell types among subjects. This phenomenon is of concern when changes in cell type proportion are associated with the phenotype itself, thereby making cell type proportion a confounder. There are several recently developed methods that attempt to correct for this confounding, including one method based on an external validation data set (Houseman et al., BMC Bioinformatics 2012), a reference-free method (Houseman et al., Bioinformatics 2014), Surrogate Variable Analysis (Leek and Storey, PLoS Genetics 2007), Independent Surrogate Variable Analysis (Teschendorff, Bioinformatics 2011), the FAST-LMM-EWASher method (Zou, Nature Methods 2014), Deconfounding (Repsilber, BMC Bioinformatics 2010), and CellCDecon (Wagner, PhD Thesis 2014). In order to compare the performance of each method, we have artificially re-combined measures of methylation obtained from cell-separated analysis of whole blood. Specifically, methylation measures are available for monocytes and CD4 T-cells. We randomly chose a subset of the samples to be disease

cases, then we designated a set of CpG sites to be associated with the disease. A new artificial set of methylation measurements was generated by combining the values from each cell type with variable proportions of each cell type in each subject. We uncovered notable differences between the methods in terms of statistical power, reduction in false discovery rate, the extent to which the confounding has been corrected, and in computational performance. The reference-based method, due to its ease of use and generally good performance, was selected as the best method under the specified circumstances. ISVA was selected as the best alternative if no external data set were available.

# ABRÉGÉ

La méthylation de l'ADN, qui a lieu à un site cytosine-guanine (CpG), empêche la liason et ainsi peut influencer la fonction et la régulation génique. Il y a donc intérêt à investiguer quels sites méthylés sont associés à une maladie ou tout autre phénotype. Malgré qu'une grande proportion de sites CpG soient méthylés chez les mammifères, la structure de la méthylation varie d'un type de cellule à un autre. Par conséquence, lorsque les niveaux de méthylation sont mesurés à partir d'un échantillon sanguin ou un autre tissu hétérogène, il peut être difficile de séparer les changements associés au phénotype étudié de ceux relevant de différents niveaux d'hétérogénéité parmis les sujets. Ce phénomène est particulièrement important lorsque le phénotype est associé avec cette hétérogénéité, créant ainsi un facteur confondant. Plusieurs méthodes ont récemment été dévelopées pour tenter de corriger ce phénomène, incluant une méthode utilisant un jeu de données externe (Houseman et al., BMC Bioinformatics 2012), une méthode sans jeu de données externe (Houseman et al., Bioinformatics 2014), Analyse de variables latentes (Leek and Storey, PLoS Genetics 2007), Analyse de variables latentes indépendantes (Teschendorff, Bioinformatics 2011), la méthode "FaST-LMM-EWASher" (Zou, Nature Methods 2014), "Deconfounding" (Repsilber, BMC Bioinformatics 2010), et "CellCDecon" (Wagner, Thèse de doctorat 2014). Afin de comparer la performance de chaque méthode, nous avons artificiellement recombinés les mesures de méthylation obtenues à partir d'un échantillon sanguin pour lequel chaque type de cellule a été analysé séparément. Plus précisément, les mesures de méthylation sont disponibles pour les monocytes et les

lymphocytes T CD4. Nous avons aléatoirement désigné une portion des échantillons comme étant affectés par, ainsi qu'une portion des sites CpG comme étant associés à, la maladie. Un nouveau jeu de données synthétique contenant des mesures de méthylation a été généré en combinant les valeurs de chaque type de cellule en proportions variables, et ce, pour chaque sujet. Nous avons découvert d'importantes différences entre les méthodes en ce qui a trait à la puissance, la réduction du taux de fausses découvertes, la capacité de corriger les biais dus à l'hétérogénéité, et la performance computationelle. La méthode basée sur un jeu de données externe a été sélectionnée comme étant la meilleure méthode, grâce à sa simplicité et sa bonne performance générale. L'analyse de variables latentes indépendantes a été choisie comme meilleure alternative lorsqu'aucun jeu de données externe n'est disponible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER 1
# Introduction

The word 'epigenetics' encompasses changes in the functionality of DNA (i.e. gene expression) not explained by DNA sequences themselves, in particular, changes procured by phenomena such as DNA methylation, RNA-silencing, and histone modification [11]. The problem addressed in this thesis is specifically concerned with DNA methylation. In recent years understanding the role of epigenetics has become vital in disease etiology at the genetic level. There is, in fact, a subset of diseases referred to as 'epigenetic diseases' that are primarily caused by the improper regulation of genes [11]. Similarly, deviant hypermethylation has been shown to fundamentally affect tumor development in humans [21]. Diseases such as Beckwith-Wiedemann syndrome, Prader-Willi syndrome, and Angelman syndrome, among others, are also associated with aberrant methylation, clearly demonstrating the importance of methylation in proper gene regulation [32]. It is therefore of great interest to study how changes in methylation can associate with diseases or other interesting phenotypes.

One study design allowing such an association to be detected is the Epigenome-Wide Association Study (EWAS). An EWAS is, at the most basic level, a regression model testing for association between methylation levels and a phenotype. Though one could conceivably run an EWAS analysis on methylation measurements recovered from any kind of cell type, it is essential that the same kinds of cells be sampled among all subjects, as methylation profiles differ notably between cell types. Herein lies

the crux of the problem addressed in this thesis: when drawing samples containing multiple cell types (e.g. whole blood or adipose tissue), between-sample variability in cell type composition can act as a confounder. This occurs when the disease being studied is associated with changes in cell type composition, something that is known to be true in blood samples coming from cancer patients [43] as well as in synovial and cartilage cells in patients affected by rheumatoid arthritis [34].

The advancement of high-throughput technologies in recent years has enabled the development of a number of methods that can account for this unmeasured confounding. Though there are numerous similarities between the approaches, there remain some fundamental differences in terms of limitations and performance. Ultimately, it would be valuable to do a careful review of how these methods perform relative to each other in terms of adjustment, statistical power, as well as computing performance under controlled conditions. In this thesis I will undertake such a review.

The methods will be compared through the use of a rare data set containing methylation measurements on separated blood cell types. This data set comes from the research of Dr. Marie Hudson at the Lady Davis Institute in Montréal, QC. The possession of this kind of data is a privilege, as separating cell types is not a trivial task. A more detailed description of the data can be found in Section 4.1.

The fundamental idea of this work is to use the measurements from the separated cell types in a cell mixture simulation. We can specify, for each cell type, associations between a phenotype and various methylation sites. Then, as one would often see in real cell mixture data, a change in the relative proportion of cell types in each

sample can be induced between cases and controls. This allows us to set simulation parameters that each cell type adjustment method will try to estimate, while simultaneously preserving the natural variation coming from the methylation data across individuals and across sites in the genome. Because the associations are prespecified, calculating power and detecting type I errors will be relatively straightforward. Furthermore, a range of different situations can be tested to see if the performance of methods is significantly affected (i.e. changing simulation parameters).

Additionally, I will examine performance on a data set of mixed cell types (blood) in order to compare results in a study with no simulation involved. These data are from the Assessment of Risk in Colorectal Tumors in Canada (ARCTIC) study. Details of this data set can be found in Section 5.2.

The thesis will be structured as follows: Chapter 2 will provide some background about DNA methylation and the different challenges researchers face when working with this kind of data, and will more concretely describe the problem of cell type heterogeneity in epigenomic studies. Chapter 3 will describe each of the cell type adjustment methods, and will function as the literature review. Chapter 4 will provide a description of the separated blood cell data set, simulation details, and define the performance metrics that will be used for method comparison. Chapter 5 will provide the results of the simulation as well as results of analysis of the ARCTIC data set. Finally, Chapter 6 will provide discussion and conclusions, and will outline possible directions for future work.

In order to properly describe the work done in this thesis, it will be necessary to delve into the details of DNA methylation and how measurements on methylation,

in conjunction with one or more phenotypes of interest, will be used to formulate an EWAS. This is addressed in the next chapter.

# CHAPTER 2
# DNA Methylation and the Epigenome-Wide Association Study

The purpose of Chapter 2 is to provide a simple description of DNA methylation which is the principal focus of the methods presented in this thesis. In doing so it is imperative to provide details of how methylation is measured, underscore issues with the quality of methylation data, and specify how to deal with such issues. Finally, the EWAS will be more precisely defined, and in due course, the underlying problem of cell heterogeneity will emerge.

## 2.1 What is DNA Methylation?

The purpose of DNA methylation is to block the binding of transcription factors to the DNA thereby acting as a regulator of DNA activity, and therefore, gene expression. This phenomenon occurs when there is a transfer of a methyl group to the fifth carbon on a cytosine nucleotide, which occurs almost always when the cytosine is succeeded by a guanine nucleotide [28]. Sites along the genome housing a cytosine-guanine dinucleotide are referred to as 'CpG' sites and measurements taken on a subset of these sites will form the basis of the EWAS. There are approximately 7.4 million CpG sites in the human genome that could be methylated or unmethylated [9].

CpG sites in vertebrates are largely methylated, and most of the unmethylated sites are found in what are referred to as 'CpG islands', regions in the genome containing a high density of CpGs [6]. Humans have been shown to possess more than

25,000 CpG islands [8]. Naturally, it is imperative to ensure methylation analyses accommodate a good coverage of CpG sites within islands.

## 2.2 Measuring Methylation

One platform from which methylation measures can be easily obtained is the Infinium HumanMethylation450 BeadChip. The 450K array features probes corresponding to 482,421 CpG sites as well as over eight hundred negative control probes that allow assessment of the quality of the signals [39]. CpGs from all chromosomes are included on the array and about 30% are located in CpG islands [33]. The CpG sites on the array were chosen to include a high proportion of CpG islands (96%) and RefSeq sites (99%), among other criteria [4].

The underlying process of the Illumina method involves performing a bisulfite conversion on the DNA which is manifested by a change from a cytosine to a uracil. Methylation can then be inferred at a specific CpG site if the cytosine remains untouched by the bisulfite reaction [5]. Among the 482,421 probes spanning the genome, two types of probes exist: Infinium I which contains both a methylated and an unmethylated probe for every CpG site, and Infinium II which contains only one probe for both the methylated and unmethylated signals [4]. Though the signals from either probe type can be used to calculate easily interpretable statistics measuring methylation, there does exist a probe type bias that must be accounted for in the normalization step, which is outlined in Section 2.4.

DNA extracted from any type of human tissue can be used to measure methylation on the Illumina platform, however, blood is the most common and is relatively simple to obtain. Methylation can also be estimated for tissues that are more difficult

to sample. In some cases, calibration models have been developed to use samples taken from tissues that are easier to obtain as surrogate measures for tissues that are hard to obtain. For example, measurements taken on peripheral blood leukocytes have been used to predict profiles for artery and atrium tissue [27].

It is also well known that methylation patterns can differ with respect to factors such as age and sex. For example, models using methylation information at selected sites in the genome have been shown to be fairly accurate age predictors in a variety of cell types [16, 42]. Also, substantial differences in methylation between the sexes exists even on the autosomes, and hence sex should be taken into account in methylation analyses.

### 2.2.1 Statistics calculated from methylation data

Measuring methylation within a cell, due to the presence of two homologous chromosomes, results in one of three outcomes: the cell could have zero, one, or two methylated signals at a CpG site. However, in practice a tissue sample contains a very large number of cells which necessitates the need for a summary measure that estimates the global level of methylation within a collection of cells. Here we define two statistics, $\beta$ and $M$, to summarize the extent to which a given CpG site is methylated in a sample. For these two definitions, we use a slightly modified version of the notation found in [10]. Firstly, we examine the more common of the two: the 'beta' value.

**Definition 2.2.1** *Let $y_{i,m}$ and $y_{i,u}$ be the methylated and unmethylated intensities at CpG site $i$, respectively. Then the 'beta' value, i.e. the ratio of methylated and overall intensities, is defined as:*

$$\beta_i = \frac{\max(y_{i,m}, 0)}{\max(y_{i,m}, 0) + \max(y_{i,u}, 0) + \tau} \tag{2.1}$$

*where $\tau \geq 0$ is an offset term to regularize in the event that the overall signal is low.*

After normalization for technical artifacts (details of which will be covered in Section 2.4), the intensities $y_{i,m}$ and $y_{i,u}$ could potentially be negative. By this definition such values are set to zero. The term $\tau$ in the denominator is a regularization term which, on the Illumina platform, is set to 100 by default.

From this definition it is obvious that beta will lie between 0 and 1, allowing it to retain a nice interpretation. It can be thought of as an estimator for the proportion of cells in the sample that are methylated at a given CpG site. The other statistic that is common in the literature, but less so than the $\beta$ value, is the $M$ value, which is defined as follows:

**Definition 2.2.2** *Using $y_{i,m}$ and $y_{i,u}$ from Definition 2.2.1, the M-value for CpG site $i$ is defined to be:*

$$M_i = \log_2 \left( \frac{\max(y_{i,m}, 0) + \tau}{\max(y_{i,u}, 0) + \tau} \right) \tag{2.2}$$

*where $\tau \geq 0$ is a regularization term.*

Du et al. note in [10] that the M-value is "more statistically valid for the differential analysis of methylation levels", as it has been observed to give a more

stabilized variance over technical replicates. Even so, a few of the methods that will be introduced in Chapter 3 specifically require the use of the $\beta$ value. This fact, combined with its ease of interpretability, means the $\beta$ value will be the *de facto* measurement of methylation within the context of this thesis.

It turns out that methylation estimates are quite reliable. The Illumina array was used to analyze technical replicates, and was shown to give highly reproducible results for normal lung tissue and lung adenocarcinoma tissue [5]. There remain, however, a number of potential technical artifacts that can arise and must be accounted for. The next two sections will outline issues in data quality one might encounter in the analysis of methylation data, and will mention some of the tools used to remedy them.

### 2.3 Data Quality Issues

A number of data quality issues exist within the realm of methylation data as one would expect when extracting measurements from high-throughput genomic data. These kinds of problems mostly entail systematic differences in methylation patterns resulting from different technical artifacts that exist in Illumina's chip design, but can also include unexplained noise possibly due to laboratory environment, technician precision, or random error.

Probe type is one source of variation in the data. It has been shown that different distributions of signal intensity can be observed between the type I and type II probes. Two possible ways of dealing with this include: transforming the observations

from one probe type to match the distribution of the other, or normalizing the data from each probe type separately [29].

There is also a significant amount of background noise observed in methylation data from the 450K array. The array's design includes a set of negative control probes that are not expected to hybridize. Readings from these control probes can be used in statistical models to perform adjustment for background noise [44]. The control probes can also be used to estimate probe type bias as mentioned above.

One of the biggest challenges in EWAS is the potential for batch effects. The Illumina BeadChip is designed so that individual chips are capable of holding up to 12 samples, with up to 8 chips per plate. Additionally, chips are manufactured in batches of several hundred at a time. Multiple chips/plates will usually be required to acquire measurements for all samples in the study, and multiple batches may also be needed. Environmental factors such as changes in laboratory conditions and time of experiment can lead to notable differences in distributions between these batches [37]. It is imperative that any kind of normalization technique take into account which samples came from the same batch in order to attempt to rectify the problem. Batch effects also create the need for prudence in study design. If, for example, all disease cases were included on a chip and controls on another, it would be difficult to normalize for batch effect without erasing the effects due to case/control status.

## 2.4 Normalization

Normalization attempts to correct artifactual or technical biases present in the methylation data. Many normalization algorithms have been posited in the last few years. One common method is Quantile Normalization (QN). This technique

involves a non-linear transformation that forces uniformity in distribution over all methylation loci [37]. It can be seen in the literature that several other methods are build upon QN. The method used for the methylation data included in this work is called 'Functional Normalization', and was presented in 2014 by Fortin et al. [13], the details of which are found in the next subsection.

### 2.4.1 Functional Normalization

Functional Normalization (FN) is a normalization technique that takes a known covariate or set of covariates and removes only the (unwanted) variation associated with that covariate. This is advantageous as the naïve version of QN, which does not take covariates into account, runs the risk of eliminating the true biological variation in the data. FN does not require the user to input information on specific experimental information (i.e. batch number). It is demonstrated in [13] that the first two principal components from the 450K control probes can act as a proxy for chip and position biases. Similarly, FN can be used on any kind of genomic data containing some kind of control probes.

First, FN calculates the empirical cumulative distribution function for each of the samples (in our case each sample contains $m$ methylation measurements). The resulting quantile vectors are used to form regression models with the empirical quantile function as a response, and the covariates as predictors (separately by probe type). The variation resulting from the covariates is then subtracted off the original set of methylation vectors. Any subsequent analysis (i.e. an EWAS) is then performed on these adjusted vectors.

It was shown that FN is able to outperform many of the other current normalization techniques. For example, the authors observed better comparability between experimental replicates, even in the presence of batch effect. It was also able to improve prediction of differentially methylated positions on the sex chromosomes. Finally, it was able to reduce variability between technical replicates in a single study. Motivated by these advantages, FN is performed on all methylation data before any of the simulation or cell type adjustment occurs.

### 2.5 Epigenome-Wide Association Studies

The Epigenome-Wide Association Study (EWAS) is a natural extension to the well-known Genome-Wide Association Study (GWAS). The GWAS tested for associations between a phnotype and a series of single nucleotide polymorphisms across the genome, wheras the EWAS implies a series of statistical tests that attempt to find associations between a phenotype and each of the CpG sites on the array. In this thesis, such associations are tested for through the use of regression models. It is important to note that though the model is framed in a way that there are predictor and response variables, we are making no claim as to the direction of causality when an association is found. Furthermore, the nature of the EWAS is exploratory; that is, any observed association should be taken as a stepping stone for further investigation, rather than absolute proof of a biological link.

The EWAS in this thesis will be designed in a somewhat peculiar way: the methylation measurements will be used as the response variable, and the phenotype of interest will be a predictor (which further stresses why no causal link is implied). One can also include other covariates such as age, sex, etc. that must be accounted

for in the context of the study. There is no specific form that the regression model for the EWAS must take, however, most of today's literature specifies the model to be a linear combination of the phenotype and covariates. The regression model used here will do the same, specifically because it is a necessary condition for the use of some of the adjustment methods. A directed, acyclic graph illustrating the assumed relation can be seen in Figure 2–1.

**Definition 2.5.1** *Let* $\boldsymbol{Y}$ *be an* $m \times n$ *matrix (CpG sites* $\times$ *Subjects), and let* $\boldsymbol{X}$ *be a matrix containing a column of ones, followed by a column corresponding to the phenotype of interest, followed by columns corresponding to other covariates. Then the regression model for the EWAS is defined to be:*

$$\boldsymbol{Y} \;=\; \boldsymbol{B}\boldsymbol{X}^{\top} + \boldsymbol{E} \tag{2.3}$$

*where* $\boldsymbol{B}$ *is the resulting matrix of associations between the CpG sites and the columns of* $\boldsymbol{X}$ *and* $\boldsymbol{E}$ *is a matrix of errors.*

This is, essentially, a set of $m$ separate linear regression models, where $m$ is the number of CpG sites in consideration. Consequently, an assumption of normality exists for each row of $\mathbf{E}$. Each of the regression models implies a statistical test; specifically, for CpG site $i \in 1, \ldots, m$ the corresponding null hypothesis is $H_0 : \mathbf{B}_{ij} = 0$, and the alternative hypothesis is $H_a : \mathbf{B}_{ij} \neq 0$, where $\mathbf{B}_{ij}$ is the $i^{th}$ entry of the $j^{th}$ column of $\mathbf{B}$ which corresponds to the phenotype of interest.

