Estimating Cell Type Proportions in Human Cord Blood Samples from DNAm Arrays

by

Louie Dinh

B. Sc., The University of British Columbia, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

The University of British Columbia
(Vancouver)

October 2017

© Louie Dinh, 2017
Abstract

Epigenome-wide association studies are used to link patterns in the epigenome to human phenotypes and disease. These studies continue to increase in number, driven by improving technologies and decreasing costs. However, results from population-scale association studies are often difficult to interpret. One major challenge to interpretation is separating biologically relevant epigenetic changes from changes to the underlying cell type composition. This thesis focuses on computational methods for correcting cell type composition in epigenome-wide association studies measuring DNa in blood. Specifically, we focus on a class of methods, called reference-based methods, that rely on measurements of DNa from purified constituent cell types. Currently, reference-based correction methods perform poorly on human cord blood. This is unusual because adult blood, a closely related tissue, is a case-study in successful computational correction. Several previous attempts at improving cord blood estimation were only partially successful. We demonstrate how reference-based estimation methods that rely on for cord blood can be improved. First, we validated that existing methods perform poorly on cord blood, especially in minor cell types. Then, we demonstrated how this low performance stems from missing cell type references, data normalization and violated assumptions in signature construction. Resolving these issues improved estimates in a validation set with experimentally generated ground truth. Finally, we compared our reference-based estimates against reference-free techniques, an alternative class of computational correction methods. Going forward, this thesis provides a template for extending reference-based estimation to other heterogeneous tissues.
Lay Summary

All cells within the human body generally contain the same genetic material in the form of DNA. However, our bodies require cells to specialize and carry out the many different tasks related to maintaining a complex biological organism. One system that our body uses to facilitate these specializations is called DNA methylation. DNA methylation acts like a series of on-off switches that allow different cells to execute different parts of their shared genetic program. Some diseases leave clues about its’ underlying cause in the differing pattern of DNA methylation between healthy and diseased individuals. However, it is difficult to differentiate between patterns due to disease and patterns due to cellular specialization. There are computational methods for separating these two types of patterns, but they do not work well for umbilical cord blood samples. This thesis studies how to improve such methods.
Preface

This dissertation is an original intellectual product of the author, Louie Dinh. The author conducted all experiments and wrote the manuscript under the supervision of Dr. Sara Mostafavi and Dr. Raymond Ng.
# Table of Contents

Abstract ................................................................. ii

Lay Summary ........................................................... iii

Preface ................................................................. iv

Table of Contents ................................................... v

List of Tables ........................................................ viii

List of Figures ........................................................ ix

Glossary ............................................................... xi

Acknowledgments ..................................................... xiii

Dedication ............................................................. xiv

1 Introduction ......................................................... 1
  1.1 Epigenome Wide Association Studies ......................... 2
    1.1.1 Epigenetics and DNAm .................................. 2
    1.1.2 Measuring DNAm with Microarrays .................. 3
    1.1.3 Blood is a Complex Tissue ......................... 4
  1.2 Cell Type Heterogeneity in Association Studies ............ 5
    1.2.1 Confounding Due to Cell Type Heterogeneity ....... 5
    1.2.2 Computationally Correcting for Confounding ....... 6
1.3 Thesis Motivation ........................................... 6
1.4 Approach and Contribution ................................. 8
1.5 Detailed Outline of Thesis ................................. 8

2 Related Works ............................................. 10
   2.1 The Linear Mixture Model for Complex Tissues .......... 10
   2.2 Reference-Based Methods ............................... 11
      2.2.1 Reference-Based Methods for Gene Expression ....... 12
      2.2.2 Reference-Based Methods for DNAm .................. 12
      2.2.3 An Adult Blood DNAm Reference .................... 14
      2.2.4 Three Cord Blood DNAm References ................ 14
   2.3 Reference Free Methods Applicable to DNAm .......... 14

3 Approach and Results ................................... 18
   3.1 Description of Datasets ................................. 19
      3.1.1 Reference Cell Type Profiles ....................... 19
      3.1.2 Validation ......................................... 21
   3.2 Evaluation Metrics ..................................... 21
   3.3 Validation of Existing Estimation Methods ............ 22
      3.3.1 Approach ......................................... 23
      3.3.2 Accuracy of Adult Reference Dataset ................ 24
      3.3.3 Effects of Using a Cord Specific Reference ......... 29
      3.3.4 Comparison of Cord and Adult References .......... 30
      3.3.5 Comparison of Three Cord Blood References ........ 34
   3.4 Normalization .......................................... 39
      3.4.1 Approach ......................................... 39
      3.4.2 Within Array Normalization .......................... 42
      3.4.3 Between Array Normalization ........................ 43
      3.4.4 Cluster Analysis of Full Normalization Pipeline .... 47
   3.5 Signature Selection .................................... 47
      3.5.1 Approach ......................................... 47
      3.5.2 Balancing of Highly and Lowly Methylated Probes ... 50
      3.5.3 Finding an Optimal Signature Size .................. 51
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5.4 Treating T-cells as Indistinguishable</td>
<td>52</td>
</tr>
<tr>
<td>3.6 Evaluation</td>
<td>53</td>
</tr>
<tr>
<td>3.6.1 Approach</td>
<td>53</td>
</tr>
<tr>
<td>3.6.2 Validating Estimation Accuracy in Cord Blood</td>
<td>55</td>
</tr>
<tr>
<td>3.6.3 Changes to the Cell Type Signatures</td>
<td>57</td>
</tr>
<tr>
<td>3.6.4 Comparison to Reference-Free Techniques</td>
<td>62</td>
</tr>
<tr>
<td>4 Conclusions</td>
<td>69</td>
</tr>
<tr>
<td>Bibliography</td>
<td>72</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1</td>
<td>Count of Purified Reference Cell Type Profiles</td>
<td>20</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Number of Differentially Methylated Positions Between Cell Types in Cord</td>
<td>34</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Summary of Evaluated DNAm Normalization Methods</td>
<td>40</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Jaccard Index of Signature Probes for Standard and Optimized Normalization</td>
<td>58</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>Pearson Correlations of Signature Probe Ranks Between Standard and Optimized Pipeline</td>
<td>60</td>
</tr>
<tr>
<td>Table 3.6</td>
<td>sPCA Explained Variance at 500 Representative Probes</td>
<td>64</td>
</tr>
<tr>
<td>Table 3.7</td>
<td>sPCA Explained Variance Across All Probes</td>
<td>65</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Overview of Evaluated Reference-Validation Pairings.</td>
<td>22</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Estimation of Adult Validation Samples Using Adult Reference. Black line is x=y.</td>
<td>25</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Estimation of Cord Validation Samples Using Adult Reference. Black line is x=y.</td>
<td>26</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Comparison of Adult Reference Performance by: (A) Correlation, (B) MAD.</td>
<td>27</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Comparison of Adult Reference Performance After Subsampling on Cord Validation Set: (A) Correlation, (B) MAD.</td>
<td>28</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>Estimation of Cord Validation Samples Using Cord Reference (UBC).</td>
<td>29</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>(A) Clustering Between Cord Cell Types on Signature Sites using Euclidean Distances (B) Sensitivity to Missing nRBC in Reference Profile.</td>
<td>31</td>
</tr>
<tr>
<td>Figure 3.8</td>
<td>MDS Plot of Adult and Cord Reference Profiles.</td>
<td>32</td>
</tr>
<tr>
<td>Figure 3.9</td>
<td>Proportion of differentially methylated positions (DMPs) between Cell Types.</td>
<td>33</td>
</tr>
<tr>
<td>Figure 3.10</td>
<td>Number of Variable Sites in Reference Cell Type Profiles.</td>
<td>35</td>
</tr>
<tr>
<td>Figure 3.11</td>
<td>Correlation (Estimated vs. Measured) by Cord Blood Reference Set.</td>
<td>36</td>
</tr>
<tr>
<td>Figure 3.12</td>
<td>Comparison of Signature Probes Between Cord Reference Datasets.</td>
<td>37</td>
</tr>
<tr>
<td>Figure 3.13</td>
<td>Estimation Performance by Alternative Cord Reference.</td>
<td>38</td>
</tr>
<tr>
<td>Figure 3.14</td>
<td>SVD Analysis Raw Samples.</td>
<td>44</td>
</tr>
</tbody>
</table>
Figure 3.15  SVD Analysis of Background Subtraction and Probe Type Normalization Methods ................................. 45
Figure 3.16  P-value Inflation When Comparing Cord Blood Samples After (A) Quantile Normalization (B) ComBat .......................... 46
Figure 3.17  Hierarchical Clustering of Samples After Each Normalization ................................................................. 48
Figure 3.18  Effect Size of DMRs By Cell Type ................................................. 50
Figure 3.19  Estimation Performance by Signature Size. Solid is Mean, Dashed is median, Dotted is 10th and 90th quartile. .............. 51
Figure 3.20  Evaluating Effect of Treating Tcells as Indistinguishable. ................................................................. 52
Figure 3.21  cell type proportion (CTP) Estimation with Cord Optimized Pipeline. Line of unity shown in black. ...................... 56
Figure 3.22  Estimation Performance By Reference And Normalization ................................................................. 57
Figure 3.23  Performance Improvements Step by Step. ................................................................. 58
Figure 3.24  Association of PCs to Estimated Cell Types ................................................................. 59
Figure 3.25  Effect of Normalizations on Signature Probe Selection ................................................................. 60
Figure 3.26  Changes to Rank Of Signature Probes Between Normalizations ................................................................. 61
Figure 3.27  Rank of Probes Unique to a Normalization ................................................................. 63
Figure 3.28  Variance Explained by sPC1 Versus Measured Gran Proportions ................................................................. 65
Figure 3.29  Variance Explained by sPC1 Versus All Measured CTP ................................................................. 66
Figure 3.30  Variance Explained by Top 6 sparse principle components (sPCs) Versus All Measured CTPs. Red is line of unity. ............ 67
Figure 3.31  Difference to Variance Explained by Measured CTP ................................................................. 68
Glossary