Due to the large number of tests being done (one for each CpG), it is important to make some kind of correction for multiple testing. Given 480,000 CpG sites, one

would expect about 24,000 false positives at significance level $\alpha = 0.05$ (assuming independence between tests), among which it would be very difficult to discern any true associations.

### 2.6 Cell Type Heterogenity as an Unmeasured Confounder

The main problem addressed in this work arises from the fact that methylation profiles can differ between cell types—enough to be able to accurately identify cell types in homogeneous samples solely based on methylation [2]. Figure 2–2 nicely illustrates the idea. Plotting a heat map for three different cell types (Monocytes, T-Cells, B-Cells) shows how drastically the methylation profile can change over cell type.

From the perspective of an EWAS, cell type heterogeneity can complicate things. If, for example, the phenotype studied were a disease such as a malignant tumor, part of the observed immune response would involve an increase in proportion of regulatory T-cells in blood [35]. Therefore, one would observe variability of cell type proportions in whole blood samples between cases and controls in addition to any between-subject variability that already exists within these groups. Hence, cell type composition can be a confounder when studying methylation on samples containing multiple cell types, i.e. whole blood or adipose tissue.

A confounding variable can usually be dealt with by inclusion in the regression model. Of course, if we had some kind of direct measure of cell type composition for each subject we would easily be able to perform such an adjustment. In most cases, however, no direct measure of cell type composition is available. It will be necessary to estimate the effects of confounding through other means. We can do so

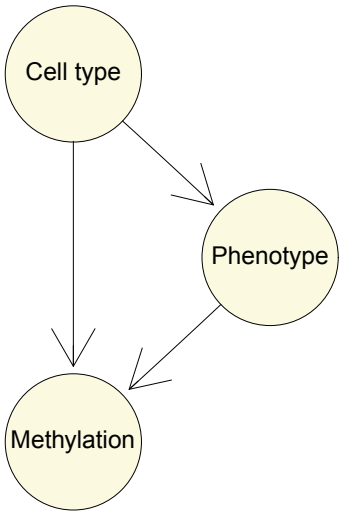


Figure 2–1: Directed acyclic graph for the EWAS regression model including cell type as a confounding variable. The model is most concerned with estimating the phenotype-methylation link. Though the model implies the direction of the arrow is from phenotype to methylation, the result of the EWAS makes no claim as to the direction of causality.

by using different methods that uncover hidden patterns in large scale, genome-wide, methylation data that can be used to construct a set of latent variables. Including these variables in the regression model will, ideally, rectify the problem at hand.

One ultimate question remains: how is it possible to use observed methylation data to capture information about cell type heterogeneity in order to form these latent structures? The next chapter will introduce a number of existing methods that address this concern.

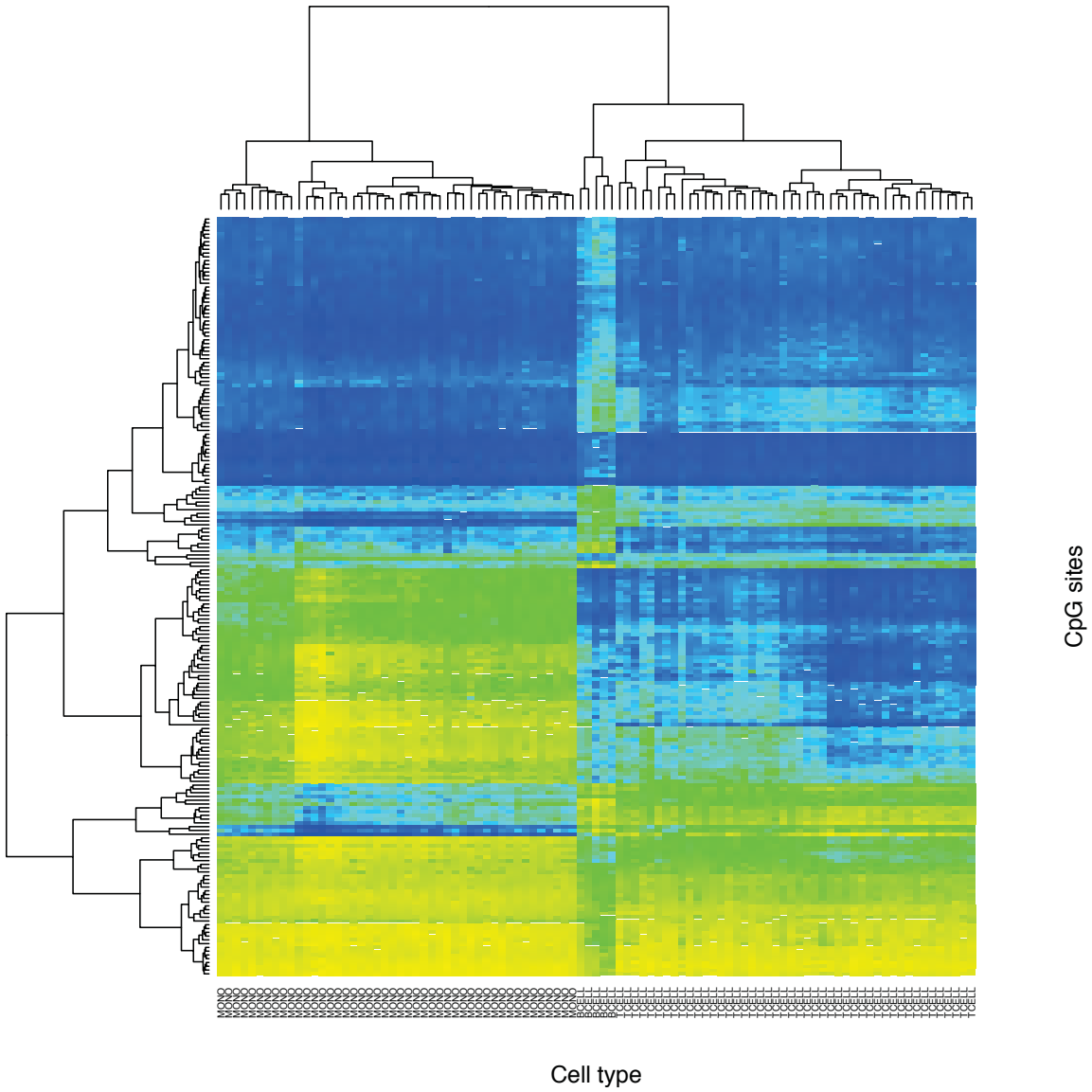


Figure 2–2: A heat map showing differences in methylation profiles at 200 CpG sites for different cell types on the Marie Hudson data set (see Section 4.1). The columns represent samples of different cell types and the rows represent the 200 CpG sites. Yellow means unmethylated, green means partially methylated, and blue means completely methylated. Dendograms show clustering, which correctly identified the differences due to cell type. Monocytes are found on the left side, B-cells are in the middle, and T-cells are on the right.

# CHAPTER 3
# Description of Adjustment Methods

Having developed a more intimate understanding of DNA methylation and the problem at hand, we now focus on providing an overview of some of the currently available methods attempting to correct for confounding by cell heterogeneity. This chapter can be thought of as a literature review, and will also provide some descriptions of the supporting mathematical and statistical techniques these methods employ. Though their performances will not be compared until Chapter 5, some advantages and disadvantages of each will naturally arise from the details.

The first two methods come from the same first author, Andres Houseman, and the more recent of the two is more generally applicable as it does not assume any kind of cellular composition. The methods SVA and ISVA are very similar conceptually, but the former applies singular value decomposition, and the latter, independent component analysis. The FaST-LMM-EWASher method is somewhat unique in its construction, and uses only a subset of the data. The Deconfounding method uses a decomposition similar to that of independent component analysis. Finally, the method 'CellCDecon' is a very new method formulated by James Wagner in his PhD thesis, which tries to explicitly estimate the cell type composition, something the other algorithms (aside from reference-based) do not pursue.

It should be noted that some methods are also applicable to data other than DNA methylation. SVA, ISVA, and Deconfounding, and CellCDecon are capable of

Table 3–1: List of adjustment methods

| Method | Abbreviation | First Author | Year |
|---|---|---|---|
| Reference-based | Ref-based | Andres Houseman | 2012 |
| Reference-free | Ref-free | Andres Houseman | 2014 |
| Surrogate Variable Analysis | SVA | Jeffrey Leek | 2007 |
| Independent Surrogate Variable Analysis | ISVA | Andrew Teschendorff | 2011 |
| Factored Spectrally Transformed Linear Mixed Model 'EWASher' | FaST-LMM-EWASher | James Zou | 2014 |
| Deconfounding | Deconf | Dirk Repsilber | 2010 |
| Cell Composition Deconvolution | CellCDecon | James Wagner | 2014 |

adjusting any kind of high-throughput data, as long as they can be summarized in the correct format. Moreover, SVA and ISVA can potentially correct for other kinds of unmeasured confounding: there is no assumption made that the observed genomic inflation is due solely to cell type heterogeneity.

Throughout the chapter, the methods' notations will be unified for the sake of consistency. The measured methylation beta values from the target study will be found in the matrix $\mathbf{Y}_{m\times n}$, where $m$ represents the number of CpG sites in consideration, and $n$ represents the number of samples. The covariates are contained in $p-1$ vectors, which include the phenotype of interest as well as other known confounders. They form the matrix $\mathbf{X}_{n\times p}$, which also contains an intercept column.

### 3.1 Reference-based method

The reference-based method was introduced by Houseman et al. in [17]. The idea here is to use an external data set containing methylation measurements on six separated blood cell types, common in whole blood, in order to 'calibrate' a regression

model. This would be expected to make the connection between methylation and the phenotype more apparent.

### 3.1.1 External data set

The six separated cell types included in the external data set are: Monocytes, CD4T-Cells, CD8T-Cells, B-Cells, Granulocytes, and Natural killer cells. Currently, any use of this method is restricted to blood samples. On this front, the authors leave the door open by providing a parameter in the R package which could house measurements on other separated cell types. Measurements on the separated blood cells were made using the Infinium HumanMethylation27K Beadchip platform, with 27 additional samples of whole blood included from the same subjects to estimate chip effect. The method has also been updated for use on the 450K platform.

### 3.1.2 Regression Model

Let $m$ be the number of CpG sites under consideration and assume $d_0$ cell types for our sample data (In this case $d_0 = 6$). Consider initially the external data on separated cell types. Let the methylation measurements for sample $h$ in this data set be the $m \times 1$ vector $\mathbf{Y}_{0h}$, and let $\mathbf{w}_h$ be a vector of length $d_0$ containing the known proportions of cell types in the sample. Since these are separated cell types, $\mathbf{w}_h$ should have the value one at exactly one entry, and zeros elsewhere.

Next, let $\mathbf{Y}_i$ be a vector of length $m \times 1$, representing the $i^{th}$ column of $\mathbf{Y}$. It contains methylation measurements for subject $i$ in the study data (containing $n$ subjects). The vector $\mathbf{X}_i$ is the $i^{th}$ row of the covariate matrix $\mathbf{X}$, containing an intercept, followed by the level of the phenotype for subject $i$, followed by other covariates to be included in the model for this subject.

Two separate regression models are formulated corresponding to the reference data, and the study data, respectively:

$$\mathbf{Y}_{0h} = \mathbf{B}_0\mathbf{w}_h + \mathbf{e}_{0h} \tag{3.1}$$

$$\mathbf{Y}_i = \mathbf{B}_1\mathbf{X}_i + \mathbf{e}_i, \tag{3.2}$$

where $\mathbf{e}_{0h}$ and $\mathbf{e}_i$ are error terms. The two matrices, $\mathbf{B}_0$ and $\mathbf{B}_1$ contain parameters, including a column corresponding to each intercept. Then these two matrices are assumed to relate to each other in the following way, which will link the two regression models:

$$\mathbf{B}_1 = \mathbf{1}\gamma_0^\top + \mathbf{B}_0\Gamma + \mathbf{U}. \tag{3.3}$$

$\Gamma$ is a $d_0 \times d_1$ matrix containing the estimated associations between $B_0$ and $B_1$ from which the estimated change in cell type compositions for subject $i$'s set of covariates can be extracted as $\Gamma\mathbf{X}_i$. There was no specific dependence structure assumed in the error vectors $\mathbf{e}_{0h}$ and $\mathbf{e}_i$. Therefore, this formulation is useful in the case where the error matrix contains dependence, i.e. if there were a known batch effect that should be included as the random component of a random effects model.

The paper also suggests a more simple (and more convenient) individual level model that should be used if one is only interested in *correcting* for cell type heterogeneity, but not explicitly measuring changes in cell type distribution associated with the covariates. Here, the left hand side of Equation 3.3 is replaced with the methylation measurement for a given subject. This suggests that the cell type inference will not consider the measured phenotype of the target data; rather, it will depend solely on the subject's methylation profile. The model is:

$$\mathbf{Y}_i \quad = \quad \mathbf{B}_0\Gamma^* + \mathbf{U}^*. \tag{3.4}$$

The estimate $\hat{\Gamma}^*$ will contain direct estimates of cell type composition for subject $i$ and can be obtained by projecting $\mathbf{Y}_i$ on the column space of $\mathbf{B}_0$. Houseman et al. also suggest using quadratic programming to ensure the components of $\hat{\Gamma}^*$ remain non-negative. Interestingly, the sum of the proportions for a sample is not constrained to one. Despite the absence of such a restriction, it turns out that the sum of the proportions of each subject is often quite close to one, as seen in the paper, as well as in the data considered in this thesis.

Once the estimated cell proportions are obtained for each individual, adjustment occurs simply by including the estimates as covariates in the usual regression model after having deleted one of the columns (due to linear dependence). One limitation of this method is the assumption of the cell type composition. At this point only a small number of validation data sets exist for this method. Therefore, inference can

only be performed if working on the same kinds of cells in the validation data, or at least a subset of these cells. In 2014, Jaffe and Irizarry extended the use of the reference-based method by providing a blood cell separated data set on the 450K platform [20]. The authors also applied the method on samples of flow sorted cells from the dorsolateral prefrontal cortex obtained from the research of Guintivano et al. [14]. The absence of a wide variety of cell separated data sets led to the development of a more general method which is similarly formulated, but contains no specific cell type assumption.

### 3.2 Reference-free method

The reference-free method was also developed by Houseman et al., in his more recent paper [18]. The motivation for this method was to have an algorithm that would not be constrained to specific cell types. Indeed, the reference-free method can be run on any kind of cell mixture and as its name suggests, there is no reference to an external data set. The algorithm was inspired by Surrogate Variable Analysis, which is introduced in the next section.

To begin, assume we have a matrix $\mathbf{Y}$ of dimension $m \times n$ containing methylation 'beta' values for $m$ CpG sites on $n$ subjects. Let $\mathbf{X}$ be a matrix of covariates; the first column is the intercept, the second is the phenotype, and the remaining columns (optional) contain other covariates. The method starts by fitting an unadjusted model, in other words, a model ignoring cell type composition:

$$\mathbf{Y} \quad = \quad \mathbf{B}^{*}\mathbf{X}^{\top} + \mathbf{E}^{*}. \tag{3.5}$$

The matrices involved in (3.5) undergo a recharacterization: $\mathbf{B}^* = \mathbf{B} + \mathbf{M}\boldsymbol{\Gamma}^\top$ and $\mathbf{E}^* = \mathbf{M}\boldsymbol{\Xi}^\top + \mathbf{E}$, where $\mathbf{M}$ $(n \times K)$ contains average methylation measurements for $K$ cell types, $\boldsymbol{\Gamma}$ contains cell type specific effects, and $\boldsymbol{\Xi}$ contains cell type specific errors. The parameter of interest is $\mathbf{B}$, i.e. the cell type independent effects between the covariates and the CpG sites. It should be noted that the reference-free method does not explicitly estimate the number of cell types present among the mixtures, rather, it attempts to estimate a set of $d$ latent vectors that will adjust for cell type heterogeneity. The latent dimension $d$ is estimated through "Random Matrix Theory" [30].

Next, a singular value decomposition (SVD) is performed on the matrix $(\hat{\mathbf{B}}^*, \hat{\mathbf{E}}^*)$, as outlined in Appendix A. The SVD from this concatenated matrix can be expressed in a factor-analytic form:

$$(\hat{\mathbf{B}}^*, \hat{\mathbf{E}}^*) \quad = \quad \boldsymbol{\Lambda}\hat{\mathbf{U}}^\top + \hat{\boldsymbol{\Phi}}, \tag{3.6}$$

where $\boldsymbol{\Lambda}$ is an $m \times d$ matrix and $\hat{\mathbf{U}}$ is $(p+n) \times d$.

The actual form of the above decomposition will depend on the chosen latent dimension $d$. The columns of $\mathbf{U}$ form the latent variable structure. The estimated associations between the covariate matrix and the methylation sites are obtained by taking the residual from regressing $\hat{\mathbf{B}}^*$ on $\boldsymbol{\Lambda}$. That is,

$$\hat{\mathbf{B}} = \hat{\mathbf{B}}^* - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^\top\hat{\mathbf{B}}^*. \tag{3.7}$$

The standard errors are more difficult to obtain for the reference-free method. As a result of the complicated (and rather improvised) estimation procedure, the authors suggest bootstrapping by sampling from the columns of $\mathbf{E}^*$ to obtain standard errors for $\hat{\mathbf{B}}$. Another limitation of the method is that it specifically stipulates that the methylation measurements be on the beta scale. Logit-transformed data would violate the assumption of linearity in the cell type mixture step.

### 3.3 Surrogate Variable Analysis

Surrogate Variable Analysis (SVA) was introduced in 2007 by Leek and Storey in [24]. Unlike the first two methods, SVA is not explicitly tailored for analysis on methylation data, nor is it particularly concerned with confounding by cell type composition. Rather, it can be run on data from any kind of high-throughput experiment provided the data can be summarized into a matrix in the correct format along with a set of covariates. It attempts to adjust for 'unmeasured' or 'unmodeled' confounders, which do not specifically need to be consequences of variability in cell type composition. Even so, methylation data turns out to be well-suited for analysis via SVA.

Here we slightly modify Leek and Storey's notation in order to remain consistent with the other methods outlined in the thesis. SVA attemps to find a set of mutually orthogonal vectors $\mathbf{h}_k$, $k = 1, \ldots, K$, that span the same linear space as $\mathbf{g}_\ell$, $\ell = 1, \ldots, L$ with $K \leq L$, which represent functions of unmeasured confounders. The vectors $\mathbf{g}_\ell$ are assumed to be additive in the underlying model, and the authors maintain that this assumption is quite general. In our case these vectors will contain

the unobserved cell type proportions among samples. The vectors $\mathbf{h}_k$ are then referred to as 'surrogate variables', and are essentially a set of vectors that try to capture the same information that is contained in $\mathbf{g}_\ell$. The algorithm begins by fitting the unadjusted model which is defined to be:

$$y_{ij} = \mu_i + f_i(x_j) + e_{ij}, \tag{3.8}$$

where $y_{ij}$ is the methylation beta value for the $j^{th}$ subject at the $i^{th}$ locus, $\mu_i$ is a baseline methylation level for locus $i$, $x_j$ contains the phenotype of interest for subject $j$, and $f_i$ is some function of the phenotype. In our case we assume $f_i$ is simply linear over all loci, though in general this assumption is not necessary. The error term in this model is then reparameterized to take the form:

$$e_{ij} = \sum_{\ell=1}^{L} \gamma_{\ell i}\mathbf{g}_{\ell j} + e^*_{ij}, \tag{3.9}$$

which, given the fact that $\mathbf{g}_\ell$ and $\mathbf{h}_k$ span the same linear space, can be rewritten as:

$$e_{ij} = \sum_{k=1}^{K} \lambda_{ki}\mathbf{h}_{kj} + e^*_{ij}. \tag{3.10}$$

The goal is to use the residual matrix of the unadjusted model to capture information that will help formulate a set of surrogate vectors. This begins by performing a SVD (see Appendix A) on the residual matrix $\mathbf{R} = \mathbf{UDV}^\top$. The singular values $d_\ell$ can be found on the diagonal of the matrix $\mathbf{D}$. Let the number of non-zero singular values be $n_0$. Then calculate the proportion of variance explained by eigengene $k$ to be:

$$V_k = \frac{d_k^2}{\sum_{\ell=1}^{n_0} d_\ell^2}. \tag{3.11}$$

Randomly permute the elements of each row of $\mathbf{R}$ and use the resulting matrix as the response in equation 3.8. Perform $B$ such permutations and use the resulting residual matrices to obtain the statistics $V_k^{0b}, b = 1, \ldots, B$ which are calulated from the singular values like in equation 3.11. Set $\rho_k = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}(V_k^{0b} \geq V_k)$, and then invoke the restriction $\rho_k = \max(\rho_{k-1}, \rho_k)$ so that if an eigengene $k$ is declared significant, all eigengenes $k'$ (such that $k' < k$) are also declared as such. Finally, declare eigengene $k$ to be significant if $\rho_k < \alpha$ for some significance level $\alpha \in (0, 1)$. The number of surrogate variables that will be estimated is then $\hat{K}$ which is defined as the number of eigengenes declared significant.

Next the surrogate variables must be constructed. Let $\mathbf{e}_k$ (dimension $n$) be the $k^{th}$ significant eigengene from $\mathbf{V}$ for $k = 1, \ldots, \hat{K}$. For each of the significant eigengenes form a series of linear models with $\mathbf{e}_k$ as the response and $\mathbf{y}_i$ (dimension $n$) as the predictor for $i = 1, \ldots, m$. Choose the $\hat{m}_1$ most significant loci from this model,

where $\hat{m}_1$ is the estimated proportion of loci truly associated with the eigengene, obtained through Storey and Tibshirani's "General Algorithm for Estimating q-Values" [36]. Form the matrix $\mathbf{Y}_r$ by selecting from $\mathbf{Y}$ only the $\hat{m}_1$ rows corresponding to significant loci. Calculate $\mathbf{e}_j, j = 1, \ldots, n$, the eigengenes of $\mathbf{Y}_r$. Choose $\hat{\mathbf{h}}_k$ to be the $\mathbf{e}_j$ that is the most correlated with the original eigengene $\mathbf{e}_k$.

The $\hat{\mathbf{h}}_k$ for $k = 1, \ldots, \hat{K}$ are the estimates of the surrogate variables. Now that they have been obtained, the analysis becomes quite simple: include the estimated surrogate vectors as covariates in the original model:

$$y_{ij} \quad = \quad \mu_i + f_i(x_j) + \sum_{k=1}^{\hat{K}} \lambda_{ki} \hat{\mathbf{h}}_{kj} + e^*_{ij}. \tag{3.12}$$

As stated earlier, within the context of this thesis we assume $f_i$ to be a linear function. Consequently, finding a fit for equation 3.12 is not difficult. The fact that we obtain estimates for $\lambda_{ki}$ is merely incidental; at the end of the day the parameters of interest are the methylation-phenotype associations contained in $\hat{f}_i$.

### 3.4 Independent Surrogate Variable Analysis

Independent Surrogate Variable Analysis (ISVA) is a method that is very similar to SVA, but uses a different kind of matrix decomposition in its approach. The method was introduced in 2011 by Teschendorff et al. [38].

In SVA the surrogate variables were assumed to be orthogonal, and were chosen to span the same linear space as some unmeasured set of confounding vectors. Similarly, ISVA involves 'Independent Component Analysis' (ICA) which constrains one

of the matrices in its decomposition to have columns that are "statistically as independent from each other as possible" [19]. ICA boils down to a decomposition of the form $\mathbf{Y} = \mathbf{SA} + \epsilon$ with $\epsilon$ as small as possible. The algorithm begins by providing inital matrices for each of the decomposition terms, and performing an orthogonal, linear transformation on both in a manner that seeks to preserve statistical independence among the columns of $\mathbf{S}$.

Akin to SVA, the algorithm begins by fitting the unadjusted model as in equation 3.8. The latent dimension $K$ is estimated through "Random Matrix Theory"[30]. Let $\hat{K}$ be the estimated value of $K$. Perform ICA on the residual matrix $\mathbf{R}$ from the unadjusted analysis. The result is:

$$\mathbf{R} \quad = \quad \mathbf{SA} + \epsilon, \tag{3.13}$$

where $\mathbf{S}$ is $m \times \hat{K}$ and $\mathbf{A}$ is $\hat{K} \times n$.

The following steps are repeated for each row of $\mathbf{A}$, denoted as $\mathbf{A}_k$, for $k = 1, \ldots, \hat{K}$. Regress $\mathbf{A}_k$ on each CpG site, $\mathbf{Y}_i$ (i.e. each row of $\mathbf{Y}$). The resulting p-value is denoted as $p_i$. The q-values are then calculated as outlined in [36] and choose CpG sites with $q < 0.05$, or the top 500 CpG sites if the former quantity is less than 500. Let the number of chosen sites be $r_k$. $\mathbf{Y}_r$ is a simplified methylation matrix which includes only the $r_k$ chosen CpGs.

The matrix $\mathbf{Y}_r$ undergoes an ICA to obtain $\mathbf{Y}_r = \mathbf{S}_r\mathbf{A}_r + \epsilon_r$. Finally, the $k^{th}$ independent surrogate variable, $v_k$, is chosen to be the column of $\mathbf{A}_r$ exhibiting the greatest correlation with $\mathbf{A}_k$.

The analysis proceeds by including the estimated independent surrogate variables:

$$y_{ij} \quad = \quad f_i(x_j) + \sum_{k=1}^{\hat{K}} \lambda_{ki} v_{kj} + e_{ij}. \tag{3.14}$$

## 3.5 FaST-LMM-EWASher

FaST-LMM-EWASher was introduced in 2014 by Zou et al. [47]. The acronym FaST-LMM stands for "Factored Spectrally Transformed Linear Mixed Models", and refers to an algorithm designed by Lippert et al. [25] to address confounding resulting from population and family structure in genome-wide association studies. The usual linear mixed model (LMM) could certainly be used to correct for confounding, but FaST-LMM sought a more computationally efficient alternative.

### 3.5.1 Description of FaST-LMM

The FaST-LMM algorithm involves an initial linear mixed model (LMM) where the phenotype of interest and other covariates (as found in the matrix $\mathbf{X}$) are included as fixed effects. Only one methylation site will be considered at a time, i.e. at locus $j \in 1, \ldots, m$, the response would be $\mathbf{Y}_j$ (the $j^{th}$ row of $\mathbf{Y}$). The variance of the error term is expressed as $\sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}$ where $\mathbf{K}$ is the similarity matrix between samples $i = 1, \ldots, n$ and $\mathbf{I}$ is the identity matrix. The terms $\sigma_g^2$ and $\sigma_e^2$ scale the genetic and

random variance, respectively. Therefore, the LMM model for CpG site $j$ is written as:

$$\mathbf{Y}_j|\mathbf{X} \sim N\left(\mathbf{X}\beta_j, \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}\right), \tag{3.15}$$

for a set of fixed effects $\beta_j$. Normally, fitting the LMM would be computationally intensive, but the authors propose a formulation that allows the model to be expressed in a similar form.

A spectral decomposition is performed on the similarity matrix. This suggests that the matrix can be expressed as $\mathbf{K} = \mathbf{USU}^\top$ where $\mathbf{S}$ is diagonal, and $\mathbf{U}$ is an orthogonal matrix. Consequently, the variance component in the normal log-likelihood from the LMM can be rewritten as:

$$\begin{aligned}
\sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I} &= \sigma_g^2\mathbf{USU}^\top + \sigma_e^2\mathbf{I} \\
&= \sigma_g^2\mathbf{USU}^\top + \sigma_e^2\mathbf{UU}^\top \\
&= \sigma_g^2\mathbf{U}\left(\mathbf{S} + \frac{\sigma_e^2}{\sigma_g^2}\mathbf{I}\right)\mathbf{U}^\top.
\end{aligned} \tag{3.16}$$

The matrices $\mathbf{U}$ and $\mathbf{U}^\top$ can be redistributed in the log-likelihood to form a transformed response $\mathbf{U}^\top\mathbf{Y}_j$ and transformed set of predictors $\mathbf{U}^\top\mathbf{X}$. Now the new variance term in the log-likelihood is diagonal. Estimation proceeds by analytically

maximizing the log-likelihood and extracting coefficient and variance estimates for a fixed $\delta = \frac{\sigma_e^2}{\sigma_g^2}$, and then using a numerical method to find the optimal $\delta$.

### 3.5.2 Extending FaST-LMM

FaST-LMM-EWASher is an extension of this algorithm specifically designed to be used on epigenomic data in the presence of cell type heterogeneity. Its approach is a bit different than the others in that it filters out loci having methylation beta values that are unilaterally above 0.8 or below 0.2 for cases and controls; that is, it considers the methylation status at these loci to remain unchanged between cases and controls. One disadvantage of the method is that the phenotype of interest must be binary. The other covariates, however, may be categorical or continuous.

The method, like the others, begins by fitting the unadjusted model (after having filtered out methylation loci as mentioned above) and ordering all the loci by their significance. The total number of loci that will be included in the similarity matrix for the LMM is determined by ten-fold cross-validation to maximize the log-likelihood, as suggested in [26]. Let the number of chosen CpG sites be $\hat{K}$. The similarity matrix $\mathbf{K}$ is calculated using the realized relationship matrix [15] from the top $\hat{K}$ loci.

Next, the FaST-LMM algorithm is performed on the chosen subset of CpGs with the calculated similarity matrix. The p-values are calculated at each CpG, and the genomic inflation factor $\lambda$ is computed as the ratio of the median of the observed p-values to the median of the theoretical distribution. This procedure considers $\lambda > 1$ to be evidence of inflation due to confounding factors. If evidence of inflation is present, then a principal component analysis is performed on the reduced methylation

matrix (leading to linear combinations of the *samples*). Subsequently include each principal component as a covariate in the LMM until the genomic inflation factor is controlled ($\lambda \leq 1$). A limit to the number of PCs included in the LMM can be specified to stop the algorithm even if the inflation factor is greater than one, as the inclusion of too many PCs would saturate the model.

### 3.6 Deconfounding

The Deconfounding method was published in 2010 by Repsilber et al. [31] as a means for correcting for differential gene expression in heterogeneous tissue samples. Its use, however, is easily adaptable to methylation data. This method will use the methylation matrix $\mathbf{Y}$ to form a non-negative matrix factorization [22]:

**Definition 3.6.1** *For a matrix $\boldsymbol{Y}_{m\times n}$, and a given $k < \min(m, n)$, a non-negative matrix factorization (NMF) finds non-negative matrices $\boldsymbol{S}_{m\times k}$ and $\boldsymbol{C}_{k\times n}$ such that*

$$\boldsymbol{Y} \;=\; \boldsymbol{SC}. \tag{3.17}$$

This decomposition itself can be calculated using the "Least squares non-negative matrix factorization" *lsqnonneg* algorithm [23]. Additionally, in order to increase the interpretability of the final estimates, deconfounding administers another set of constraints on the decomposition:

1. $\mathbf{S}$ must be normalized.
2. All elements $c_{ij}$ of the matrix $\mathbf{C}$ must satisfy $0 \leq c_{ij} \leq 1$.
3. Columns of $\mathbf{C}$ must sum to one: $\sum_i c_{ij} = 1$.

So, the algorithm proceeds by finding a set of starting values for $\mathbf{S}$ and $\mathbf{C}$ and then applying the above constraints. Then the matrix $\mathbf{S}$ is held constant, while the

*lsqnonneg* algorithm finds **C**. The constraints are reapplied for **S**. Now **C** is held constant and **S** is found using *lsqnonneg*, and the constraints are reapplied on **C**. Finally, calculate the absolute differences between the elements of the matrices **Y** and **SC**, and stop the algorithm if a prespecified threshold has been met. Otherwise, continue from the step where **S** is held constant using the current values.

Deconfounding attemps to directly estimate the proportions of cell types found in the sample. The columns of **C** form the subject-specific estimates for the proprotions of cell types, while the rows of **S** estimate average cell-specific methylation levels at the different CpG sites. However, this interpretation is only valid under the assumption that cell type heterogeneity is the only source of unmeasured confounding in the sample. It should also be noted that the number of cell types present in the sample is not estimated by the Deconfounding method; the parameter is user-specified.

### 3.7 CellCDecon

The method CellCDecon was written very recently by James Wagner in his PhD thesis [40]. The underlying approach is a bit different than the other methods in that it doesn't apply a predefined matrix decomposition, but rather uses numerical least squares estimation on a specific model. Also, the creation of the latent vectors does not consider the phenotype of interest or the covariates. It is run instead on the normalized methylation beta value matrix alone. It could, however, be run on the residual matrix from a linear model that was not adjusted for cell type composition.

The parameters in the model assume there is an overall mean methylation level $\mu_{ic}$ for a given CpG site $i \in 1, \ldots, m$ and cell type $c \in 1, \ldots, K$. Next, suppose

that sample $j \in 1, \ldots, n$ contains a proportion $p_{jc}$ of cell type $c$. Then the partition of the observed beta matrix due to cell type composition at location $i$ for subject $j$ is $\sum_{c=1}^{K} \mu_{ic} p_{jc}$. Then the regression model for the methylation beta value $\beta_{ij}$ is parameterized as:

$$\beta_{ij} = \sum_{c=1}^{K} \mu_{ic} p_{jc} + e_{ij}. \tag{3.18}$$

CellCDecon does not internally provide an estimate for $K$, the number of cell types present in the samples, and at this point the parameter is user-specified. So, given the value of $K$, the procedure begins by populating the vectors $p_j$ ($K$-dimensional) for each subject $j$ with random proportions that sum to one. The initial values for the overal mean methylation parameter are taken to be the observed mean methylation values at each probe plus some random noise.

The algorithm allows random perturbations to occur within each $p_j$, subject to the original constraints. The change is accepted if the sum squared residuals from the model in (3.18) decreases. The same procedure occurs for the vectors $\mu_i$. These random perturbations alternate between the cell type composition and mean probe vectors for 1000 iterations, at which point the current values of the parameters are considered to be the final estimates. The vectors $p_j$ can then be used in a regression model among the phenotype of interest and other covariates and should now adjust for cell type composition.

The method is intended to give a direct estimate of the cell type composition for each subject. However, if there are other sources of unmeasured confounding in the data, these estimates are unlikely to be accurate as the model does not take them into account. Also, it can be difficult to infer the true cell type composition when the number of cell types was itself specified by the user.

### 3.8 Other methods

This chapter by no means forms an exhaustive list of all methods available at this time. It simply covers a number of novel approaches that have appeared in the last few years in the world of genomics/epigenomics. In fact, there exist a plethora of other methods under the guise of 'deconvolution' providing the same kind of correction for unmeasured confounding both in other high-throughput data sources that could be adapted for use on methylation data [45]. It is not our goal to introduce every single method available on the market; rather, we would like to choose methods in a way that allows us to touch on a variety of underlying statistical concepts.

Now that the methods have been more concretely described, it will be necessary to test how they perform. The next chapter will provide a medium of comparison by using a real methylation data set and modifying it to produce an association with a simulated phenotype. This semi-artificial data will be used as input for each of the methods, and the performance of each will be measured through several metrics.

# CHAPTER 4
# Test Data and Simulation Details

Having established an ensemble of cell type adjustment methods in Chapter 3 our focus now shifts to finding an effective means of comparison. Though part of the proposed solution involves simulation, the simulated effects will be superimposed on a real set of methylation measurements. Such a simulation allows the freedom of choosing the true association parameters that will be estimated in addition to providing a realistic distribution of methylation measurements for the different cell types. This chapter will first provide details relevant to the cell type separated methylation data that will be used in the simulation. Next, the simulation details will be specified. Finally, a few tools will be introduced that will be used later to compare performance of the different methods.

## 4.1 Cell Type Separated Methylation Data Set

### 4.1.1 Description of Data Set

We have available a data set containing methylation measurements on three separated blood cell types obtained through flow cytometry: Monocytes, CD4T-Cells, and B-Cells. This data is a unique resource coming from the research of Dr. Marie Hudson at the Lady Davis Institute in Montréal, Québec. The samples were taken from patients newly diagnosed with one of four types of autoimmune diseases: Rheumatoid arthritis (15 patients), Scleroderma (21 patients), Systemic lupus erythematosus (11 patients), and Myositis (4 patients). There were several control

samples available as well. Purity measures were provided to ensure good quality of the cell separation process. Some samples failed quality control or insufficient tissue was available so we do not have all three cell types for all patients. Patients having only one cell type available for measurement were not included in the simulation or analysis as they could not be used to form an artificial cell mixture. There were very few samples of B-cells available, and even fewer samples where all three cell types were available. Consequently, my simulation only focuses on mixtures from monocytes and T-cells. In total 46 patients were included in my simulation each with measurements from monocytes and CD4 T-cells.

### 4.1.2 Data Quality Issues

Data normalization was performed as in Section 2.4, with functional normalization as the method used. Other measures of quality control were performed in order to identify any bad samples. Also, a number of probes were removed, specifically those on the sex chromosomes as well as probes close to SNPs. In order to attempt to remove remaining batch effects, a regression model was run on the remaining samples which included the following covariates: sample plate (i.e. batch number), position on plate, and the bisulfate conversion rate for the sample. This procedure made some improvements, but some batch effect remained.

The main issue with the data set is, in fact, the study design. Samples coming from patients having the same autoimmune diseases were included in the same batch. Therefore, effects due to technical factors are impossible to remove without affecting those coming from differences in the autoimmune diseases.