BMIQ  beta mixture quantile normalization
CBC  complete blood count
CTP  cell type proportion
CTH  cell type heterogeneity
DNAm  DNA methylation
DMP  differentially methylated position
DMR  differentially methylated region
EH  expression heterogeneity
EWAS  epignome-wide association study
FACS  fluorescence activated cell sorting
GWAS  genome wide association study
ICA  independent components analysis
ISVA  independent surrogate variable analysis
ILBG  Illumina background correction
LMM  linear mixed model
MAD  mean absolute deviation
MDS  multidimensional scaling

NOOB  normal-exponential using out-of-band probes

PC  principle component

PCA  principle components analysis

RUV  remove unwanted variation

SWAN  subset within array normalization

sPC  sparse principle component

SD  standard deviation

SNP  single nucleotide polymorphism

SVD  singular value decomposition

SVA  surrogate variable analysis
Acknowledgments

First, I’d like to thank my supervisors Sara Mostafavi and Raymond Ng for their endless patience during my graduate studies. Without their wisdom, this document would not exist.

Next, I would like to thank Meaghan Jones for being a wonderful collaborator. Also, thank you to Magda Price, Kabor lab, and Robinson lab for lively discussions on the proper way to analyze methylation data.

Thank you to all the friends that I made throughout. You kept me from becoming a hermit. My fellow Mostafavi lab members: Bernard Ng, Halldor Thorhallsson, Farnush Farhadi, Emma Graham, and Hamid Omid. My computer science cohort: Kimberly Dextras-Romagnino, Meghana Venkatswamy, Dilan Ustek, Antoine Ponsard, Jacob Chen, Yasha Pushak, Robbie Rolin, Neil Newman, Clement Fung, Alistair Wick, and Kuba Karpierz. My adoptive lab and graduate program: Phil Richmond, Emma Hitchcock, Shams Bhuiyan, Sam and Amelia Hinshaw, Oriol Fornés and Robin van der Lee. Special thanks to Rachelle Farkas for chatting whenever I couldn’t be bothered to work.

I’d like to thank my oldest friends: Raymond Huang, Anthony Dong, Simon Tso, Derek Cho, Nelson Wang and Gilbert Leung.

Most of all, thank you to my family and to my partner. Many thanks to my parents and my sister for always being there. Finally, I’d like to thank my partner, Crystal Wong, for her help in this and all the things to come.
Dedication

To my parents
Chapter 1

Introduction

The question of how molecular changes manifest as observable phenotypes has been asked since the discovery of DNA as the biological mechanism for inheritance. With the advent of genome sequencing, researchers began probing for genetic changes, like single nucleotide polymorphisms (SNPs), that were enriched within populations with particular traits. This paradigm, called an association study, is increasingly popular as the cost of probing the genome continues to drop [32].

DNA microarray technology is one particular development that facilitated the popularity of association studies. While early microarrays were limited to reading nucleotide base pairs, their capabilities have since been extended to other genomic features. Among other things, microarrays can now be used to quantify gene expression profiles or epigenetic marks genome-wide at the population scale.

While cost-effective and convenient, microarrays are not without drawbacks. Multiple studies have identified key challenges to ensuring high quality microarray-based analyses [33, 35]. These challenges originate from factors like measurement error and shortcomings in statistical methodology [7, 16].

Researchers have overcome a number of these challenges, but one remaining issue is the confounding effects of cell type heterogeneity (CTH) in microarray association studies [16, 25]. The literature contains many algorithms to computationally correct for CTH, but the problem is not satisfactorily resolved in all cases. One particularly troubling case arises in the study of human infant cord blood DNA
methylation (DNAm). Cord blood is unusual because of its relatedness to adult blood, a case where CTH in DNAm studies has been successfully resolved [22]. In adult blood samples, effects due to CTH are successfully mitigated through estimation of each sample’s cell type proportion (CTP). But, it was observed that existing CTP estimation techniques applied to cord blood suffer a severe degradation in performance [53]. Several previous attempts to close this adult-infant gap through better characterization of cord blood’s constituent cell types did not fully succeed [5, 12, 19]. In this thesis, we investigate the reasons behind estimation performance degradation and present a pipeline for more accurately estimating cell type proportions for cord blood DNAm measurements.

1.1 Epigenome Wide Association Studies

Association studies are used to link patterns in the epigenome to human phenotypes and diseases. These studies are called epignome-wide association study (EWAS). Often, association studies target DNAm, an epigenetic mark, due to its’ response to environmental factors, role in gene regulation and implications in development [6]. These studies measure the DNAm level of up to millions of sites across the genome to find loci that exhibit a relationship to disease status or phenotype.

EWAS must employ large cohorts to detect small effects in a large number of measurements. This drives researchers towards accessible biological tissues like buccal or blood. These complex tissues, called complex tissues, consist of multiple cell types with distinct methylation profiles [45]. This mixing of distinct cell types makes inferring associations between methylation level and phenotypes a challenge.