## 4.2 Simulation

The goal of the simulation is to create methylation measurements on an artifical cell mixture. The simulation uses the separated blood cell data, modifes it and then combines data across cell types in order to obtain an artificial mixture of cell specific methylation profiles. The simulation is comprised of three elements: a simulated case/control phenotype, a change in methylation at certain sites associated with phenotype, and a change in cell type proportion associated with the phenotype. There are also precision parameters that can adjust distributional variability at most stages of the simulation. The details are outlined in the next section.

### 4.2.1 Simulation Steps

1. **Simulation of phenotype:** Take random sample from the $n$ patients to be cases/controls with $P(case) = P(control) = 1/2$.
2. **Simulation of cell distributions:** For each individual I assume values of the proportions of cell types in that individual's sample. Specifically, I assume variability around an average proportion separately for cases and controls. The average proportions for controls will make up the vector $\alpha_1$, and for cases $\alpha_2$. Then, for subject $i \in 1, \ldots, n$ we simulate cell type proportions from the following distribution:

$$Dirichlet(\rho\alpha_1), \quad \text{if subject } i \text{ is a control} \tag{4.1}$$

$$Dirichlet(\rho\alpha_2), \quad \text{if subject } i \text{ is a case,} \tag{4.2}$$

where $\rho > 0$ is the so-called 'precision' parameter. (i.e. a greater precision parameter corresponds to less variation in the observed values).

3. **Simulation of associated sites:** Choose at random $S$ CpG sites from the $m$ sites in the original matrix of methylation beta values. These will be differentially methylated with the phenotype in the simulation. Draw a cell type specific effect at each chosen CpG site from $N(\mu_{k,overall}, \sigma^2_{k,overall})$ for $k \in 1, \ldots, K$ where $K$ is the number of cell types in consideration. Set $\mu_{j,k}$ to be the effect (i.e. the observed value from the aforementioned distribution) for CpG site $j$ and cell type $k$.
4. **Simulation of individual effects:** If subject $i$ is a case and site $j$ is one of the $S$ selected sites from the last step, generate a subject specific effect $e_{i,j,k}$ from $N(\mu_{j,k}, \sigma^2_{j,k})$ for each cell type $k$. Let the corresponding methylation value be $\beta_{i,j,k}$ then the new (affected) methylation value will be:

$$\beta'_{i,j,k} \quad = \quad \text{logit}^{-1}\left(\text{logit}(\beta_{i,j,k}) + e_{i,j,k}\right). \tag{4.3}$$

For controls and non-differentially methylated sites, set $\beta'_{i,j,k} = \beta_{i,j,k}$.

5. **Combining measurements:** Let $p_k$ for $k = 1, \ldots, K$ be vectors of length $n$ containing the simulated cell proportions for each subject of cell type $k$. Then

the final beta value for person $i$ at CpG site $j$ will be:

$$\beta^f_{i,j} = \sum_{k=1}^{K} p_{i,k}\beta'_{i,j,k}, \tag{4.4}$$

where $p_{i,k}$ is the $i^{th}$ entry of $p_k$, i.e. the proportion of cell type $k$ for subject $i$.

The use of the Dirichlet distribution is advantageous, as it allows us to both specify the mean for each cell type over all cases or controls, as well as the precision in the observed values. The support of the distribution is constrained to a vector with elements in the interval (0,1) such that the elements sum to one. It is therefore well-suited for simulating a set of proportions for each subject [17].

The simulation exhibits variability at multiple levels. Over all differentially methylated regions one would expect to see a range of positive and negative associations with the phenotype. We also allow these associations to differ between cell types in order to specify an association between change in methylation and cell type. After having specified each of the site and cell type specific associations, we add some between-subject variability to each site (in step 4). In order to stay within the expected bounds of the beta value, we work on the logistic scale. However, after simulating the effects, it is necessary to revert back to the original scale as a few of the cell type adjustment methods cannot be run on values on the logistic scale. Finally, the sum in the final step simulates how methylation measurements coming from $K$ different cell types with proportions $p_{i,k}, (k \in 1, \ldots, K)$ would appear had they been present simultaneously in a sample.

The simulation will be run under different values of parameters in order to capture different scenarios that might occur in real methylation data. The parameters that will be modified between simulation runs are: $\mu_{k,overall}$ and $\sigma^2_{k,overall}$ which change the mean and variance of the overall association with the phenotype among the chosen CpG sites, and $\rho$ which changes the precision of the patient-level cell type distributions between cases and controls. We can specify how different the cell type distributions are relative to each other, and we can change the amount of variation present among either the CpGs or subjects. The motivation behind altering parameters is to see if the performance of the adjustment algorithms differs according to the underlying biological structure, as well as in the presence of technical noise (e.g. batch effect).

It is evident that a simulated dataset should be generated in a way that embodies the statistical properties of real data. In order to verify whether our simulated data are statistically realistic, we extracted the patients' ages from the cell separated data and fit a linear model between age and methylation separately for each cell type. In looking at the top age-related CpG sites (on the logistic scale) for each cell type, the effect distributions looked approximately normal. In some cases the effect distributions showed bimodality, but there was always either a strong positive or negative net effect. Therefore, using a normal distribution on logit-transformed methylation data to simulate a cell type's effect distribution seems to be an appropriate simplification. Admittedly, the parameters in the simulation are set in a way to generate a large number of associated CpGs with strong effect sizes, perhaps more than would

be expected for a real covariate. However, this facilitated a more informative power and false discovery rate analysis (see Subsection 4.3.3).

It is important that the simulation be run after the normalization step. Batch effects and other biases would make measurements from different samples inherently incomparable and therefore difficult to combine. We also only combine cell type samples for the same subject. Doing so for permutations of different subjects would certainly be a viable way to increase sample size, but as one would expect larger variability between subjects, we believe this idea is beyond the scope of the thesis.

We now have a strategy for creating a simulated data set on which the adjustment methods can be tested. The next section will address a few tools that will be used to compare the perfomance of each method.

## 4.3 Performance Metrics

The best evaluation measure of a method's performace is not always obvious. Though one method may result in a more powerful regression model (identifying associations between methylation levels and the phenotype), it may invoke a number of false positives. Additionally, there may be limitations on the type of data on which it can be used. This section will outline several tools, both graphical and numerical, that will be used together to obtain a more clear vision as to the performance and usefulness of each adjustment method. The results of using each of these tools on the simulated data set are found in Section 5.

### 4.3.1 QQ-Plot for P-values

Genomic inflation refers to a systematic decrease in p-values across most or all loci considered in the study, thereby running the risk of false positives. It is one of

the most common symptoms of confounding due to cell mixtures. In order to check for genomic inflation, we need see if there exist a set of observed p-values that are smaller than expected by chance. In other words, even if there were no association between phenotype and methylation, we would expect to see a fair number of very small p-values as a result of the large number of tests done (on 480,000 regression models linking CpG sites to the phenotype). By comparing to a 'null' distribution (i.e. the distribution of the test statistic given no association), we can identify p-values that are significantly smaller than one would expect by chance. The QQ-Plot (Quantile-Quantile) is an excellent way to spot this kind of departure from the null distribution. The idea is to take the quantiles from a theoretical null distribution and plot them against the quantiles from the observed data. Should the data truly come from the null distribution, we would expect the points to lie on or near the diagonal line '$y = x$'.

We did, of course, simulate associations between the phenotype and methylation at several probes. So in this case, we have the benefit of knowing that certain CpG sites will indeed be associated with the phenotype. We would therefore expect to see, after correcting for cell heterogeneity, a large set of p-values closely following the null distribution, and a select few p-values at the top of the plot after a $-log_{10}$ transformation.

To illustrate what one would expect to see in a QQ-Plot showing genomic inflation, we refer to a toy data set whose results are shown in Figure 4–1. Imagine a test statistic whose distribution under the null hypothesis is $N(0, 1)$. Then, if the null hypothesis is true, the p-values calculated from the tails of this null distribution should

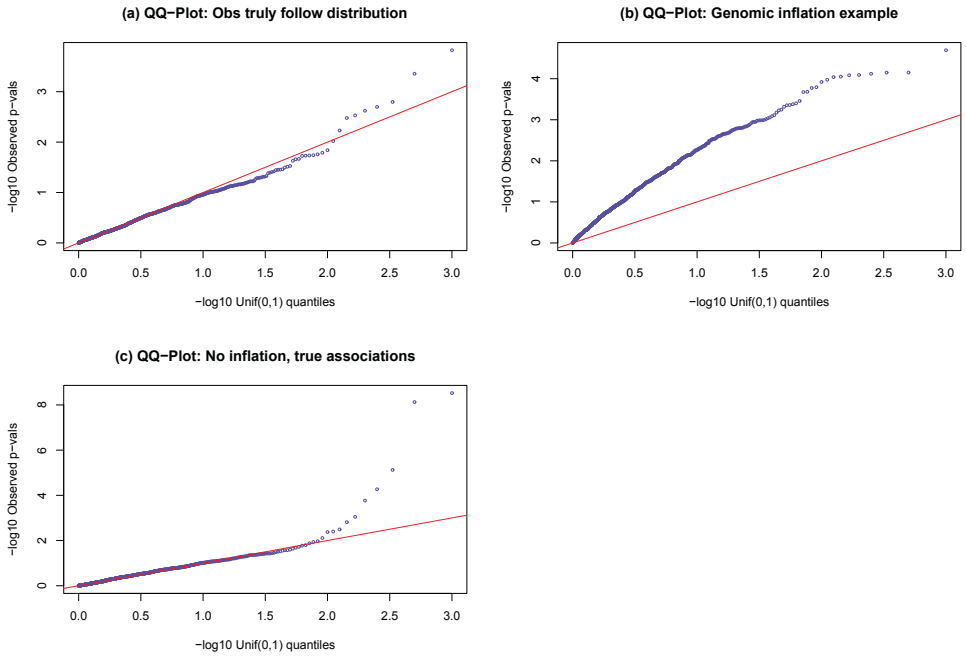


Figure 4–1: Examples of different scenarios for QQ-Plots. On each plot the horizontal axis represents theoretical quantiles from a Unif(0,1) distribution and the vertical axis shows observed p-values. Plot (a) contains p-values corresponding to observations truly coming from the null distribution. Plot (b) contains p-values corresponding to genomic inflation, i.e. they are smaller than one would expect under the null. Plot (c) contains p-values from observations not showing genomic inflation, but indeed exhibiting some true associations. Note that both axes are on the $-log_{10}$ scale.

follow the $U(0,1)$ distribution. In plot (a) we generated 1000 observations from the standard normal distribution, and calculated the p-values (based on the $N(0,1)$ distribution) to see if they would appear to be uniformly distributed. It can be seen that, because the observations fall close to the diagonal line that they are indeed uniformly distributed, and there is no evidence of inflation. In plot (b) observations were drawn from the distribution $N(0,1) + U(0,2)$. After calculating the p-values (still based on the N(0,1) distribution), all points fall well above the diagonal, showing inflation. Finally, in plot (c) we drew observations from $0.99N(0,1)+0.01U(0,5)$. Plotting p-values as before shows most fall on the diagonal, with a small number of points falling above the line. This is what one would expect to see after correcting for genomic inflation in a data set exhibiting a small number of true associations.

### 4.3.2 ROC Curves

Receiver operating characteristic (ROC) curves are rather commonplace in today's literature, but remain an integral part of the analysis of a model's predictive power. The main function of the ROC curve is to examine the tradeoff between a model's sensitivity (the probability that the null hypothesis is rejected given that it is false) and false positive rate (the probability that the null hypothesis is rejected given that it is true), with respect to the adjustment of a discriminatory threshold [12].

In the simulation study, we have knowledge of which CpG sites are truly associated with the phenotype of interest. Hence, at a given p-value threshold, calculating sensitivity and specificity is straightforward. The threshold is allowed to vary between the smallest and largest p-values at some predetermined step size. The

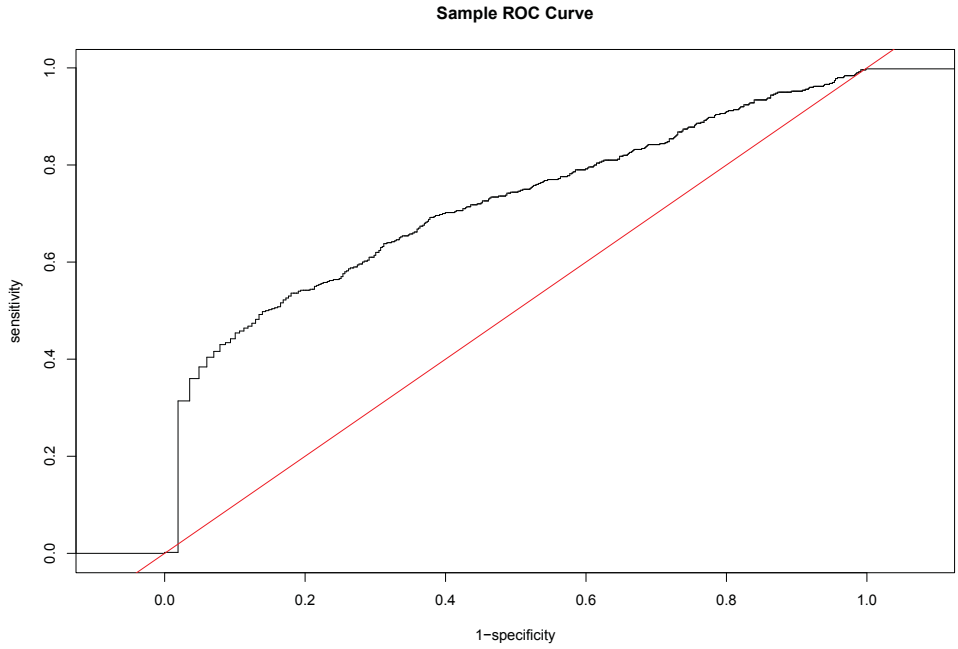


Figure 4–2: An example of an ROC curve. The horizontal axis corresponds to the false positive rate (1-specificity), and the vertical axis corresponds to the true positive rate (sensitivity). The curve shows the tradeoff between sensitivity and specificity among different levels of a p-value threshold. A method that lies on the diagonal performs as well as a random guess (i.e. $P(\text{reject null}) = P(\text{do not reject null}) = 1/2$). A method performs well if the curve lies well above and to the left of the diagonal.

sensitivity and false positive rate are calculated at each threshold level and plotted against each other. Of course, in an EWAS one has to be careful as even a small gain in sensitivity could be accompanied by a large number of false positives. Therefore, when viewing the ROC curves for the different adjustment methods, the main region of interest will be the far left hand side of the curves where the false positive rate is small. An example of an ROC curve can be seen in Figure 4–2.

### 4.3.3 Power After Controlling for False Discovery Rate

A useful tool for evaluating a method's performance is the calculation of its statistical power. Power is defined to be the probability that the null hypothesis is rejected given that the null hypothesis is false. In the context of the EWAS, power specifically refers to the probability that a methylation locus that is truly associated with the phenotype is declared as such. Of course, an increase in power is not very useful if it is accompanied by a large increase in the type I error rate (i.e. the probability that the null hypothesis is rejected when the null is true). In multiple testing, it is often necessary to control for the Family-Wise Error Rate (FWER), which is the probability of making at least one type I error among all the tests [3]. Procedures correcting for the FWER (e.g. Bonferroni correction) are often overly conservative in the context of genomics given dependence between tests. Therefore, in these kinds of data, a less stringent criterion is needed.

The False Discovery Rate (FDR) is an important measure in genomics and epigenomics, and is indeed less stringent a criterion than FWER. FDR is defined to be the expectation of the proportion of hypotheses that are incorrecly rejected among all rejections. More precisely, if we let $R$ be the total number of rejections, and $R_0$ be the number of rejections when the null hypothesis is true, the FDR is defined to be $E(R_0/R)$.

In 1995 Benjamini and Hochberg suggested an algorithm that adjusted p-values in a way that controlled the FDR [3]:

**Definition 4.3.1** *If $T$ is the total number of tests done, and the ranked p-values are $P_{(i)}$, then find $c$ to be the largest $i$ such that*

$$P_{(i)} \leq \frac{i}{T}q, \tag{4.5}$$

*and reject each null hypothesis corresponding to $i \leq c$.*

It was shown that among the rejected hypotheses, $E(R_0/R) \leq q$, thus controlling the FDR. Alternatively, one could simply report the adjusted p-values, $\frac{T}{i}P_{(i)}$ for all $i \in 1, \ldots, T$. It was also demonstrated that this procedure generally leads to a higher power than do FWER adjustment algorithms. Given how mainstream this kind of procedure is in genomic studies, it seems comparing power under an FDR adjustment would be a good measure of the performance of each of the cell type adjustment methods.

### 4.3.4 Kolmogorov-Smirnov Statistic

The Kolmogorov-Smirnov test allows us to compare whether an observed distribution differs from some reference distribution. It compares the distribution function of the observed data to the theoretical distribution function for the reference distribution. The calculated statistic considers the maximum absolute difference between these two functions [7].

Assume iid observations $X_1, \ldots, X_T$. First, calculate the empirical distribution function from the $T$ observations: $\hat{F}_T(x) = \frac{1}{T}\sum_{\ell=1}^{T} \mathbf{1}(X_\ell \leq x)$. Then suppose the distribution function from the comparison distribution is $F(x)$. Define the (one sample, two-sided) Kolmogorov-Smirnov statistic to be:

$$D = \sup_x |\hat{F}_T(x) - F(x)|. \tag{4.6}$$

So, computing the Kolmogorov-Smirnov statistic over the observed p-values relative to the $U(0,1)$ distribution should give a good idea as to whether there is genomic inflation present. Evidently, the calculated p-values from the EWAS are not independent, thus the iid assumption is violated. However, we are only interested in the raw value of the statistic for each method for the sake of comparison. We will not calculate the p-values associated with the Kolmogorov-Smirnov test and the statistic $D$ will serve only as a measure of a method's genomic inflation relative to the others.

Though the simulation and performance metrics have been described, it remains to see how the adjustment algorithms will perform relative to each other given the circumstances we have specified. The next chapter will provide the results of this comparison.

# CHAPTER 5
# Results

Here we present the results of testing the different cell type heterogenity adjustment methods. Each method is run on different sets of parameters in the simulation from Section 4.2.1, and results are compared using the previously defined performance metrics. We also test the adjustment methods on a real data set (no simulation involved) and attempt to compare performance without knowing the true parameters. Finally, computing performance will be measured using a benchmark data set, and it will be seen how computing time scales according to number of samples, and other parameters.

It would also be worthwhile to look at how the power and FDR change with respect to sample size. In any regression model, an increased sample size should certainly improve power, assuming the increase in samples does not introduce any new biases. In our case, however, it would be valuable to see if any of the methods responds more positively to a modest increase in sample size. Unfortunately, since we only have 46 samples from which to form cell mixtures, we do not believe we can do a meaningful analysis on power versus sample size; therefore, this is left as future work.

## 5.1 Simulation Results

Five different scenarios have been simulated. The first assumes distinct degrees of association with the phenotype in the two cell types. The second assumes there

is no such confounding present. The third specifies that the two cell types have, on average, opposing associations with the phenotype, i.e. one cell type has a net positive effect, and the other, a net negative effect. The final two scenarios contain a distinct difference in cell type effect, but simulate high and low precision in cell type composition among subjects. In every case, the number of CpG sites randomly chosen to be associated with the phenotype is 500. A summary of the important parameters the simulations is seen in Table 5–1.

Table 5–1: Summary of parameters in simulations

| Parameter | Description |
|---|---|
| $\alpha_1$ and $\alpha_2$ | Vectors containing average cell type proportions for cases and controls, respectively |
| $\rho$ | Precision of simulated cell mixture distributions. Greater value corresponds to more precision |
| $S$ | Number of CpG sites chosen to be associated with the phenotype (500 in all scenarios) |
| $\mu_{k,overall}$ | Average simulated effect over all CpG sites for cell type $k$ |
| $\sigma^2_{k,overall}$ | Variance of simulated effects over all CpG sites for cell type $k$ |

### 5.1.1 Scenario 1: Distinct Associations with Phenotype in Cell Types

The first run of the simulation specified parameters that allowed the distribution of phenotype-methylation associations to differ greatly between the two cell types. That is, the realized values of $e_{i,j,k}$ in equation (4.3) will differ greatly between the two cell types. The values of the parameters in this specific scenario were chosen to induce significant confounding. Therefore, performing a regression analysis unadjusted for cell type composition should result in many p-values that are smaller than expected, or equivalently a greatly inflated p-value QQ-plot.

More specifically, the parameters (as in section 4.2.1) were set to the following:

- Average cell type proportions for controls (i.e. entries of $\alpha_1$) were 0.43 for T-cells, and 0.57 for monocytes. For cases the average proportions (entries of $\alpha_2$) were 0.35 for T-cells, and 0.65 for monocytes. The initial control proportions were chosen to preserve the same T-cell to monocyte ratio as found in [17].
- The net effect for T-cells was chosen to be positive, with $\mu_{T,overall} = 0.5$ and $\sigma^2_{T,overall} = 0.75^2$. The net effect for monocytes was chosen to be negative, but with a lesser magnitude than that of T-cells, with $\mu_{Mono,overall} = -0.05$ and $\sigma^2_{Mono,overall} = 0.05^2$.

The generated distributions of the simulated associations between methylation and the phenotype (for 500 differentially methylated positions) can be seen in Figure 5–1 which shows drastic differences between monocytes and T-cells, both in terms of net effect, as well as variability.

Similarly, the distribution of the simulated proportions of T-cells over all subjects can be seen in Figure 5–2. On average, cases reduce the proportion of T-cells by 0.08 compared to controls. The precision parameter $\rho$ in the Dirichlet distribution as seen in Section 4.2.1 was set to 100, specifying a marked difference between cases and controls. A higher value of $\rho$ corresponds to smaller variation in the simulated cell type proportions among the samples.

The QQ-plots showing the distributions of p-values for the phenotype-methylation tests in the different adjustment methods can be found in Figure 5–4. Figure 5–4 (a) shows the p-values from a regression analysis that has not been adjusted for cell type heterogeneity. As expected, there is genomic inflation present, as almost all the

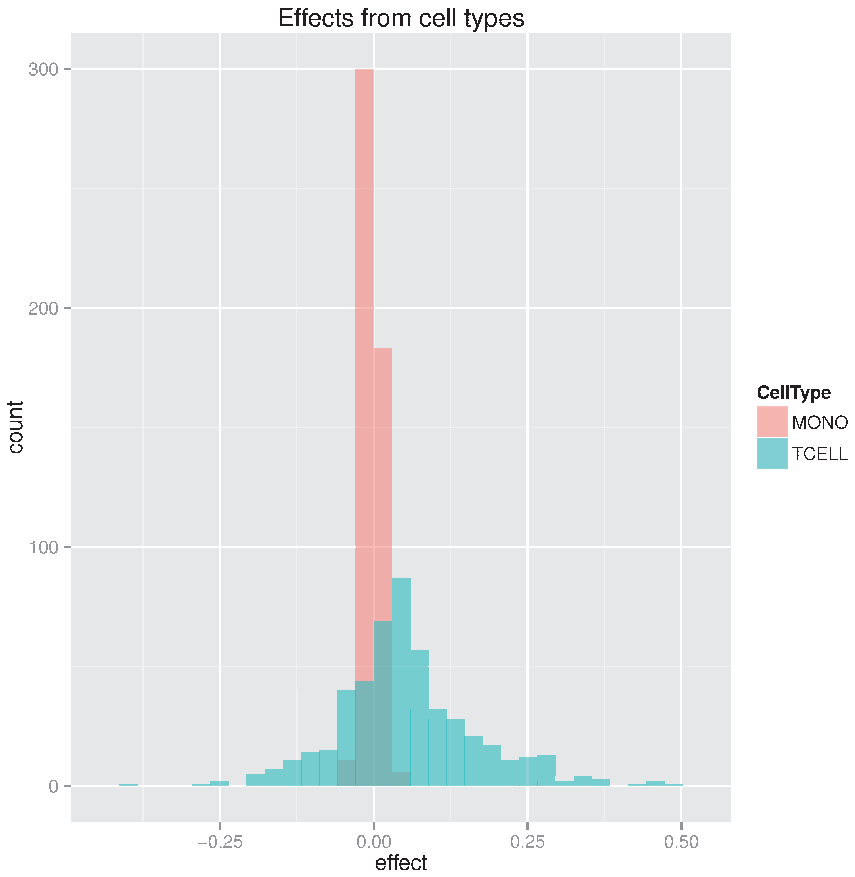


Figure 5–1: Distributions of simulated effect sizes at the 500 chosen CpG sites for different cell types. Green represents the distribution for T-Cells and exhibits a positive net effect, but a high amount of variation. Red represents the distribution for monocytes and exhibits a small negative net effect, but with a less variation. Dark green represents where the distributions overlap.

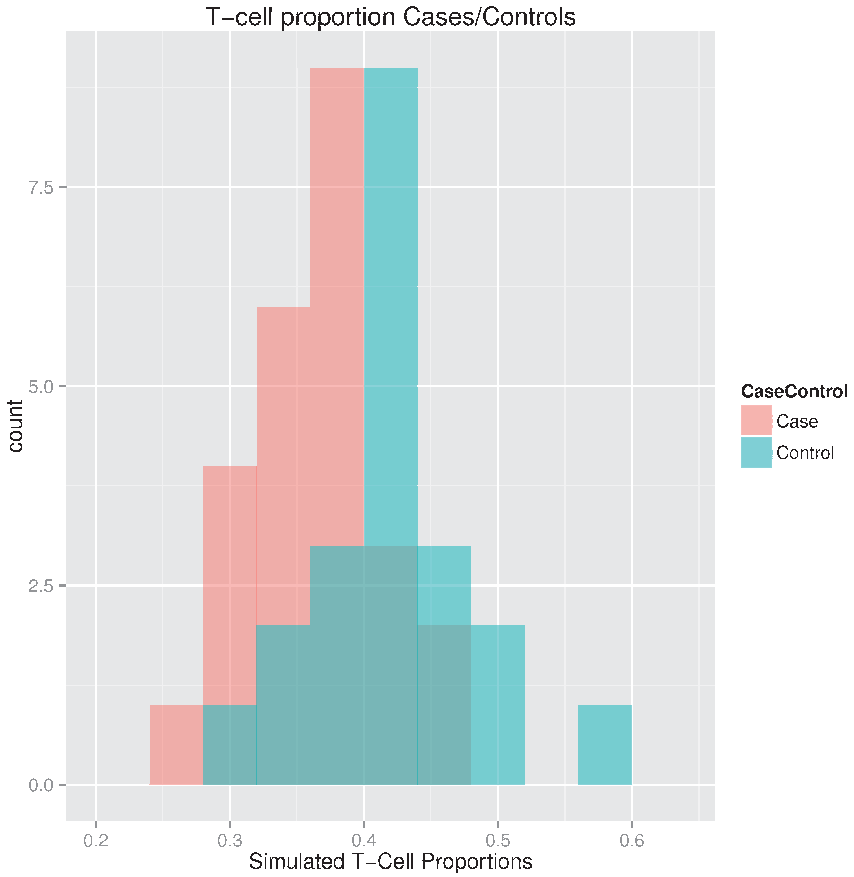


Figure 5–2: Histogram for the simulated distributions of T-cell proportions among cases and controls. On average, cases were subject to a 0.08 reduction in T-cells. Dark green represents where the distributions overlap

points fall above the line '$y = x$'. In plot (b) it can be seen that the reference-based method does a good job in reducing the inflation, but perhaps does a bit of over-correction on closer inspection. The reference-free method, however, still appears to have a little bit of inflation present. The methods SVA, ISVA, CellCDecon, and Deconf all appear to satisfactorily control for confounding.

The FaST-LMM-EWASher method was originally run with no covariates, however, the resulting p-value plot appeared to show that the method performed a significant overcorrection. This can be seen in Figure 5–3. However, when including the patients' autoimmune disease type (from the original clinical data; not simulated) as an additional covariate, the resulting p-value plot (Figure 5–4 (f)) appears to adjust as one would expect.

One interesting finding is the fact that, though we only included two cell types in our simulation, the methods that estimate latent dimensions all estimated a much higher value. For example, the reference-free method estimated the latent dimension to be $d = 13$. SVA and ISVA estimated the number of surrogate variables to be 10 and 12, respectively. In both SVA and ISVA it is assumed that the number of surrogate variables is less than or equal to the number of true confounders whose linear space they span. Perhaps there is some unexplained variation present in the original data (not specified in the simulation) that these methods pick up on. In fact, fitting linear models on the separated cell types using the patient age as the phenotype showed that the latent dimension is high even when no cell mixture is present. For example, Random Matrix Theory (used for dimension estimation in the reference-free method and ISVA) estimated a latent dimension of 10 for both

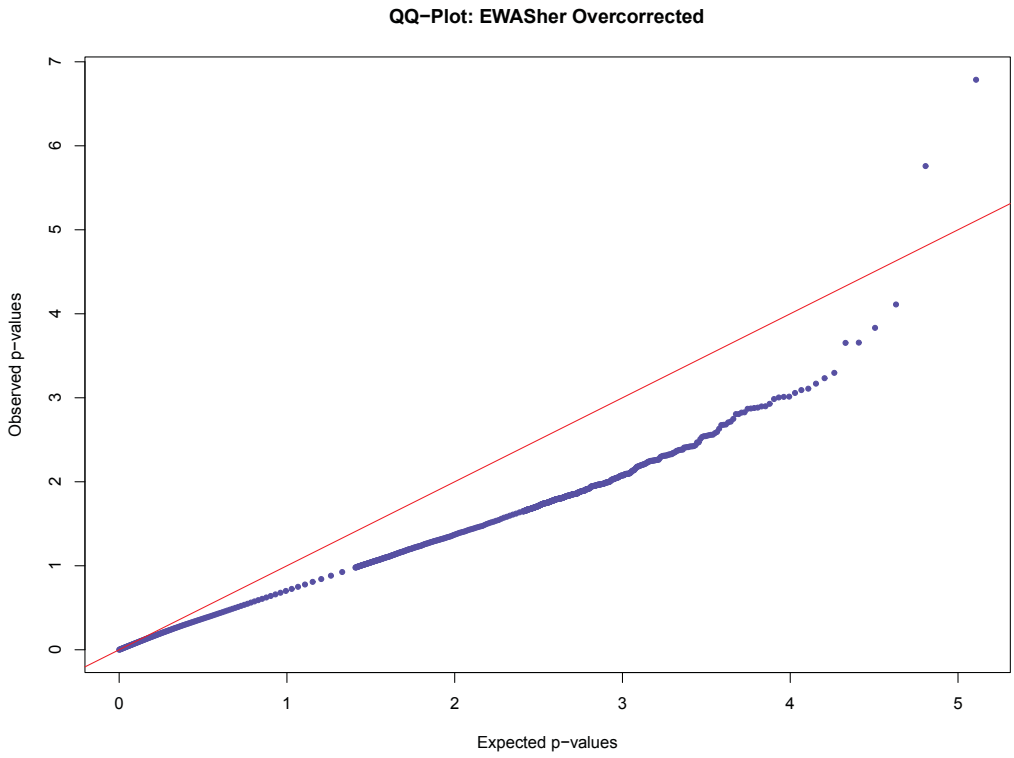


Figure 5–3: QQ-Plot for the distribution of p-values for the method FaST-LMM-EWASher without including additional covariates in the model.

T-cells and monocytes. Similarly, SVA estimated the latent dimensions of T-cells and monocytes to be 7 and 9, respectively.

The CellCDecon and Deconf algorithms were both run multiple times assuming the number of cell types to be 2 through 8, inclusive. For both methods the worst result occurred when the assumed number of cell types was 2. This result is analogous to the high estimates of latent dimension in reference-free, SVA, and ISVA. Results were almost indistinguishable for the values 3 through 8. Motivated by this observation and the desire for simplicity, results shown in this thesis for these two methods assume three cell types.

It is also necessary to examine whether the methods correctly identified the CpG sites chosen to be associated with the phenotype. First, we examine the ROC curve in Figure 5–5. The curves for all adjustment methods lie above that of the unadjusted regression model. The reference-based method is, overall, the best performing method, however, the most important section of the plot is the leftmost side, as even a small increase in the false positive rate is unacceptable among such a large number of statistical tests. It is clear that it will be necessary to examine the other performance metrics in order to better evaulate how well these methods are picking out the simulated associations. Additionally, though the curve for EWASher appears to be on par with the other adjustment methods, it is worth noting that the filtering step did remove a number of the chosen CpG sites.

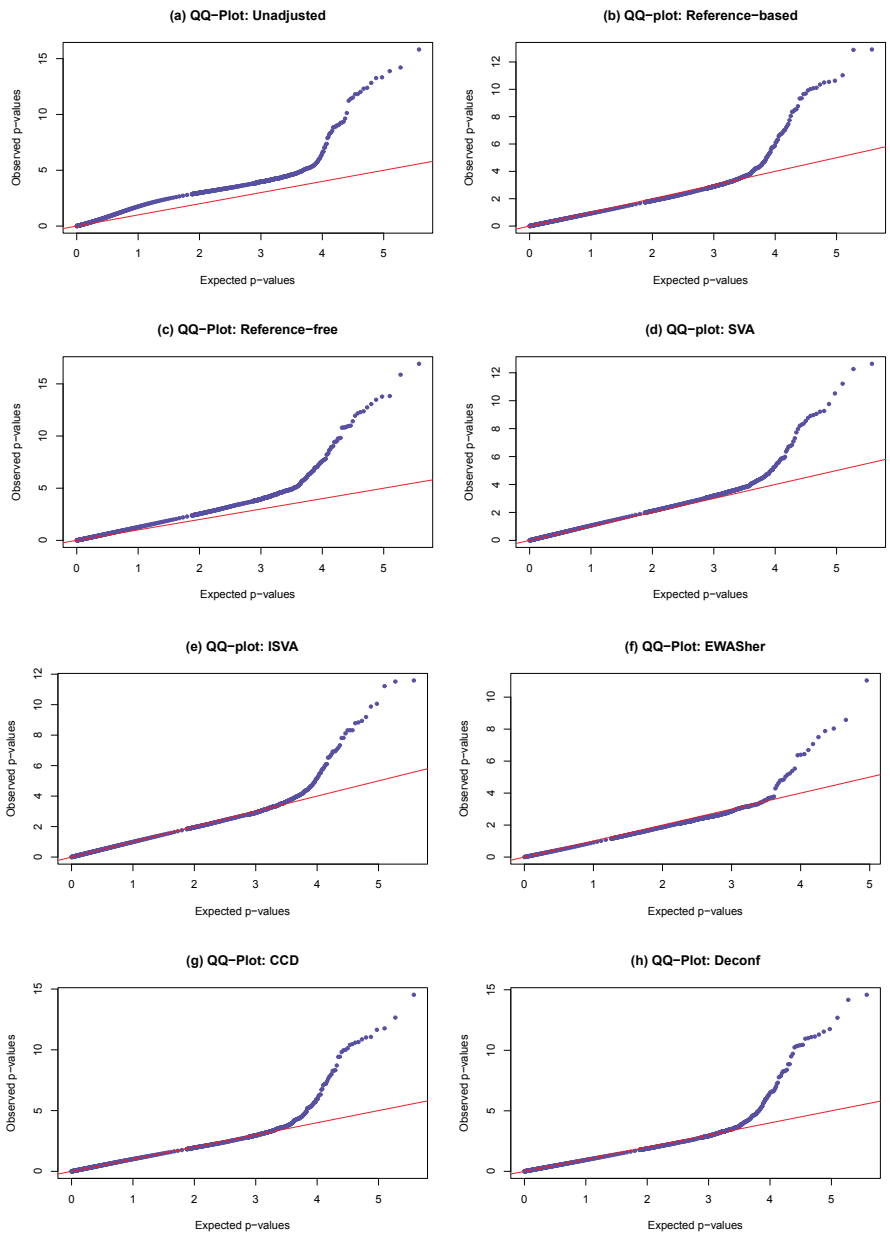


Figure 5–4: QQ-plots showing the distributions of p-values for (a) no cell type adjustment, (b) reference-based adjustement, (c) reference-free adjustment, (d) SVA, (e) ISVA, (f) EWASher, (g) CellCDecon, and (h) Deconf. Note both axes are on the $-log_{10}$ scale.

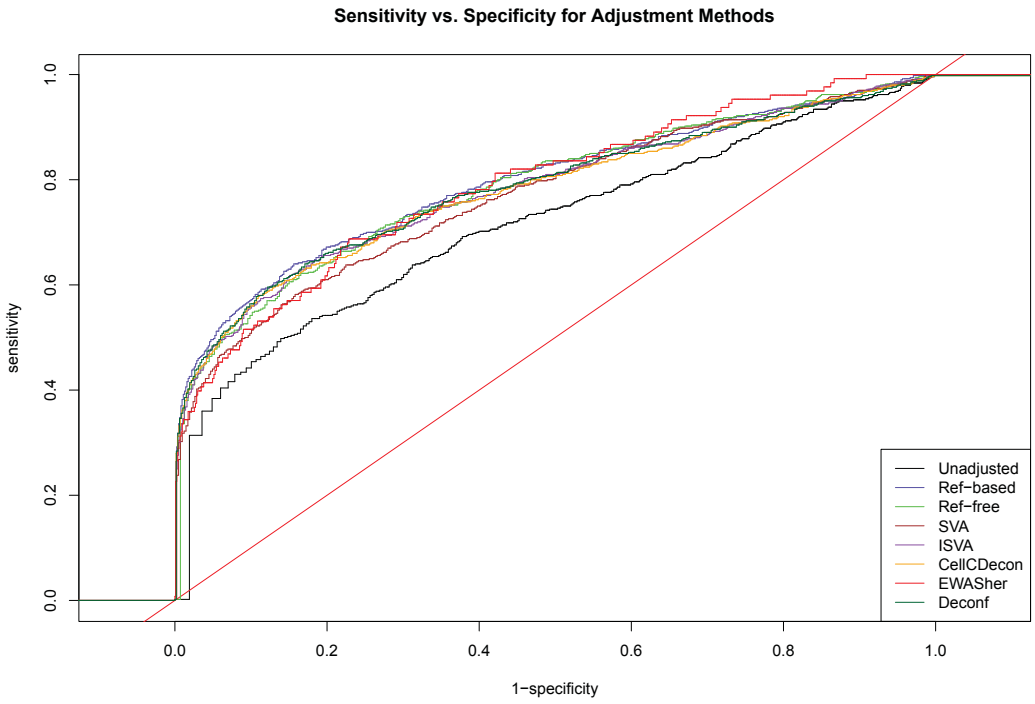


Figure 5–5: The ROC curves for the different adjustment methods in the distinct associations in cell types scenario.