1.1.1 Epigenetics and DNAm

Scientific disciplines currently disagree on the definition of epigenetics [13]. Some use the term to describe changes to gene expression, while others are explicitly referring to inheritance of expression patterns. In this thesis, we will follow the latter convention. Specifically, we use the term epigenetics to refer to mitotically heritable modifications to DNA that does not involve modifying the underlying base pair sequence [28 46]. One such heritable epigenetic mark, and the focus of
this thesis, is DNA methylation (DNAm).

DNA methylation is when a methyl group attaches to a DNA molecule. This process is crucial to development and involved in many biological processes like genomic imprinting, X-chromosome inactivation, and aging [28, 41]. In mammals, like humans, DNA methylation occurs almost exclusively at the DNA base Cytosine immediately followed by Guanine, denoted a CpG site. Functionally, high levels of CpG methylation in promoter regions have been shown to negatively correlate with gene expression in multiple species [21]. DNAm offers a mechanism by which somatic cells, all sharing the same DNA, can execute different genetic programs.

DNAm plays a large role in cell differentiation and aging [25, 28]. Patterns of DNAm are so closely tied to aging that it can be used to accurately predict biological age in humans [21]. In addition, DNAm is highly sensitive to environmental and psychosocial factors such as stress, car exhaust or neonatal exposure to maternal cigarette smoke [26, 34]. Thus, many studies analyze DNAm due to its key biological role, sensitivity to environmental factors, and heritability.

1.1.2 Measuring DNAm with Microarrays

To study DNAm’s role in gene regulation, aging and disease, we measure methylation levels across the genome. One platform for high throughput genome-wide measurement of DNAm is Illumina’s Infinium 450k. The 450k is a microarray based platform that interrogates the methylation status of 485,577 CpGs across the human genome [42]. The accuracy, sample requirements and relatively low cost make it the platform of choice for many EWAS studies [33].

To measure methylation at CpG sites, the 450k uses probes designed to hybridize with DNA from a specific genomic locus [14]. Hybridization is followed by an elongation step that causes differential fluorescence in methylated versus unmethylated sites. The amount of methylation is inferred from the strength of fluorescence. Due to the design of the 450k and inherent noise of physical experimentation, interpretation of DNAm data requires care [11, 33].

One complication arising from the 450k’s design is the distinction between Type 1 and Type 2 probes. To increase coverage of CpGs, the 450k includes two
types of probes; here denoted as Type 1 and Type 2. Type 1 probes use two different physical beads to measure DNAm at a CpG; one each for methylated and unmethylated states. Type 2 probes use one physical bead and competitively binds to methylated and unmethylated DNA. Furthermore, the two probe types use different binding chemistries. This distinction is similar to the difference between two-colour and one-colour gene expression microarrays. As a result, Type 2 probes are less sensitive in the detection of extreme methylation values, and have greater variance between replicates [15].

Once measured, these raw DNAm fluorescence signals are transformed into either Beta values or M-values to facilitate analyses [8, 17]. Beta values can be intuitively interpreted as proportion of DNA molecules methylated at a particular site, say $s$. Let $F$ be the measured intensity of the fluorescence due to methylated molecules and let $U$ be the measured intensity of fluorescence due to unmethylated molecules. Then

$$Beta_s = \frac{\max(F, 0)}{\max(F, 0) + \max(U, 0)}$$

Another representation used in DNAm analysis is the M-value. Beta values can be easily transformed to M-values as follows:

$$M_s = \log_2 \left( \frac{Beta_s}{1 - Beta_s} \right)$$

This logit transformation results in better statistical properties such as meeting Gaussian assumptions, approximate homoscedasticity, and no longer being restricted to the interval between 0 and 1 [17]. Such properties are desirable when performing statistical procedures like t-tests. In this thesis, both Beta values and M-values are used.

1.1.3 Blood is a Complex Tissue

As mentioned previously, one of the most common target tissues for EWAS is whole blood. While accessible, the heterogenous nature of blood complicates the interpretation of EWAS analyses. Human blood is a mixture of cell populations with distinct methylation profiles [45]. In this thesis, we focus on the 7 major cell types found in human infant cord blood: granulocytes (Gran), CD14+monocytes
Mono), CD4+ T-cells (CD4T), CD8+ T-cells (CD8T), CD19+ B-cells (Bcell), CD56+ natural killer cells (NK) and nucleated red blood cells (nRBC). Previous findings show that these cell types can be differentially methylated at over 20% of measured CpGs [45]. This differential methylation, paired with variability in CTP, can make interpretations difficult, confound statistical associations and cause spurious discoveries. The next section discusses statistical confounding in detail.

Cord blood and adult blood, while similar, must be treated as distinct tissues. Adult blood contains only 6 of the 7 cord blood cell types mentioned; nucleated red blood cells are unique to cord blood [12]. In adults, red blood cells do not contain nuclei and therefore do not contribute to the methylation measurements. In cord blood, red blood cells are still in the process of extruding their nuclei and many still contain genetic material. Previous assays of nRBC DNAm revealed an unusual methylation profile [12]. Their methylome did not exhibit a strong bimodal distribution like most cell types, and instead nRBCs had many intermediately methylated CpGs. In certain pregnancy complications, nRBCs can contribute up to 50% of the genetic material measured [2]. Studies on cord blood must account for this unusual cell type.

1.2 Cell Type Heterogeneity in Association Studies

Researchers have long recognized that varying cell type composition across samples can dramatically affect the interpretation of association studies [3]. DNAm studies are particularly susceptible to this type of confounding due to its tissue-specificity and highly variable nature [22, 34]. To compound the issue, CTPs in whole blood are not static over time. These proportions can change with environmental exposures, disease, and particularly age [25, 26]. Thus DNAm studies using blood samples from different ages must be particularly vigilant in correcting for inter-individual differences in CTP.

1.2.1 Confounding Due to Cell Type Heterogeneity

Systemic differences in cell type proportion has long been recognized as a source of DNAm variability [22, 35]. Left unaccounted for, CTP differences can lead to many false positive associations [25]. This problem is known as statistical con-
founding due to CTH. For convenience, we sometimes refer to the problem as just CTH.

CTH arises when comparing measurements from mixtures of cell types, when the underlying CTPs differ between samples. A detailed mathematical description of the problem can be found in Section 2.1.

Experimental techniques for purifying cell populations can be used to resolve CTH. For example, fluorescence activated cell sorting (FACS) can be used to isolate pure cell populations before taking DNAm measurements. Directly comparing cells of the same type eradicates CTH. However, experimental approaches suffers some limitations: affected cell types not known apriori, additional overhead, labour intensive, and cannot be performed post-hoc. These drawbacks make computational correction methods appealing.

### 1.2.2 Computationally Correcting for Confounding

Techniques for correcting CTH can be divided into two classes: reference-based CTP estimation and reference-free surrogate variables [38]. Reference-based methods seek to accurately estimate the proportions of constituent cell types in a sample. These methods require reference cell type profiles - experimental measurements of DNAm from purified constituent cell types. Reference-free methods do not require reference cell type profiles. However, this saving of experimental labour comes at the cost of interpretability. Reference-free methods result in surrogate, also called latent, variables that are a function of cell proportions rather than direct estimates.

Once computed, association studies can incorporate these reference-based estimates or reference-free surrogate variables in the same way [35]. Associations between DNAm and phenotypes are typically inferred using linear models. Specifically, a linear model is fit to each methylated position with the phenotype as an explanatory variable. Correcting for CTH is done by expanding this set of explanatory variables to include either the CTP estimates or the surrogate variables.