Table 5–2 outlines the results from the performance metrics. Here, the null hypothesis for a given CpG site is rejected if the Benjamini-Hochberg corrected p-value is less than 0.05. The shown values are: the false discovery rate, power, and the Kolmogorov-Smirnov test statistic.

It must be stressed that the power for these methods will appear to be extremely low, as some of the CpG associations were simulated to be quite close to zero, and cannot be identified among the thousands of CpG sites that correlated with the phenotype by chance alone. However, the values in the table can be compared across methods to evaluate performance.

The power from the unadjusted analysis is one of the highest for these data, but its FDR is particularly high, which underscores the need for these correction methods. The reference-based method results in power almost as high as that for the unadjusted analysis, but with a satisfactory reduction in the FDR. An odd result can be seen for the reference-free method: the power is greatly increased, but is accompanied by a very large increase in FDR. This result is analogous to its failure to account for genomic inflation as well as the other methods. ISVA slightly outperforms SVA both in regards to power and FDR. CellCDecon and Deconf both perform significantly better than the unadjusted model. FaST-LMM-EWASher's power is quite low, as some of the associated CpGs were removed in the filtering step. When considering the top 100 CpG sites chosen by each method, there were 19 sites chosen simultaneously by all the methods (including unadjusted). When removing the results from EWASher, this number increases to 46.

All methods showed a reduction in the Kolmogorov-Smirnov (KS) statistic, implying that in most cases (excluding reference-free) confounding was more or less accounted for. The method that best corrects for inflation is CellCDecon assuming three cell types present. The KS statistic for the reference-based method and EWASher were both slightly higher than the others. This is a testament to the fact that both methods have overcorrected a bit: their points actually lie below the line that corresponds to the null distribution. The KS statistic calculated here is, in fact, meant to be used for a two-sided test.

Table 5–2: Performance metrics comparison for the distinct associations in cell types scenario

| Method | FDR (0.05 Thresh.) | Power | KS Stat. |
|---|---|---|---|
| Unadjusted | 0.179 | 0.096 | 0.1679 |
| Ref-based | 0.021 | 0.094 | 0.0255 |
| Ref-free | 0.53 | 0.206 | 0.1459 |
| SVA | 0.054 | 0.07 | 0.0205 |
| ISVA | 0 | 0.078 | 0.0035 |
| EWASher | 0.071 | 0.026 | 0.0437 |
| CellCDec | 0.021 | 0.094 | 0.0084 |
| Deconf | 0.037 | 0.104 | 0.023 |

### 5.1.2 Scenario 2: No Confounding

The next simulated scenario assumes that the distributions for the associations between the chosen CpG sites and methylation do not differ between the two cell types. All parameters were chosen to be the same as in Subsection 5.1.1, except the following $\mu_{T,overall} = \mu_{Mono,overall} = 0.25$ and $\sigma^2_{T,overall} = \sigma^2_{Mono,overall} = 0.5^2$. The goal here is to understand how the methods perform when there is no confounding present.

The resulting distribution of associations (effects) for the two cell types is illustrated in Figure 5–6. The figure clearly shows that the observed distributions of effects are very similar to one another.

The QQ-plots showing the distributions of p-values for the no confounding scenario can be found in Figure 5–7. The unadjusted model shows no inflation, which is expected. SVA and reference-free both show signs of a little bit of inflation, and ISVA shows a fair amount of inflation compared to the unadjusted model, though looking at the vertical scale shows the smallest p-values are not quite as small as in the unadjusted model. Looking at the numerical performance metrics should make things more clear.

The values for the different performance metrics can be found in Table 5–3. In several cases, the FDR actually ends up being worse after performing an adjustment method. In some cases, power has actually gone down as well. There has not been a large departure from the null distribution in any of the methods, except for once again, the reference-free method. The implication of these results is that in the event there is no confounding present, running a cell type adjustment method could actually be deleterious.

### 5.1.3 Scenario 3: Opposing Effects

Another interesting scenario is monocytes and T-Cells having opposite associations with the phenotype. One would expect this to make finding differentially methylated positions difficult as any effects would be cancelled out. Once again, most of the parameters remain the same as in Subsection 5.1.1, with exception to

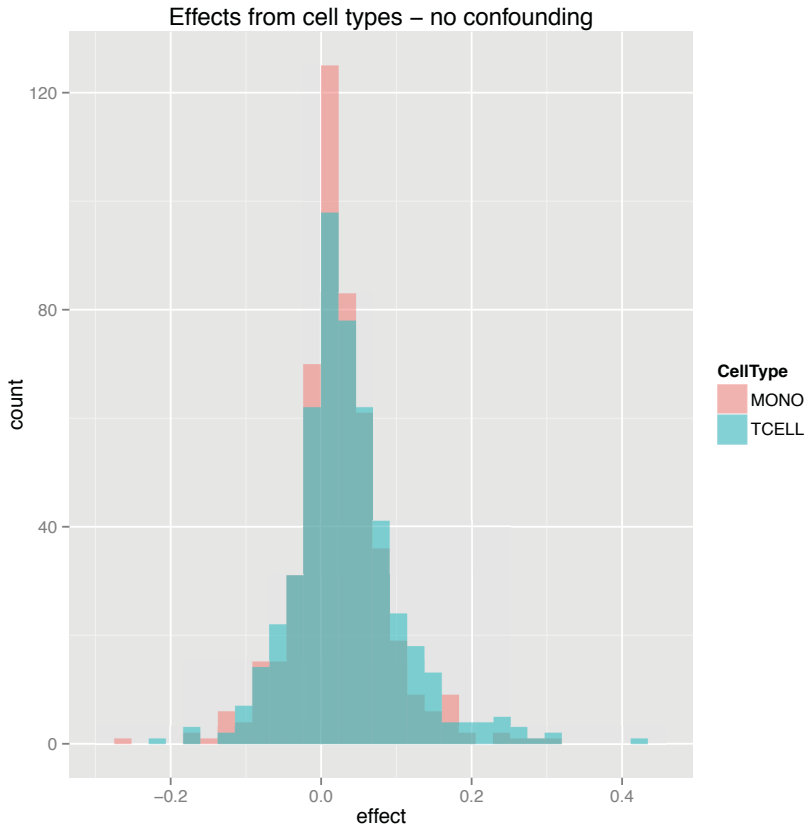


Figure 5–6: Distributions of simulated effect sizes at the 500 chosen CpG sites for different cell types. Green represents the distribution for T-Cells and red represents the distribution for monocytes. Dark green represents where the distributions overlap. Both sets of effects were generated from the same distribution to have a small net positive effect.

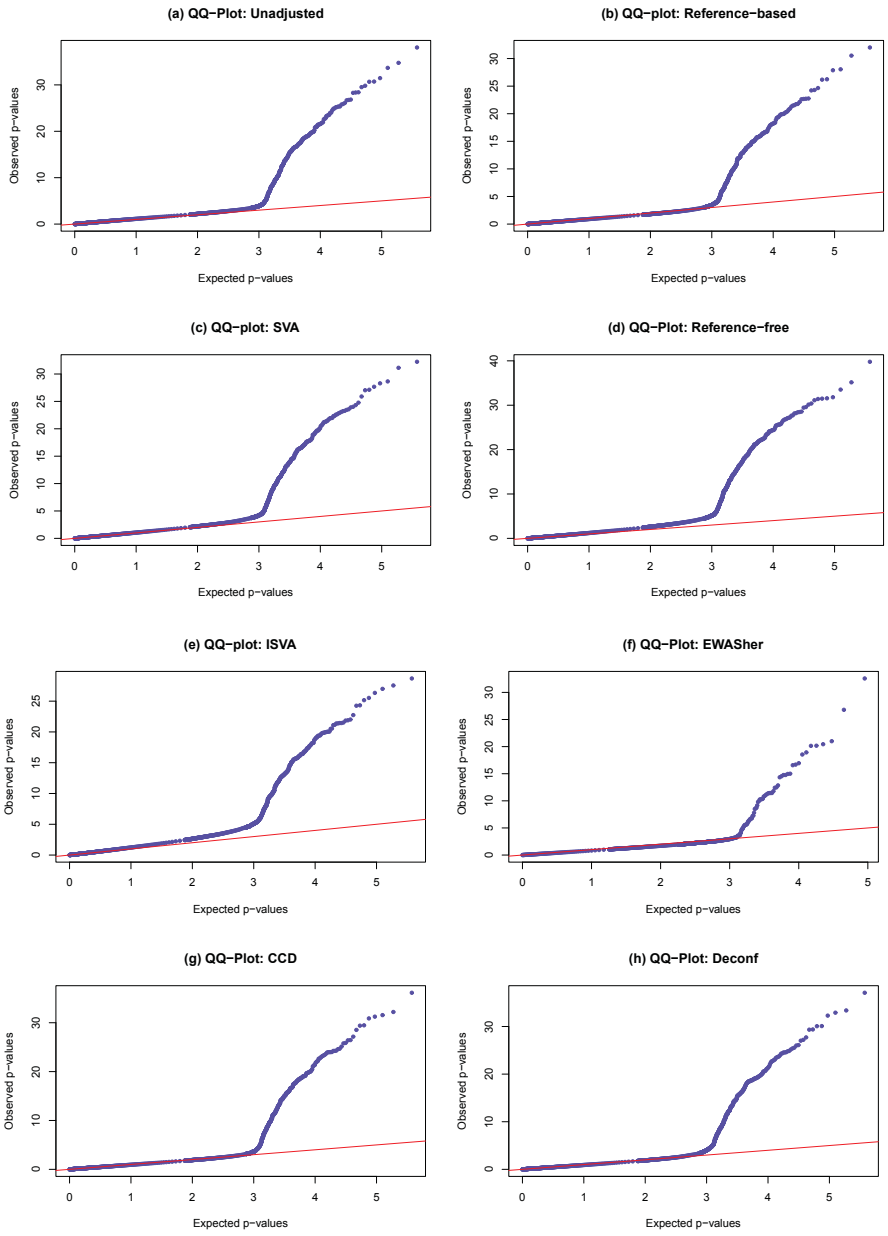


Figure 5–7: QQ-plots showing the distributions of p-values under the no confounding scenario for (a) no cell type adjustment, (b) reference-based adjustement, (c) reference-free adjustment, (d) SVA, (e) ISVA, (f) EWASher, (g) CellCDecon, and (h) Deconf. Note both axes are on the $-log_{10}$ scale.

Table 5–3: Performance metrics comparison for the no confounding scenario

| Method | FDR (0.05 Thresh.) | Power | KS Stat. |
|---|---|---|---|
| Unadjusted | 0.0246 | 0.632 | 0.059 |
| Ref-based | 0.0171 | 0.572 | 0.0296 |
| Ref-free | 0.499 | 0.702 | 0.1591 |
| SVA | 0.120 | 0.628 | 0.0148 |
| ISVA | 0.551 | 0.654 | 0.0613 |
| EWASher | 0 | 0.122 | 0.0662 |
| CellCDec | 0.0126 | 0.626 | 0.0205 |
| Deconf | 0.0792 | 0.628 | 0.0313 |

the following parameters: $\mu_{T,overall} = 0.75$, $\mu_{Mono,overall} = -0.75$ and $\sigma^2_{T,overall} = \sigma^2_{Mono,overall} = 0.1^2$.

The generated cell effect distributions can be seen in Figure 5–8. T-cells have been assigned a net positive effect, and monocytes a net negative effect, with the sign 'flipped' between the average for each cell type. The variances were chosen to be equal.

Next we present in Figure 5–9 the QQ-plots showing the distribution of p-values for each of the adjustment methods. The unadjusted model does not show signs of confounding here. The reference-free method appears to have slighly inflated p-values. The other methods seem to be as one would expect, except for FaST-LMM-EWASher, which has overcorrected. Once again, it will be necessary to look at the numerical performance metrics.

The realized values of the performance metrics are seen in Table 5–4. Because there were not as many effects around zero, the FDR for several of the methods (including unadjusted) was zero. More specifically, there are likely more large associations than before, which makes it less likely that non-differentially methylated

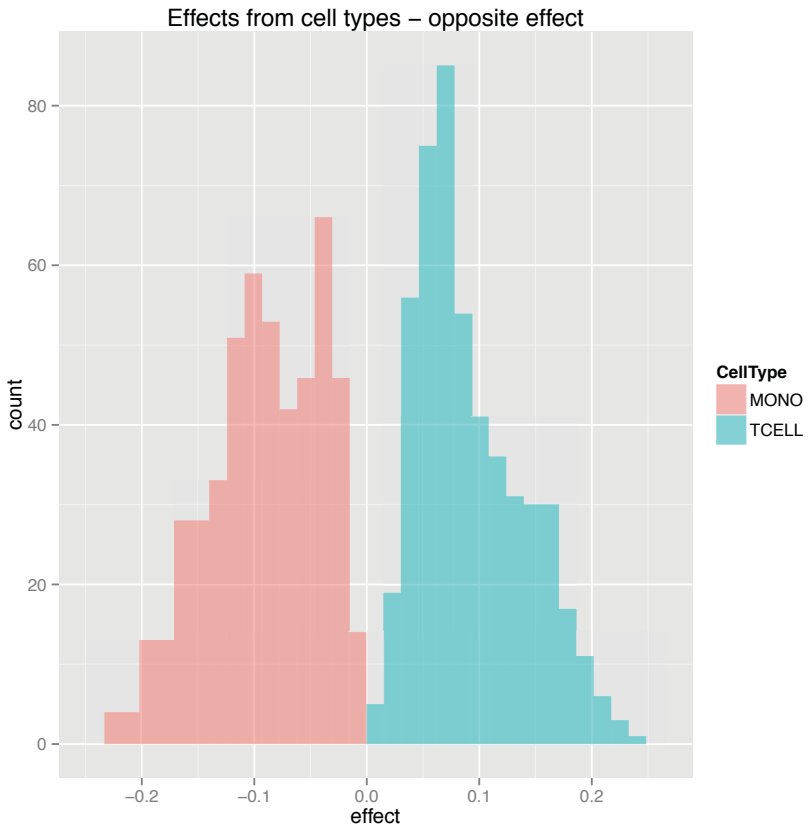


Figure 5–8: Distributions of simulated effect sizes at the 500 chosen CpG sites for different cell types. Blue represents the distribution for T-Cells and red represents the distribution for monocytes. The two sets of effects were generated from the distributions with the same variance, but means with the opposite sign.

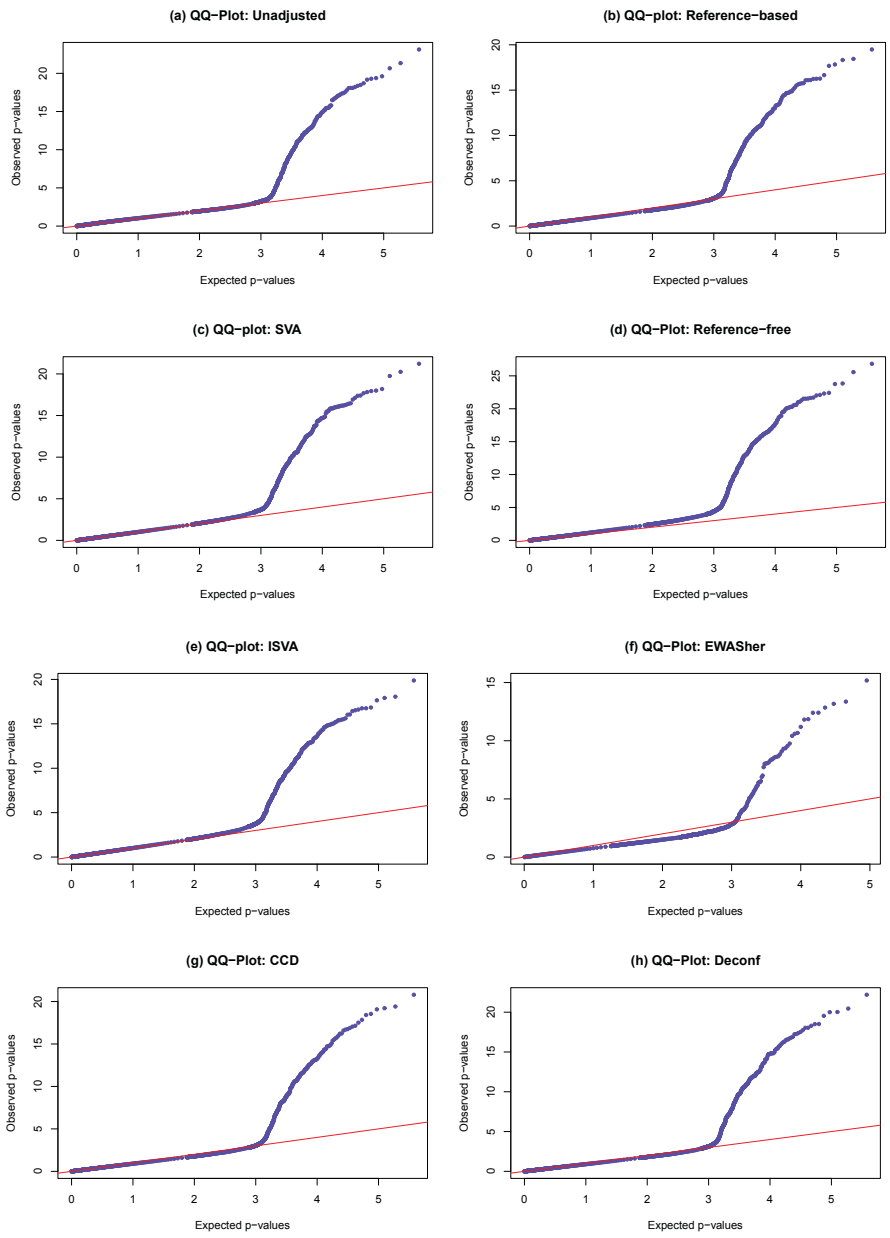


Figure 5–9: QQ-plots showing the distributions of p-values under the opposite effect scenario for (a) no cell type adjustment, (b) reference-based adjustement, (c) reference-free adjustment, (d) SVA, (e) ISVA, (f) EWASher, (g) CellCDecon, and (h) Deconf. Note both axes are on the $-log_{10}$ scale.

positions that happen to correlate with the phenotype are declared signifcant after performing FDR adjustment. However, despite these large effects, power is still lower for each method than it was in the no confounding scenario. With the exception of the reference-free method, all the adjustment methods achieve a higher power than the unadjusted model. Additionally, the p-values for all methods except reference-free do not show a large departure from the null distribution.