### 1.3 Thesis Motivation

Computational methods that accurately estimate cell type proportions in DNAm microarrays are convenient, highly interpretable and can be performed post-hoc
on DNAm EWAS. Methods correcting for blood samples are of particular interest due to their prevalence in population scale studies [25]. For adult whole blood, accurate methods for estimation of CTP exists. However, attempts to develop an analogous method for human infant cord blood, a closely related tissue, have shown a persistent degradation in performance [19, 53].

Currently, the most accurate methods for estimating CTP in adult whole blood rely upon the availability of reference cell type profiles. In 2013, Koestler et al. [30] used these reference-based methods to estimate CTPs for 94 adult samples with matched blood counts; correlation for monocytes and lymphocytes were 0.6 and 0.61 respectively. In 2016, Koestler et al. [31] used an improved method to estimate CTP for 6 adult samples with more detailed cell counts. They observed correlations over 0.99 for all cell types. However, it was observed that methods based on adult reference profiles failed to produce accurate estimates in cord blood [53]. For cord blood, Yousefi et al. [53] observed correlations for Monocytes and Lymphocytes of -0.01 and -0.03 respectively.

Since methylation of cord blood cell types are known to be distinct from adult blood cell types, there have been several attempts at remedying this situation by developing cell type reference profiles specific to cord blood [5, 12, 19]. However, even with cord blood specific reference profiles, estimation accuracy in cord blood was still low compared to their adult whole blood analog. In their 2016 study, Gervin et al. [19] estimated cell counts for 195 cord blood samples using cord blood reference profiles. They observed CTP correlations between 0.51 and 0.57 for low abundance cell types like Bcells and Monocytes. While better than adult references, cord blood estimation performance has room for improvement. Furthermore, this raises the concern that existing estimation methodologies cannot be extended to new tissues simply by characterizing that new tissue.

An understanding of the culprits behind this loss of accuracy would benefit all cord blood based DNAm EWAS. Furthermore, pinpointing culprits will assist in developing a better CTP estimation method for cord blood DNAm. Finally, as the application of EWAS broadens to new tissue types, this understanding will help extend reference-based CTP estimation methods to these new targets.
1.4 Approach and Contribution

The objective of this thesis is to investigate how to improve the low estimation accuracy of cell type proportions in DNAm array measurements from human infant cord blood. Specifically, we aimed to improve upon a reference based method that accurately estimates cell type proportions in adult whole blood. Previous attempts at solving this problem by generating cord specific cell type reference profiles still has poor estimation accuracy in lowly abundant cell types. We assessed each step of the estimation procedure, identified problematic steps, resolved each issue specifically for cord blood samples, and validated the improved estimation method.

First, we confirmed how the adult estimation procedure is unsuitable for infant cord blood. To do so, we estimated cell type proportions using the same adult-calibrated procedure for both adult and infant samples. Estimates were compared to experimentally measured proportions to corroborate previously reported results. Indeed, adult-calibrated procedures are unsuitable for use on cord blood samples.

Next, we explored how the same procedure, but with cord blood specific cell type profiles, improves estimation accuracy. We used three different sets of cord blood specific cell type profiles and observed consistently low estimation accuracy, especially for low abundance lymphoid cells. Therefore, we concluded that the low accuracy is partially caused by the procedure itself.

In order to improve estimation accuracy, we tailored the procedure to cord blood in a step-by-step fashion. A detailed outline of this thesis follows.

1.5 Detailed Outline of Thesis

The remainder of this thesis formalizes the problem of estimating cell type proportions in complex tissues [Chapter 2], surveys existing methods used to correct for cell type heterogeneity in association studies [Chapter 2], diagnoses the issues of applying existing methods to cord blood [Chapter 3], proposes resolutions and presents validation results for a new cord blood estimation pipeline [Chapter 3], and discusses directions for future work [Chapter 4].

A detailed chapter breakdown of this thesis is as follows:
• Chapter 2 surveys the existing literature on the problem of confounding due to cell type heterogeneity in association studies. First, we present a linear model generating the mixed signal observed in complex tissues. We show how this model leads to the two classes of solutions for CTH: reference-based and reference-free. We give reference-based estimation in DNAm a more formal description and trace its origins to early applications in gene expression. Previous works on improving estimation of CTP in cord blood are also described. Finally, we touch upon reference-free techniques both specific to DNAm and generally applicable to microarray studies.

• Chapter 3 presents the datasets, approach and experimental results used to improve the accuracy of CTP estimation in cord blood. First, we describe the two validation sets of mixed tissues and four reference sets of cell type profiles. We also describe how performance of CTP estimation was measured. Then we report results confirming previously reported results of low estimation accuracy using existing methods on our validation datasets. Next, we present a series of diagnostics used to determine the issues behind this low estimation accuracy. We show how resolving these issues lead to improved estimation accuracy. Finally, we detail a comparison between our reference-based estimates of cell type proportions to a PCA-based reference-free surrogate variables from the perspective of variance explained.

• Chapter 4 summarizes our results, identifies the limitations of reference-based estimation and discusses directions for future work.
Chapter 2

Related Works

In this chapter, we discuss prior works on correction of CTH in DNAm studies. Section 2.1 mathematically describes how measurements on a complex tissue can be modelled as a linear mixture. We show how the two classes of CTH correction techniques, reference-based and reference-free, arises naturally from this description. Section 2.2 summarizes the most successful reference-based methods for CTP estimation in blood. Section 2.3 summarizes existing reference-free techniques.

2.1 The Linear Mixture Model for Complex Tissues

In 2001, Venet et al. [52] presented the idea of computationally quantifying CTP directly from the mixed microarray measurements. This was originally presented in the context of gene expression. Here, we summarize that framework in the context of DNAm.

Assume that DNAm measurements were made on a mixture of distinct cell types using a microarray. Then, for each DNAm locus, the total measurement is the sum of signals from each cell type alone. We assume that the signal from each cell type is proportional to its relative abundance in the sample. This is called the linear mixing assumption, and it serves as a good model for the fluorescence intensities measured on microarrays [24].

Mathematically, we can represent the data generation process as follows:

- **L**: The number of measured mixed samples.
• **M**: A matrix of measurements from DNA methylation microarrays. One column per mixed sample and one row per probe.

• **G**: A matrix of reference cell type profiles. One column per cell type and one row per probe.

• **C**: A matrix of concentrations. One row per cell type and one column per mixed sample.

We assume that the mixed signal is generated only from the represented cell types in the columns of **G**:

\[ M = GC \]

Since each column of **C** represents concentrations of cell types in a sample, each entry is between 0 and 1, and each column must sum to 1. We recognize that each column of **M** to be a linear mixture of the columns of **G** with weights defined by the corresponding column of **C**.

In these microarray experiments, **M** is always observed. The problem of quantifying CTP can be specialized into whether **G**, or a noisy representation of **G**, is available. Note that if **C** was observed, then we already have cell type proportions for all mixed samples which solves our original problem. When **G** is available, we are dealing with reference-based methods. If **G** is not available, then we must rely on reference-free methods.

### 2.2 Reference-Based Methods

In the reference-based context, we have observed **G** and seek only to find an approximation of **C**, denoted \( \hat{C} \). This is done on a sample-by-sample basis. The algorithm proceeds in two steps:

1. **Signature Selection**: Select an appropriate subset of rows, corresponding to probes, of **G**. The selected probes are called the cell type signature probes and the resulting submatrix of **G** is called the cell type signature.

2. **Optimization**: For each column of **M**, denoted \( M_i \), solve the following for \( 1 \leq i \leq L \):

\[
\min_{\hat{C}_i} \| M_i - G\hat{C}_i \|_1
\]
such that \( \hat{C}_i \) constrained to the interval \([0,1]\) and elements sum to less than or equal to 1. Here the norm denotes some appropriate measure of distance between two vectors, for example Euclidian distance.

All reference-based estimation techniques rely on this formulation but differ in choice of signature selection heuristic, distance function, optimization procedure and enforcement of constraints on \( \hat{C}_i \).

### 2.2.1 Reference-Based Methods for Gene Expression

One of the early reference-based methods for cell type estimation from gene expression microarrays was presented by Abbas et al. [3]. Their method quantified the proportion of immune cell in adult whole blood samples. In the framework laid out above, signature selection was done by maximizing the condition number of the resulting \( G \) matrix, optimization was done with Euclidian distances, and the estimates had no enforced constraints. Their optimization is exactly the least squares method used in linear regression. As a post-hoc step, to get positive proportions, the process is run iteratively each time removing the most negative coefficient. Finally the results are normalized to sum to 1.

The least-square method was subsequently extended to explicitly enforce the constraints as part of the optimization. One extension used non-negative least squares to enforce the lower bound of the constraint to get estimated proportions greater than 0 [43]. Another extension used quadratic programming to ensure that the proportions were in the interval \([0,1]\) and each sample’s proportions summed to one [20]. More recent methods move away from the least squares framework to more robust methods like Support Vector Regression [39].

### 2.2.2 Reference-Based Methods for DNAm

Correcting for CTH in DNAm studies, while conceptually similar, requires acknowledgement of some DNAm-specific realities. In 2012, Houseman et al. [22] proposed a method for estimating cell type proportion from DNAm microarrays based. This method, once again, attempts to estimate the cell type proportions, \( \hat{C} \), that best approximates \( M \) with \( G\hat{C} \). Signature selection was done based on ordering F-statistics for CpGs. F-statistics were computed from independently fitting a
linear mixed-effects model to each probe in the reference data to identify between cell type differentially methylated positions (DMPs). The method uses Euclidian distances and quadratic programming to enforce constraints.

Houseman’s method was successfully applied to validation blood samples from a cohort of 94 healthy adult individuals [30]. The study was done using Illumina’s 27k technology, the predecessor to the 450k. Each sample was subject to complete blood counts (CBCs) and assayed on the DNAm microarray. The DNAm data, along with the reference data described in subsection 2.2.3, was used to estimate CTP. Since CBCs can only resolve cell types to the Lymphocyte, Monocyte and Granulocyte level, detailed estimates were aggregated appropriately. Estimates were shown to have low root-mean-squared-error (Lymphocytes: 5%, Monocytes:6%) and medium-high correlations (Lymphocytes:0.6, Monocytes:0.61). Thus, computational estimation of CTP in DNAm was demonstrated to be reasonably accurate in adult whole blood.

In 2014, Jaffe and Irizarry [25] extended Houseman’s method to the Illumina 450k. This thesis builds upon Jaffe’s method; so we carefully describe their approach here. Their method proceeds as follows:

1. **Remove Bad Probes**: Remove probes on the 450k that are known to be problematic due to SNPs in the hybridizing sequence.

2. **Signature Selection**: To form the matrix \( \mathbf{G} \), 100 probes were selected to distinguish each cell type. For each cell type, probes that were differentially methylated compared to all other cell types were identified using two-group t-tests. Probes with p-values \(< 10^{-8}\) were ranked in order of effect size. The 100 most differentially methylated probes were selected, balanced between highly methylated (most positive difference in mean methylation) and lowly methylated (most negative difference in mean methylation).

3. **Optimization**: Estimation of \( \hat{\mathbf{C}} \) is done by solving

\[
\min_{\hat{\mathbf{C}}} \| \mathbf{M}_i - \mathbf{G}\hat{\mathbf{C}} \|_2
\]
2.2.3 An Adult Blood DNAm Reference

Currently, the most widely used adult whole blood reference dataset for DNAm was published in 2012 [45]. Since then, it has been cited over 400 times. The dataset contains DNAm profiles for 6 cell populations in adult whole blood: Gran, Mono, Bcell, NK, CD4T and CD8T. In their comparison, Reinius et al. [45] showed that, between cell types, 85% of human genes have at least one differentially methylated probe. They conclude that interpretation of whole blood methylation measurements should be done with great caution.

2.2.4 Three Cord Blood DNAm References

As DNAm studies began focusing on cord blood, it became apparent that the existing CTP estimation methods were not performing well in cord blood samples. In 2015, Yousefi et al. [53] demonstrated that the Jaffe’s method, paired with the adult reference-data, had very poor prediction accuracy for blood samples from new borns. The suggested culprit was a mismatch between the adult originated reference data and infant originated target samples.

To rectify this problem, three different cord blood reference datasets were generated [5, 12, 19]. To date, only the dataset from The University of Oslo has been validated for CTP estimation accuracy. In their study, Gervin et al. [19] validated their reference dataset on an independent cohort of 195 individuals. Their results showed that using a cord-specific reference dataset improves correlation between estimates and cell counts, but low abundance cell types like Bcell and CD8T show only moderate correlations.

In the next chapter, we report the CTP estimation performance of these three references on a set of cord blood samples with matched cell counts. We recommend one reference based on performance and coverage of cell types. Then, we show how estimation performance can be further improved by modifying the reference-based procedure itself.

2.3 Reference Free Methods Applicable to DNAm

Reference-free methods allow for correction of CTH when there is no available reference dataset. That is, these methods estimate both $G$ and $C$, usually through
matrix decomposition. Instead of directly estimation cell proportions, reference-free methods return surrogate variables that are functions of the cell proportions. These surrogate variables, similar to CTP estimates, can be directly incorporated into downstream analyses to correct for CTH.

Reference-free methods must estimate both the profile, $G$, and the proportions, $C$, simultaneously from a set of DNAm microarray samples. Estimating both $G$ and $C$ is hard because there are many more loci measured by these microarrays than there are number of samples. Formally this is a highly under-determined problem. Also, reference-free methods will capture other systemic sources of variation, like batch effects, in addition to variation from CTH. Together, this makes reference-free results difficult to interpret directly.

In this section, we briefly outline the reference-free methods applicable to DNAm studies. First, we cover methods first developed for gene expression studies, but applicable to DNAm studies: SVA, ISVA, and RUV [18, 36, 48]. Then we summarize methods developed in the context of DNAm EWAS studies: LM-MEWasher, RefFreeEWAS, and ReFACTOR [23, 44, 54]. ReFACTOR is given a detailed treatment because it serves as the reference-free comparison in this thesis.

In 2007, Leek and Storey [36] described surrogate variable analysis (SVA), a SVD-based method for estimating unmodeled sources of variation. They called variation of expression due to these unmodeled sources expression heterogeneity (EH). Many of the following reference-free techniques build upon SVA. The idea is to capture shared sources of variation between different observations of gene expression or DNAm. Conceptually, SVA proceeds through 3 steps:

1. Remove the signal attributable to the main variables of interest to identify an orthogonal basis for EH.

2. Find a subset of measurements associated with each basis of EH.

3. Estimates the surrogate variables from the identified subsets using the original data.

These surrogate variables are then incorporated into downstream regressions to control for the effects of confounding.
In 2011, Teschendorff et al. [48] extends upon SVA from orthogonal surrogate variable to statistically independent surrogate variables. This is achieved by replacing singular value decomposition (SVD), which enforced orthogonality, with independent components analysis (ICA). They show that independent surrogate variable analysis (ISVA) is more effective in cases when confounding is uncorrelated with the primary variables of interest in a non-linear fashion.