Table 5–4: Performance metrics comparison for the no confounding scenario

| Method | FDR (0.05 Thresh.) | Power | KS Stat. |
|---|---|---|---|
| Unadjusted | 0 | 0.462 | 0.0408 |
| Ref-based | 0 | 0.484 | 0.0362 |
| Ref-free | 0.259 | 0.594 | 0.1529 |
| SVA | 0.0582 | 0.55 | 0.0034 |
| ISVA | 0.0521 | 0.546 | 0.0063 |
| EWASher | 0 | 0.108 | 0.0915 |
| CellCDec | 0.0083 | 0.48 | 0.0326 |
| Deconf | 0 | 0.492 | 0.0346 |

### 5.1.4 Scenarios 4 and 5: High vs. Low Precision of Cell Type Heterogeneity

The final two scenarios that have been simulated change the precision of the individuals' cell type distributions between cases and controls. That is, a higher precision corresponds to a more pronounced divide in cell type distributions between cases and controls, while a lower precision makes the two more difficult to distinguish. In this simulation, both T-cells and monocytes were chosen to have distinct, positive net association with the phenotype. However, we now adjust the precision parameter $\rho$ from the Dirichlet distribution that simulates cell type composition. The low precision situation corresponds to $\rho = 10$, and the high precision situation

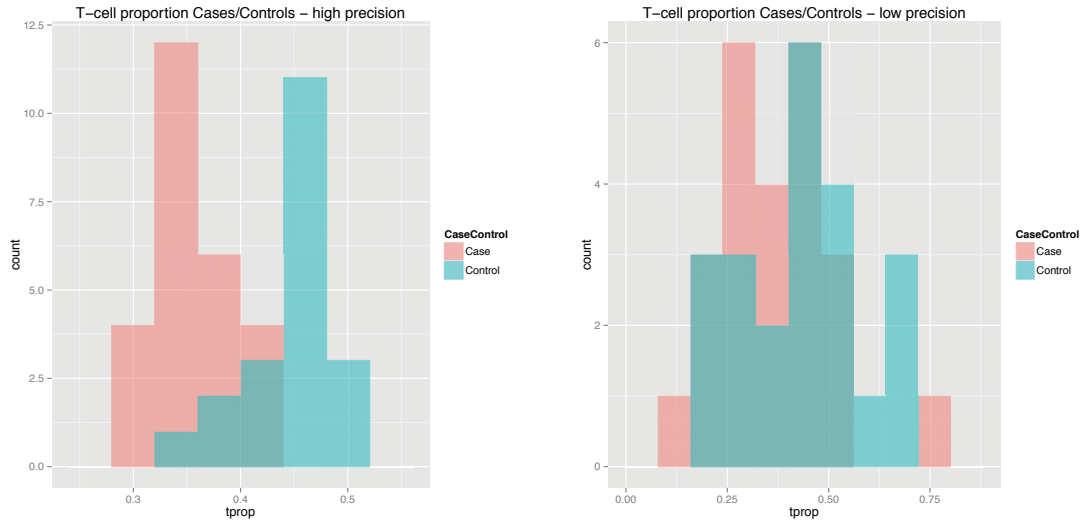


Figure 5–10: The distributions among subjects of T-cell proportions. The left plot shows the high precision scenario, and the right plot shows the low precision scenario. Green represents controls, red represents cases, and dark green represents where the distributions overlap.

corresponds to $\rho = 200$. The resulting distributions for T-cells among cases and controls can be observed in Figure 5–10.

In this section the results will be evaluated through the numerical performance metrics as the number of p-value plots here would be overwhelming. The results can be seen in Table 5–5 and Table 5–6.

In the high precision simulation, FDR is generally reduced. Power is improved in some cases, but reduced in others. Looking at the KS statistic shows all the methods, except reference-free, correct towards the null distribution.

In the low precision simulation, FDR is generally increased. Power is improved in some cases, and reduced in others. The KS statistic shows that for the methods: reference-based, reference-free, and Deconfounding, there is not much correction towards the null distribution. It seems the methods have a more difficult time performing the adjustment when the cell type compositions between cases and controls are not clearly differentiated.

Table 5–5: High precision - performance metrics

| Method | FDR (0.05 Thresh.) | Power | KS Stat. |
|---|---|---|---|
| Unadjusted | 0.524 | 0.594 | 0.1423 |
| Ref-based | 0.161 | 0.406 | 0.0432 |
| Ref-free | 0.564 | 0.698 | 0.1363 |
| SVA | 0.136 | 0.494 | 0.0318 |
| ISVA | 0.297 | 0.482 | 0.0618 |
| EWASher | 0.118 | 0.06 | 0.0383 |
| CellCDec | 0.0781 | 0.472 | 0.0312 |
| Deconf | 0.0927 | 0.47 | 0.0138 |

Table 5–6: Low precision - performance metrics

| Method | FDR (0.05 Thresh.) | Power | KS Stat. |
|---|---|---|---|
| Unadjusted | 0.141 | 0.586 | 0.241 |
| Ref-based | 0.456 | 0.636 | 0.1975 |
| Ref-free | 0.686 | 0.7 | 0.209 |
| SVA | 0.201 | 0.614 | 0.0911 |
| ISVA | 0.0876 | 0.396 | 0.0404 |
| EWASher | 0 | 0.042 | 0.0299 |
| CellCDec | 0.243 | 0.552 | 0.0082 |
| Deconf | 0.249 | 0.62 | 0.2038 |

## 5.2 Using methods on ARCTIC data set

Here we present the results of trying the adjustment methods on the Assessment of Risk in Colorectal Tumors in Canada (ARCTIC) data. Information on these data can be found in [46], and key researchers include: TJ Hudson, BW Zanke, M Lemire, S Gallinger, and M Cotterchio. These data include 450K methylation measurements from 2203 subjects consisting of 1152 colorectal cancer patients and 1051 controls. The DNA was obtained mostly from lymphocyte pellets, except for a smaller number of samples coming from lymphoblastoid cell lines. The subjects included in our analysis include only those whose DNA was obtained from lymphocyte pellets. However, many samples were identified as poor quality after the QC step and, consequently, have not been included in the analysis. The final numbers of cases and controls in the analysis are 209 and 48, respectively.

The model applied here includes the case/control status as the phenotype of interest. Additional covariates included in the model are age and smoking status (binary). The resulting QQ-plots showing the distributions of p-values from the different adjustment methods can be seen in Figure 5–11. The unadjusted model shows a great departure from the null distribution, so there certainly could be some confouding due to cell type heterogeneity occuring.

The results for the other methods are rather interesting, several have not done much to force the p-values towards the null distribution. The ISVA method has certainly done some correction, as its smallest p-value is on the order of magnitude of $10^{-8}$ (as opposed to $10^{-24}$ in the unadjusted model). FaST-LMM-EWASher has brought almost all of the p-values towards the null distribution, which is unsurprising

as the algorithm does not stop until the genomic inflation factor has been sufficiently reduced.

Also shown are the distributions of the estimated effects (associations) over all CpG sites considered by each adjustment method (Figure 5–12). The Deconfounding and reference-based methods do not change effect sizes by much. The methods CellCDecon, ISVA, reference-based, and SVA all shrink the estimated associations towards zero. For FaST-LMM-EWASher, the distribution of effects has much heavier tails than that of the unadjusted model, however, there are fewer CpG sites still in consideration after the filtering step.

It is also interesting to look at whether the methods flag the same CpG sites as significant. There are no CpG sites that can be found simultaneously in the top 100 CpG sites selected by each method. In fact, the methods only selected three CpG sites in common when considering each method's top 1000 CpG sites. This number increases to 18 when EWASher is excluded. Next, we see whether the same CpG sites were being chosen in the unadjusted model and in the results each of the adjustment methods. Table 5–7 presents the number of CpGs each method has in common with the unadjusted model from among the top 100 sites for each method.

Table 5–7: Number of CpGs from top 100 in common with unadjusted model

| Method | CpGs in common |
|---|---|
| Ref-based | 76 |
| Ref-free | 16 |
| SVA | 52 |
| ISVA | 7 |
| EWASher | 1 |
| CellCDec | 40 |
| Deconf | 75 |

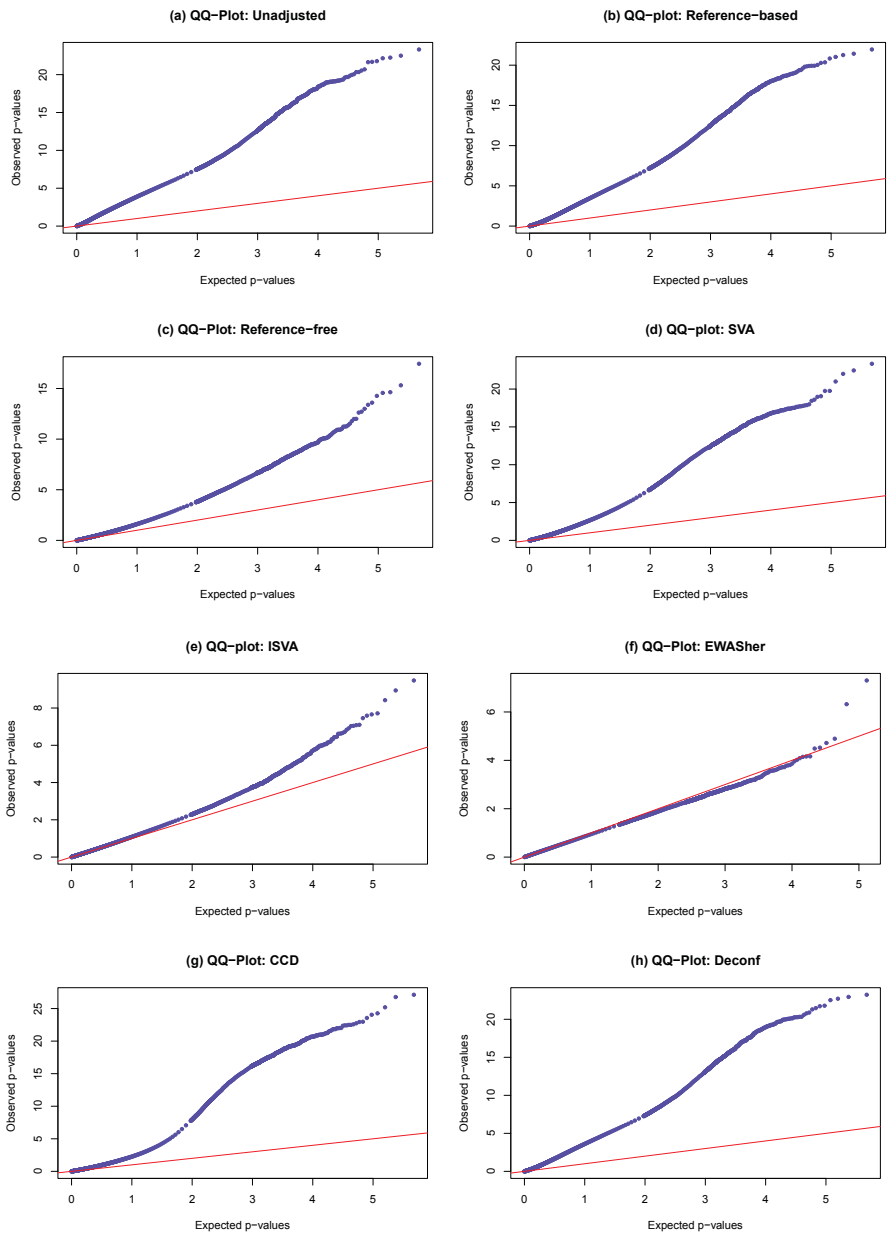


Figure 5–11: QQ-plots showing the distributions of p-values from the ARCTIC data analysis for (a) no cell type adjustment, (b) reference-based adjustement, (c) reference-free adjustment, (d) SVA, (e) ISVA, (f) EWASher, (g) CellCDecon, and (h) Deconf. Note both axes are on the $-log_{10}$ scale.

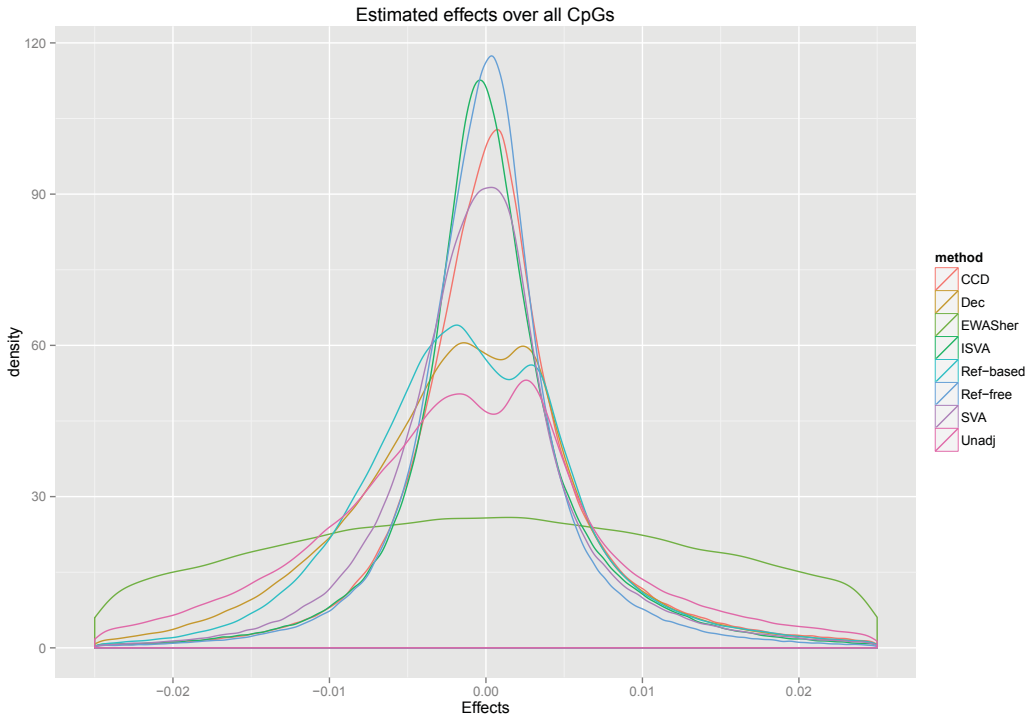


Figure 5–12: Estimated effect sizes (i.e. associations with phenotype) over all CpG sites considered by each adjustment method.

The two methods that appeared to do the most correction (as seen in the QQ-plots) also do not select the same CpG sites as the unadjusted model as being the most significant. This is particularly true for the FaST-LMM-EWASher method, which shares only one CpG site in common with the unadjusted model. EWASher's result is rather curious, as the other methods did not even come close to bringing the distribution of p-values towards the null, yet EWASher did so without issue. This begs the question: what if there truly were global change in methylation between cases and controls? The EWASher algorithm would continue until the genomic inflation factor (artificially) is pushed below one. The method would actually be removing variation that was associated with the phenotype.

### 5.3 Computing Performance

The completion times of the different algorithms will be compared in this section. The time complexities of the methods will be compared based on increasing sample size as well as increasing latent dimension (where applicable). A bit of thought is required to choose appropriate start/end points as some of the methods calculate coefficient and p-value estimates internally, while others require the use of an external function to perform a linear fit. To avoid this issue, the start time is defined to be when the adjustment method is first called, whereas the end time is when all estimates and p-values have been obtained.

The ARCTIC data are used as the benchmark data set. For the sake of preserving a large sample size, we include samples that were flagged as poor quality after the QC step. This should not be problematic as we are not interested in obtaining parameter estimates. The methylation measurements have, once again,

undergone Functional Normalization, and the phenotype of interest is chosen to be the case/control status. We also randomly choose a subset of size 10000 from the CpG sites to be considered. First, we look at how time to completion scales over sample size.

### 5.3.1 Scaling over Sample Size

We randomly sample from the patients in the ARCTIC study in order to obtain an array of different sample sizes. Ten sample sizes are chosen, and range from 50 to 500 in increments of 50. Figure 5–13 shows the resulting times.

The method that stands out the most is the reference-based method, as it finishes in significantly less time than the other methods. The method itself does not do any kind of decomposition or dimension estimation. Because it only performs a (relatively) simple regression model linking the validation data to the target data, the reference-based method is not as sensitive to increases in sample size as the others.

The completion times for the other methods are quite sensitive to increases in sample size—a fact that is unsurprising given the complexity of each algorithm. Table 5–8 shows the ratio of the time for the sample size 500 run to the time for the sample size 50 run; i.e. if time complexity were linear over sample size, one would expect this value to be about 10. The method whose time increased the most relative to its start time was ISVA. The method that had the smallest relative increase (other than reference-based) was Deconfounding, though overall the algorithm takes significantly longer than the others. EWASher did well overall, but its results are more difficult to interperet as the number of principal components it includes in the

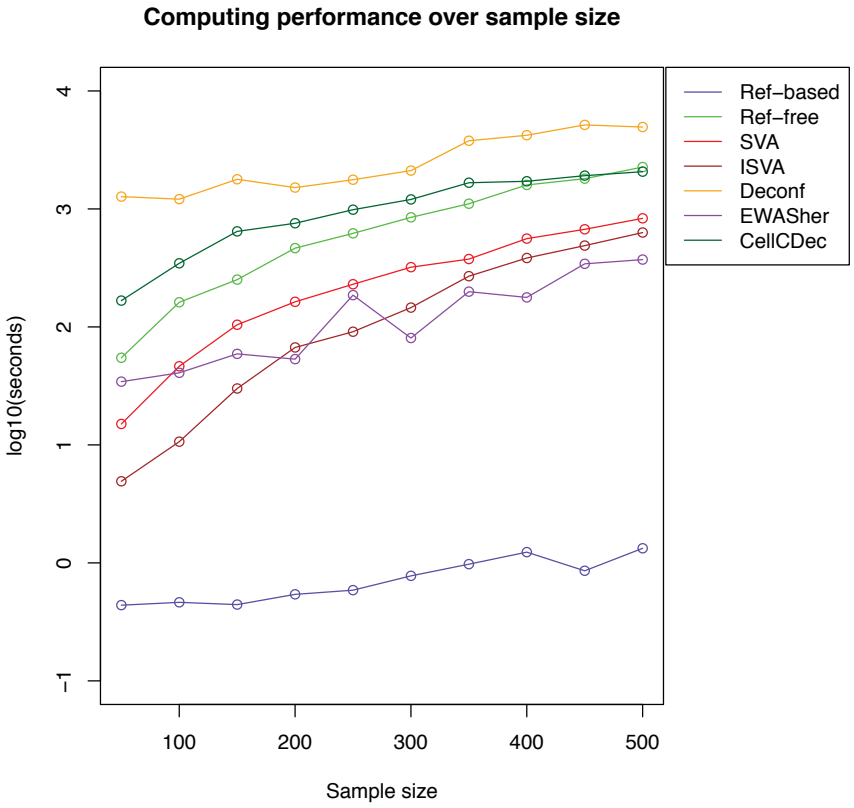


Figure 5–13: Computing performance for the different adjustment methods when considering sample size. Note the vertical axis is on the $log_{10}$ scale.

model can greatly affect the computing time. This phenomenon is easily observed as the curve for EWASher increases less steadily than the rest.