Another method for capturing unmodeled variation was presented by Gagnon-Bartsch and Speed [18] in 2012 and relies upon prior of knowledge genomic loci unaffected by the primary variable. This method, called remove unwanted variation (RUV), restricts the estimation of surrogate variables to a-priori unaffected negative control loci. By restricting the loci under consideration, RUV mitigates the problem of overcorrecting for biological variation of interest.

In 2014, Houseman et al. [23] proposed a reference-free method called Ref-FreeEWAS. This method expanded upon SVA by including the estimated covariates of the unadjusted model for differential methylation in the decomposition. They show algebraically how this expanded matrix better models the linear mixing assumption. Their method out performed SVA when technical errors are small and variability is dominated by CTH.

Also in 2014, Zou et al. [54] described a reference-free approach based on linear mixed models (LMMs) called EWASher. Originally, LMMs were used to control association study test statistic inflation due to genetic relatedness among inbred strains of model organisms like mice Kang et al. [29]. By explicitly estimating the genetic relatedness of individuals, the model can better account for correlated measurements. To capture relatedness between samples, this approach computes pairwise methylome similarity between samples, and then includes this as the covariance component of the linear mixed model as a proxy for cell type composition. The differential expression model is fit to observe inflation of test statistics. If there is inflation, the process is run iteratively with increasing number of principle component until test statistic inflation is controlled.

In 2016, Rahmani et al. [44] described ReFACTOR, a reference-free method based on sparse principle components analysis (PCA). Intuitively, since cell type composition effects should be shared across many CpGs, ReFACTOR tries to find a subset of probes that are well represented by a low-dimensional approximation of
the observed samples. These resulting sparse factors should represent large scale
effects like variation due to CTH. The REFACTOR algorithm proceeds in 3 steps:

1. Find a k-rank approximation of the sample matrix. Call this matrix $\tilde{O}$.

2. Look for the top $d$ CpGs that are best approximated. If $O_i$ and $\tilde{O}_i$ represent
the $i^{th}$ row of $O$ and $\tilde{O}$ respectively, then find the $d$ rows that have the smallest
$\text{distance}(O_i, \tilde{O}_i)$.

3. Run PCA on the subset of $d$ sites from (2), and return the scores for the top
$k$ principle components.

The resulting principle components are the sparse factors that should be a func-
tion of the CTP and can be used in downstream correction of CTH.

Since reference-free methods do not rely upon experimentally generated cell
type profiles, they must make assumptions to bound the under-constrained solution
space. These assumptions can differ quite substantially between algorithms, and
usually an algorithm’s performance depends greatly upon how well its’ assump-
tions correspond to biological reality. For example, ReFACTOR and EWASher
assume that the top components of variation are caused by cell type composition.
In situations where this assumption is unfounded, these methods can overfit and re-
move true biological signal [46]. In contrast, SVA-based methods explicitly model
out variation associated with the phenotype of interest before decomposing the
residuals into surrogate variables. Thus, SVA-based methods no longer assumes
that the largest components of variation are due to cell type. However, SVA-based
methods rely upon having a well-specified model, which may not be available. In
summary, reference-free methods are applicable to a wider range of tissues, but
suffer from limitations like overfitting, unrealistic assumptions and model avail-
ability.

In the next chapter, we compare reference-based estimates to reference-free
surrogate variables in cord blood. Specifically, we examine the amount of variance
captured by CTP estimates versus surrogate variables. We show how ReFACTOR
is able to accurately model variation attributable to abundant cell types like Gran,
but tends to overcorrect when accounting for minor cell types.
Chapter 3

Approach and Results

This chapter details our approach to improving reference-based cord blood CTP estimation. We first describe our materials and evaluation criteria followed by three results: confirmation of low estimation performance on cord blood samples, diagnoses/resolution of problems, and evaluation of the improved method.

For the first result, we treat the CTP estimation technique as a black-box. We confirm the previously reported results that adult-ref is unsuitable for CTP estimation in cord blood samples. Then, we compare the estimation performance of three different cord blood references. We comment on how updating the reference to be cord blood specific still leaves room for improvements to estimation.

To diagnose the degraded performance, we critically examined the steps involved in CTP estimation. The estimation procedure proceeds by normalizing the data, constructing a signature for each cell type, and optimizing for CTP. Normalization consists of removing unreliable probes, removing noise associated with per-sample measurement, and bringing samples to a common measurement scale for comparison. We explored the space of normalization techniques and provide recommendations. Next, constructing a signature requires filtering for probes that can discriminate between cell types. We examined the measure of discriminability and tuned the size of the cell type signatures. Finally, we evaluated our improved pipeline by comparing against using the unmodified pipeline with adult reference data, the unmodified pipeline with cord reference data, and the reference-free technique ReFACTOR.
3.1 Description of Datasets

This thesis used six DNAm datasets obtained from blood samples. The datasets consists of four reference datasets and two validation datasets. Reference datasets contained DNAm measured in purified cell populations, called reference cell type profiles, used for CTP estimation. Validation datasets contained DNAm measurements from mixed blood, and had matching CTP quantification done experimentally. All samples had DNAm measured on the Illumina 450k microarray. Three reference profiles were from human infant cord blood and one was from human adult whole blood. There was one validation set each for cord blood and adult blood. A summary of all reference datasets can be found in Table 3.1. Each dataset is described in detail below.

3.1.1 Reference Cell Type Profiles

Reference-based estimation techniques rely upon DNAm measurements from purified cell type populations. Since human blood consists of multiple cell types, work must first be done to isolate specific cell types. Reference datasets tend to be quite small, with sample sizes ranging from six to fifteen profiles per cell type. These sample sizes are often limited by the experimental cost and requirements of isolating cell populations. To carry out cell population purification, experimenters ideally have access to fresh samples and highly specialized equipment. Our reference datasets relied on two isolation technologies: Fluorescence Activated Cell Sorting (FACS) and Magnetic Activated Cell Sorting. (MACS). While different in implementation, the output of both technologies are samples highly enriched for one cell type.

Adult Whole Blood Reference Cell Type Profiles

Our adult reference dataset, referred to as adult-ref, contained cell type profiles for six cell types: Gran, Mono, Bcell, CD8T, CD4T and NK. We downloaded adult-ref from bioconductor with R (FlowSorted.Blood.450k). These cell type profiles were isolated from blood samples donated by six healthy adult males. For full details see Reinius et al. [45].
Table 3.1: Count of Purified Reference Cell Type Profiles

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Blood</th>
<th>Gran</th>
<th>Mono</th>
<th>CD4T</th>
<th>CD8T</th>
<th>NK</th>
<th>Bcell</th>
<th>nRBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>UBC-ref</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Oslo-ref</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>JHU-ref</td>
<td>15</td>
<td>12</td>
<td>15</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>4</td>
</tr>
</tbody>
</table>

Infant Cord Blood Reference Cell Type Profiles

The three cord reference datasets, all publicly available, were generated by three different research groups. For convenience, we will refer to them by the name of their respective universities of origin. That is, we will refer to them as University Of British Columbia reference (UBC-ref), Oslo reference (Oslo-ref), and John Hopkins University Reference (JHU-ref). Mentions of cord blood reference refers to UBC-ref unless otherwise specified.

UBC-ref was obtained from seven cord blood samples obtained from elective caesarian deliveries at BC Women’s Hospital. Each sample was fractionated into seven cell types using FACS. This resulted in seven samples each of Gran, Mono, Bcell, CD4T, NK and nRBC. Due to poor sample quality, there were only six CD8T profiles. In addition, the infant cord blood samples were also assayed for DNAm prior to fractionation. For full details, refer to de Goede et al. [12].

JHU-ref was obtained from fifteen cord blood samples from full term, healthy vaginal births at John Hopkins Hospital. Fractionation was done using MACS and resulted in a variable number of profiles per cell type – see Table 3.1 For full details, refer to Bakulski et al. [5]

Oslo-ref was obtained from eleven cord blood samples from uncomplicated births at Oslo University Hospital. Fractionation was done using FACS. All samples were successfully fractionated, resulting in one profile each for Bcell, CD4T, CD8T, Gran, Mono and NK. Notably, nRBC was not isolated in this reference dataset. For full details, refer to Gervin et al. [19].
3.1.2 Validation

Two sets of samples, one each for cord and adult blood, was used to evaluate estimation performance of reference-based methods. We will refer to them as cord-validation and adult-validation respectively. All samples were measured for DNAm using the Illumina 450k as mixed tissue. These DNAm measurements are then used in cell type proportion estimation. In addition, each sample had cell counts experimentally generated. Normalized cell counts are considered ground truth against which estimation performance was evaluated.

**Adult Whole Blood Validation**

Adult-validation was obtained from GEO (GSE77797). This dataset contains six adult whole blood samples with matched cell counts generated from flow cytometry. See Koestler et al. [31] for full details.

**Infant Cord Blood Validation**

Cord-validation consists of 24 cord blood samples were obtained from deliveries at BC Women’s hospital. All samples were from term healthy elective caesarian deliveries. A small aliquot of each validation whole cord blood sample was sent to a pathology lab for CBC. A second aliquot was prepared as the reference samples above, with the same markers and antibodies, with a single exception: mononuclear cell fractions were run on a FACS machine for cell counting instead of sorting. Final counts were a combination of CBC and FACS data. CBC data provided nRBC, monocyte, and granulocyte counts, as well as total lymphocytes. These counts were scaled to total 1, then lymphocyte breakdown (relative proportions of Bcell, NK, CD4T, and CD8T from FACS data) was multiplied by the calculated total lymphocyte proportion to generate final lymphocyte numbers.

3.2 Evaluation Metrics

We used the validation datasets to evaluate estimation performance of all methods. For each sample, we computationally estimated the CTP. These estimates were compared against the experimentally generated CTP, which we considered
Figure 3.1: Overview of Evaluated Reference-Validation Pairings.

the ground truth. Performance was measured with Spearman correlation (Rho) and mean absolute deviation (MAD). Correlation is our primary measure of performance because downstream correction only requires accuracy up to a scaling factor. We use MAD as a secondary measure of performance. A method with both high correlation and low MAD implies good estimation of CTP in both magnitude and ordering. Such estimates can be used for qualitative insights in place of experimentally quantified cell type proportions. In this thesis, estimation performance without any qualification refers to Spearman correlation.

3.3 Validation of Existing Estimation Methods

To confirm the previously reported loss of accuracy, we estimated cell type proportion for both adult-validation and cord-validation. In this section, all estimates are from applying a previously validated methodology, only varying the reference dataset provided [22, 25]. First, we used the adult reference set to estimate CTP in the adult validation set. Next we used the adult reference set to estimate CTP
in the cord validation set. We then compared the estimation performance of the adult and cord reference sets. Finally, we compared the performance of all three publicly available cord blood references on our cord validation set. For a summary, see Figure 3.1.

3.3.1 Approach

We compared the performance of an existing reference-based CTP estimation pipeline on both our adult and cord validation samples. We applied the widely used method described by Houseman et al. [22], and implemented in R by Jaffe and Irizarry [25]. This method is available as the function estimateCellCount in the minfi package. The method takes in two parameters, a target set of DNAm measurements for which to estimate cell proportions and a reference set of cell type profiles. In this section, we treated estimateCellCount as a black-box and only modified these two parameters.

The estimateCellCount function consists of three steps: normalization, signature construction and optimization. First, quantile normalization [50] is applied to make the reference and target datasets comparable. Then, using the reference dataset, the algorithm selects a number of probes that are differentially methylated between cell types. Finally, an optimization procedure is run to find a linear mixture that best reconstructs the mixed signal in the target samples from the signature. See Section 2.2 for a more detailed treatment.

Measuring Estimation Performance Of Reference Datasets

For the adult reference, we measured CTP estimation performance on both the adult and cord validation sets. For both, we used an unmodified version of estimateCellCounts. Proportions for the Gran, Mono, Bcell, CD4T, CD8T and NK were estimated for each validation sample. Estimates of nRBC were set to 0, because adult-ref does not contain a representation for this cell type. Estimates were compared against measured cell proportions using both Spearman correlation and MAD.

We measured CTP estimation performance of the three cord blood reference datasets. We used an unmodified estimateCellCounts to estimate CTP for the cord
validation set. Estimates for nRBC were made with UBC-ref and JHU-ref, both of which contained reference cell type profiles for that cell type. Estimation of nRBC was omitted when using Oslo-ref because it doesn’t contain representation for this cell type.

Comparing Reference Datasets

To compare reference datasets, we used dimensionality reduction, DMP calling and hierarchical clustering. For dimensionality reduction, we used multidimensional scaling (MDS) plots as implemented in the minfi package under the function mdsPlot [10]. This function performs multidimensional scaling (MDS), projecting the top 1000 most variable probes of the reference dataset onto two dimensions before plotting. To call differentially methylated positions between two datasets, we used probe-wise two-group t-tests on m-values. P-values were adjusted with the Benjamini-Hochberg procedure, and the significance threshold was set to 1% false discovery rate [7]. For hierarchical clustering, we used the default hclust procedure in R with Euclidian distances of beta values between signature probes subsequently used for CTP estimation.

3.3.2 Accuracy of Adult Reference Dataset

We first confirm previous findings that CTP estimation with adult reference cell type profiles performs well with adult samples but not cord samples. To do so, we used adult-ref set to estimate CTP for both our adult and cord validation sets.

Estimates for the adult-validation, when compared to measured proportions, were highly correlated and had low MAD. Figure 3.2 shows that for adult samples, all spearman correlation coefficients were either moderate or high: Gran (Rho=.94), Mono (Rho=.71), NK (Rho=.60), Bcell (Rho=.94), CD4T(Rho=.89) and CD8T (Rho=1.00). Furthermore, we observed that points fell close to the line of unity (black solid line) for all cell types except NK, indicating that estimates were close in magnitude to the measured proportions. This is in line with the previously reported estimation performance on adult samples as reported by Koestler et al. [31].

Estimation performance using adult reference for cord-validation was notably
Figure 3.2: Estimation of Adult Validation Samples Using Adult Reference.
Black line is x=y.

degraded. Figure 3.3 shows that only CD4T retains high correlation (Rho=.79). Gran (Rho=.49), Mono (Rho=.52) and NK (Rho=.59) are now only moderately correlated. Bcell (Rho=.43) and CD8T (Rho=.3) are lowly correlated. Proportions for nRBCs were not estimated because the cell type is not present in adult whole blood and therefore not represented in the adult reference.