Table 5–8: Ratio of 500 sample size time to 50 sample size time

| Ref-based | Ref-free | SVA | ISVA |
|---|---|---|---|
| 3.03 | 41.37 | 55.35 | 128.36 |
| Deconf | EWASher | CellCDec | |
| 3.89 | 10.82 | 12.39 | |

### 5.3.2 Scaling over latent dimension

The sample size is fixed at 50 here, and the algorithms are run over different values of latent dimension (or, in some cases, assumed number of cell types present) for the methods that allow that parameter to be specified by the user. Each method was run 9 times, specifying the parameter to be 2 through 10, inclusive. Figure 5–14 shows the results.

The reference-free method and SVA are not very sensitive to increases in the latent dimension. ISVA showed a small increase. Both CellCDecon and Deconfounding showed great sensitivity to increasing values of latent dimension (or assumed number of cell types under the nomenclature of these two algorithms).

Table 5–9 shows the ratio of the runtime from latent dimension 10 to that of latent dimension 2. These figures confirm what was seen graphically: Deconf and CellCDecon are quite sensitive to increases in latent dimension.

Table 5–9: Ratio of latent dimension 10 time to latent dimension 2 time

| Ref-free | SVA | ISVA | Deconf | CellCDecon |
|---|---|---|---|---|
| 0.659 | 0.343 | 3.325 | 15.730 | 13.853 |

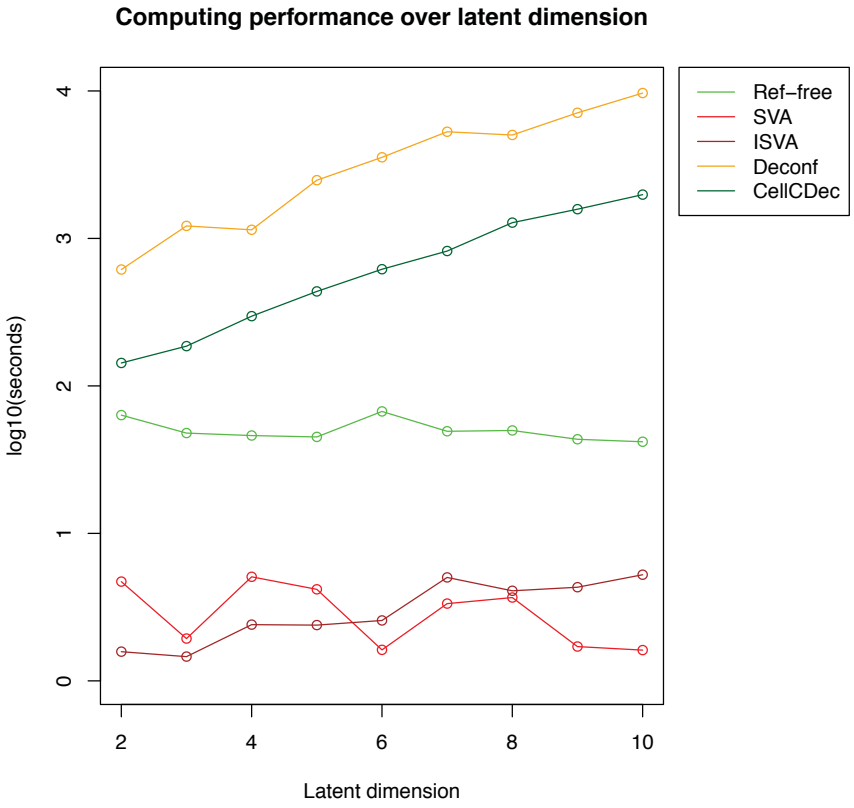


Figure 5–14: Computing performance for the different adjustment methods when considering latent dimension. Note the vertical axis is on the $log_{10}$ scale.

# CHAPTER 6
# Discussion and Conclusion

This review has shown that a number of things must be considered when deciding which adjustment method to use. Though some perform noticeably better than others, there are some limitations that must be taken into account. Each of the methods will be summarized below:

## 6.1 Summary of methods

**Reference-based method.** The method is very easy to implement, and as seen in the computing performance section, it runs very quickly, even on larger sample sizes. It usually achieved good statistical power, and, with one exception, reduced the FDR from the unadjusted model. It also has the advantage of being able to directly estimate the cell type composition of each sample. One disadvantage is this method can only be used on methylation measurements coming from cell mixtures for which there is an external, cell separated methylation data set available. However, if such a data set is available, the reference-based method can be a very powerful tool.

**Reference-free method.** This method is also fairly easy to implement. It generally resulted in increased power compared to the unadjusted model. However, it can be seen in the previous chapter that this increase in power was also accompanied by a significant increase in the FDR. Looking at the KS statistic shows it also did not adjust for genomic inflation as well as the other methods.

**Surrogate Variable Analysis.** SVA generally performed well. Its power was on par with the other methods, and in the distinct associations in cell types scenario it drastically reduced the FDR. It was also very easy to implement.

**Independent Surrogate Variable Analysis.** ISVA was very easy to implement. It generally achieved as high power as the other methods, and in the no confounding scenario actually improved power (contrary to some of the other methods), though the FDR was, admittedly, greatly increased. It was also one of the only methods to account for p-value inflation in the ARCTIC data set. However, its computing time was quite sensitive to increases in sample size.

**FaST-LMM-EWASher.** EWASher is, without a doubt, the most interesting method to review. In every case it seemed to do a very good job in reducing p-value inflation. However, the fact that it forces the genomic inflation factor towards one makes one wonder if that is truly the way to go about adjusting for cell type heterogeneity. If there were, say, global hypermethylation associated with a disease, adjustment using EWASher would be overly conservative.

Additionally, part of the algorithm involves filtering out loci that are unilaterally high or low among all subjects. The assumption here is that these loci are, for all intents and purposes, completely methylated or unmethylated and any associations between these probes and the phenotype are not interesting. At this time, the justification behind this assumption is not clear.

The final caveat in the EWASher method is it is quite difficult to implement. All the above methods can be run inside R, but EWASher requires the user to output three separate files, which are then used as input to an executable. Any further

analysis requires the user to take the output from this executable (containing the final estimates of associations and their respective p-values) to load back into R. The process becomes further complicated as the portion of the program contained in R required some editing, as the version provided on the internet returned error messages.

**CellCDecon.** The CellCDecon method generally achieved good power, and managed to reduce the FDR. It exists as a C++ program, and was quite easy to implement. The run time was longer for this algorithm than the others, and it was sensitive to increasing the assumed number of cell types. Additionally, it would be interesting to see how this program would perform if it took the phenotype and other covariates into account. The need to specify the number of cell types in this method is another issue, as this is not generally something known beforehand.

**Deconfounding.** The Deconfounding method was comparable to the other methods in terms of power and FDR. The biggest issue with the method was the running time. In all cases, it took longer to run than the other adjustment methods. It was also sensitive to both increases in sample size and number of cell types. Akin to CellCDecon, the fact that it does not internally estimate the number of cell types is an issue.

It seems that, because of its good performance, ease of use, and low computation time, the reference-based method is the way to go. Of course, this is only true if cell specific methylation measurements are available for the types of cells present in the sample. If this is not the case then the next best choice would be ISVA. Overall, ISVA

did a very good job when considering the numerical performance metrics, and was one of the only methods that managed to adjust for genomic inflation in the ARCTIC data set. Though the computational time increases significantly over sample size, it remained the third fastest method at sample size 500. Being in an easy to use R package also makes it a desirable method.

One concerning result is that the top CpG sites chosen by each method tend to differ substantially. In the first scenario of the simulation it was apparent that EWASher was the "odd one out", however, in the ARCTIC data, removing EWASher from consideration resulted in only 18 CpG sites in common among the top 1000 CpG sites chosen by each method.

## 6.2 Limitations

One of the biggest challenges in this project was data quality in the methylation data separated for blood cell types. Even after implementing the steps to address batch effect (as outlined in Section 4.1.2) there was still some kind of confounding occurring. In fact, in the study design, disease subtype (the autoimmune diseases) was confounded by batch and chip, making it impossible to statistically separate the technical factors from biological effects among these diseases. These other confounders could very well be the reason the reference-free method, SVA, and ISVA, all estimated the latent dimension to be so high. The potential presence of these confounders makes our results more difficult to interpret. Nevertheless, the fact that these methods are capable of detecting this confounding is encouraging, as there could be important confounders in epigenetic data that the investigator did not take into consideration.

Another limitation is the fact that only two cell types were used in our simulation. This was the only viable option as there were not enough good quality B-cell samples to include a third cell type analysis. Of course, in true whole blood samples, one would expect many more cell types to be present. The results of my simulation are really only applicable if the total number of cell types in the sample is small (e.g. sample of lymphocytes).

Finally, the ARCTIC data set contained many samples flagged as poor quality in the QC step. This was of particular concern as the sample size for controls was reduced from over one thousand to a mere 48 samples. It is evident that more investigation is required to decipher why so many controls were not usable. This makes one question whether there could be additional data quality problems in the remaining samples.

### 6.3 Future work

There are a number of potential directions one could go with this work. If measurements from more kinds of separated cell types were available, it would be advanageous to measure the methods' performances for more than two cell types. It would also be interesting to see if any of the adjustment methods could be modified for use in more advanced models such as random effects models. In addition, one could allow the cell type mixtures to vary with factors that have epigenome-wide effects on methylation, such as age or smoking status. Finally, since some of the existing methods are not easy to run, software could be improved to better streamline the process so that a given method is performed via a single function call.

## Appendix A - Singular Value Decomposition

Singular value decomposition (SVD) is a common technique in the literature in statistical genetics and genomics. It is related to principal component analysis in that it allows us to uncover underlying structures in data that would otherwise be difficult to visualize. Several sections in Chapter 3 will reference SVD as it is, undoubtedly, an indispensable tool. Here is the definition as found in [41]: Let $X$ be an $m \times n$ matrix. Then the singular value decomposition of $X$ is:

$$X = USV^\top$$

where $U$ is an $m \times n$ matrix, $S$ is a diagonal $n \times n$ matrix, and $V$ is an $n \times n$ matrix. The columns of $U$ are called the "left singular vectors" and the columns of $V$ are called the "right singular vectors". The diagonal elements of $S$ are called "singular values". Interestingly, in the context of genomics, the individual elements of the SVD are given unique names—a fact that underlines the importance of SVD in these kinds of analyses. The right singular vectors are referred to as 'eigengenes' and the left singular vectors, 'eigenarrays'. Additionally, Alter et al. note that the singular values characterize the "relative significance of the [corresponding] eigengene and eigenarray in terms of the fraction of the overall expression that they capture" [1]. Some of the methods presented in this thesis do indeed make use of the singular values in order to identify significance among an abundance of information.

## Appendix B - Available Software

Table 6–1: Available software

| Method | Type | Source |
|---|---|---|
| Ref-based | R | http://people.oregonstate.edu/~housemae/software/TutorialLondon2014<br>http://bioconductor.org/packages/release/bioc/html/minfi.html |
| Ref-free | R | http://cran.r-project.org/web/packages/RefFreeEWAS/index.html |
| SVA | R | http://bioconductor.org/packages/release/bioc/html/sva.html |
| ISVA | R | http://cran.r-project.org/web/packages/isva/index.html |
| EWASher | R | http://research.microsoft.com/en-us/downloads/472fe637-7cb9-47d4-a0df-37118760ccd1 |
| Deconf | R | http://web.cbio.uct.ac.za/~renaud/CRAN |
| CellCDecon | C++ | https://github.com/jameswagner/CellCDecon |

References

[1] Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[2] Udo Baron, Ivana Turbachova, Alexander Hellwag, Florian Eckhardt, Kurt Berlin, Ulrich Hoffmüller, Paul Gardina, and Sven Olek. DNA methylation analysis as a tool for cell typing. *Epigenetics*, 1(1):56–61, 2006.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[4] Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, et al. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295, 2011.

[5] Marina Bibikova, Zhenwu Lin, Lixin Zhou, Eugene Chudin, Eliza Wickham Garcia, Bonnie Wu, Dennis Doucet, Neal J Thomas, Yunhua Wang, Ekkehard Vollmer, et al. High-throughput DNA methylation profiling using universal bead arrays. *Genome research*, 16(3):383–393, 2006.

[6] Adrian P Bird. CpG islands as gene markers in the vertebrate nucleus. *Trends in Genetics*, 3:342–347, 1987.

[7] Gregory W Corder and Dale I Foreman. *Nonparametric statistics for non-statisticians: a step-by-step approach*. John Wiley & Sons, 2009.

[8] Aimée M Deaton and Adrian Bird. CpG islands and the regulation of transcription. *Genes & development*, 25(10):1010–1022, 2011.

[9] Aimée M Deaton, Shaun Webb, Alastair RW Kerr, Robert S Illingworth, Jacky Guy, Robert Andrews, and Adrian Bird. Cell type–specific DNA methylation

at intragenic CpG islands in the immune system. *Genome research*, 21(7):1074–1086, 2011.

[10] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, 2010.

[11] Gerda Egger, Gangning Liang, Ana Aparicio, and Peter A Jones. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457–463, 2004.

[12] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[13] Jean-Philippe Fortin, Aurelie Labbe, Mathieu Lemire, Brent W Zanke, Thomas J Hudson, Elana J Fertig, Celia MT Greenwood, and Kasper D Hansen. Functional normalization of 450K methylation array data improves replication in large cancer studies. *Genome biology*, 15(12):503, 2014.

[14] Jerry Guintivano, Martin J Aryee, and Zachary A Kaminsky. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*, 8(3):290–302, 2013.

[15] Ben John Hayes, Peter M Visscher, and Michael E Goddard. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91(01):47–60, 2009.

[16] Steve Horvath. DNA methylation age of human tissues and cell types. *Genome biology*, 14(10):R115, 2013.

[17] Eugene A Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13(1):86, 2012.

[18] Eugene Andres Houseman, John Molitor, and Carmen J Marsit. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439, 2014.

[19] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[20] Andrew E Jaffe and Rafael A Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*, 15(2):R31, 2014.

[21] Peter A Jones and Stephen B Baylin. The fundamental role of epigenetic events in cancer. *Nature reviews genetics*, 3(6):415–428, 2002.

[22] Hyunsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[23] Charles L Lawson and Richard J Hanson. *Solving least squares problems*, volume 161. SIAM, 1974.

[24] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

[25] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

[26] Christoph Lippert, Gerald Quon, Eun Yong Kang, Carl M Kadie, Jennifer Listgarten, and David Heckerman. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific reports*, 3, 2013.

[27] Baoshan Ma, Elissa H Wilker, Saffron AG Willis-Owen, Hyang-Min Byun, Kenny CC Wong, Valeria Motta, Andrea A Baccarelli, Joel Schwartz, William OCM Cookson, Kamal Khabbaz, et al. Predicting DNA methylation level across human tissues. *Nucleic acids research*, 42(6):3515–3528, 2014.

[28] Lisa D Moore, Thuc Le, and Guoping Fan. DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, 2012.

[29] Ruth Pidsley, Chloe CY Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C Schalkwyk. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC genomics*, 14(1):293, 2013.

[30] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luis A Nunes Amaral, Thomas Guhr, and H Eugene Stanley. Random matrix approach to cross correlations in financial data. *Physical Review E*, 65(6):066126, 2002.

[31] Dirk Repsilber, Sabine Kern, Anna Telaar, Gerhard Walzl, Gillian F Black, Joachim Selbig, Shreemanta K Parida, Stefan HE Kaufmann, and Marc Jacobsen. Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC bioinformatics*, 11(1):27, 2010.

[32] Keith D Robertson. DNA methylation and human disease. *Nature Reviews Genetics*, 6(8):597–610, 2005.

[33] Juan Sandoval, Holger Heyn, Sebastian Moran, Jordi Serra-Musach, Miguel A Pujana, Marina Bibikova, and Manel Esteller. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, 6(6):692–702, 2011.

[34] David L. Scott, Frederick Wolfe, and Torn W. J. Huizinga. Rheumatoid arthritis. *LANCET*, 376(9746):1094–1108, SEP-OCT 2010.

[35] Jalid Sehouli, Christoph Loddenkemper, Tatjana Cornu, Tim Schwachula, U Hoffmuller, A Grutzkau, Philipp Lohneis, Thorsten Dickhaus, J Grone, Martin Kruschewski, et al. Epigenetic quantification of tumor-infiltrating T-lymphocytes. *Epigenetics*, 6(2):236–246, 2011.

[36] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[37] Zhifu Sun, High S Chai, Yanhong Wu, Wendy M White, Krishna V Donkena, Christopher J Klein, Vesna D Garovic, Terry M Therneau, and Jean-Pierre A Kocher. Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC medical genomics*, 4(1):84, 2011.

[38] Andrew E Teschendorff, Joanna Zhuang, and Martin Widschwendter. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27(11):1496–1505, 2011.

[39] Timothy J Triche, Daniel J Weisenberger, David Van Den Berg, Peter W Laird, and Kimberly D Siegmund. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic acids research*, 41(7):e90–e90, 2013.

[40] James Wagner. *Computational approaches for the study of gene expression, genetic and epigenetic variation in humans*. PhD thesis, McGill University School of Computer Science, 2014.

[41] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

[42] Carola I Weidner, Qiong Lin, Carmen M Koch, Lewin Eisele, Fabian Beier, Patrick Ziegler, Dirk O Bauerschlag, Karl-Heinz Jöckel, Raimund Erbel, Thomas W Mühleisen, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome biology*, 15(2):R24, 2014.

[43] Georg Wieczorek, Anne Asemissen, Fabian Model, Ivana Turbachova, Stefan Floess, Volker Liebenberg, Udo Baron, Diana Stauch, Katja Kotsch, Johann Pratschke, et al. Quantitative DNA methylation analysis of FOXP3 as a new method for counting regulatory T cells in peripheral blood and solid tissue. *Cancer research*, 69(2):599–608, 2009.

[44] Yang Xie, Xinlei Wang, and Michael Story. Statistical methods of background correction for Illumina BeadArray data. *Bioinformatics*, 25(6):751–757, 2009.

[45] Vinod Kumar Yadav and Subhajyoti De. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Briefings in bioinformatics*, page bbu002, 2014.

[46] Brent W Zanke, Celia MT Greenwood, Jagadish Rangrej, Rafal Kustra, Albert Tenesa, Susan M Farrington, James Prendergast, Sylviane Olschwang, Theodore Chiang, Edgar Crowdy, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature genetics*, 39(8):989–994, 2007.

[47] James Zou, Christoph Lippert, David Heckerman, Martin Aryee, and Jennifer Listgarten. Epigenome-wide association studies without the need for cell-type composition. *Nature methods*, 2014.