Adult reference cell type profiles performed much worse when used to estimate CTP in cord blood samples. Figure 3.4A shows that correlation between measured and estimated proportions were worse for all cell types present in both samples. Similarly, Figure 3.4B shows that the MAD between measured and estimated CTPs is higher for all cell types except NK. This degradation was not as dramatic as those reported by Yousefi et al. [53]. Their results, showing all CTP correlations between 0.01 and 0.03, were surprising. Adult and cord blood are closely related tissues; so our results of partial degradation in estimation performance more closely aligns
with expectations. Similar to our results, Gervin et al. [19] reported that switching from an adult to a cord reference leads to only moderate changes in CTP estimates.

Since correlations are not directly comparable across sample sizes, we measured adult reference performance on subsamples of cord-validation. To do so, we subsampled cord-validation, matching the number of adult validation samples, before estimating CTP using adult-ref. Figure 3.5, shows a similar pattern of adult reference performance degradation. For correlations (Figure 3.5A), Bcell, CD8T and Gran showed the largest degradation when comparing adult against median or 75th percentile performance over cord subsamples. CD4T and Mono estimation performance fell at the top end of the inter-quartile range. Only NK estimates showed similar correlation performance, with the adult correlations landing on the median performance over cord subsamples. Comparisons of MAD performance (Figure 3.5B), also showed Bcell, CD8T and Gran to be significantly degraded.
Figure 3.4: Comparison of Adult Reference Performance by: (A) Correlation, (B) MAD.
Figure 3.5: Comparison of Adult Reference Performance After Subsampling on Cord Validation Set: (A) Correlation, (B) MAD.
We conclude that adult-reference dataset is unsuitable for estimation of CTP in cord blood samples. A major assumption of the reference-based paradigm is that the validation sample is a mixture of the reference cell types profiles. Thus, a parsimonious potential solution to this degraded performance is to use a set of cord blood specific reference cell type profiles.

### 3.3.3 Effects of Using a Cord Specific Reference

Switching from the adult reference to a cord blood specific reference improves CTP estimation performance. This effect is unlikely due to sample size changes, since the adult reference has six profiles per cell type and the cord reference has seven. Figure 3.6 shows estimates for the cord validation set using UBC-ref. All cell types showed improved correlations between estimated and measured proportions (Figure 3.22). Notably, correlations for nRBC, previously not estimated, is
now quite high (Rho=.72). This overall improvement was expected to come from two sources: more accurate representation of DNAm in shared cell types and representation of the previously missing nRBCs.

To isolate the effect of missing nRBC in the reference profile, we compared estimation of the same cord validation set under two conditions: UBC-ref with nRBC and UBC-ref without nRBC. Figure 3.7B shows the percentage difference in estimated CTP between these two conditions. The x-axis is the percentage difference in estimated CTP for the same sample when estimation is done with and without nRBCs reference profile. The y-axis is the number of validation samples with that percentage difference. We observed the sensitivity of cell type proportion estimation to be unevenly distributed across cell types. Dramatic variations in estimates for Bcell (max=50%, mean=20%), Mono(max=52%, mean=10%) and NK(max=62%, mean=21%) were observed when the nRBC profile was added. Other cell types, like CD4T(max=15%, mean=5%), CD8T(max=10%, mean=3%) and Gran(max=23%, mean=13%) were much less sensitive. To visualize how the relationship between cell types affects sensitivity to missing references, we constructed a dendrogram using Euclidian distances. Figure 3.7A shows that sensitivity to missing nRBC profiles is related to both relative cell type abundance and distance to nRBC. This suggests that the optimization constraints from our reference based method assigns DNAm signal from missing cell types to other closely related cell types.

### 3.3.4 Comparison of Cord and Adult References

To investigate the difference in estimation performance for cord blood samples, we compared the adult reference and cord reference cell type profiles. First, we examined analogous cell types between the adult and cord references. Then we looked at the number of pairwise differentially methylated regions (DMRs), which are crucial to our reference-based algorithm. Finally, we looked at how probe-level variability within cell types might play a role in estimation accuracy.

First, we visualized reference cell type profiles showed analogous cell types to be distinct. In Figure 3.8, we projected the thousand most variable probes onto two dimensions using MDS. We observed that the same cell types within cord
Figure 3.7: (A) Clustering Between Cord Cell Types on Signature Sites using Euclidean Distances (B) Sensitivity to Missing nRBC in Reference Profile
Figure 3.8: MDS Plot of Adult and Cord Reference Profiles
and adult, while similar, are visually distinct. In particular, Bcells are particularly different between cord and adult, corroborating that cell type’s loss in accuracy. Furthermore, we noticed that nRBCs are located between the major clusters, a result of their intermediate methylation status (See Section 1.1.3). This indicates that adult cell type profiles are an imperfect proxy for their cord blood equivalent, explaining why adult references performed poorly when predicting CTP for cord blood samples.

Next, we test the hypothesis that prediction of CTP in adult samples is more accurate because their cell types are much more distinct. Here, we quantified distinct as proportion of measured CpGs that are differentially methylated between cell types. Figure 3.9, shows the proportion of DMPs between pairs of cell types within each reference set. In adult, the cell types cleanly separate into lymphoid (CD4T, CD8T, NK, Bcell) and myeloid (Gran, Mono) cell lineages. Furthermore, Table 3.2 shows that in cord blood CD4T and CD8T cells have relatively few dif-

![Figure 3.9: Proportion of DMPs between Cell Types](image-url)
Table 3.2: Number of Differentially Methylated Positions Between Cell Types in Cord

<table>
<thead>
<tr>
<th></th>
<th>Bcell</th>
<th>CD4T</th>
<th>CD8T</th>
<th>Gran</th>
<th>Mono</th>
<th>NK</th>
<th>nRBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bcell</td>
<td>44505</td>
<td>45930</td>
<td>62836</td>
<td>58946</td>
<td>48853</td>
<td>183586</td>
<td></td>
</tr>
<tr>
<td>CD4T</td>
<td>44505</td>
<td>7122</td>
<td>80726</td>
<td>77772</td>
<td>44693</td>
<td>189751</td>
<td></td>
</tr>
<tr>
<td>CD8T</td>
<td>45930</td>
<td>7122</td>
<td>85685</td>
<td>84161</td>
<td>43557</td>
<td>191093</td>
<td></td>
</tr>
<tr>
<td>Gran</td>
<td>62836</td>
<td>80726</td>
<td>85685</td>
<td>17400</td>
<td>58792</td>
<td>152116</td>
<td></td>
</tr>
<tr>
<td>Mono</td>
<td>58946</td>
<td>77772</td>
<td>84161</td>
<td>17400</td>
<td>55965</td>
<td>155700</td>
<td></td>
</tr>
<tr>
<td>NK</td>
<td>48853</td>
<td>44693</td>
<td>43557</td>
<td>58792</td>
<td>55965</td>
<td>163982</td>
<td></td>
</tr>
<tr>
<td>nRBC</td>
<td>183586</td>
<td>189751</td>
<td>191093</td>
<td>152116</td>
<td>155700</td>
<td>163982</td>
<td></td>
</tr>
</tbody>
</table>

differentially methylated sites. Since the reference-based algorithm sub-selects from these DMPs to create the cell type signatures (2.2) for CTP estimation, we expect a lack of candidate probes to have negative downstream effects.

Finally, we examined the variability of the probes because that determines variability of the estimated proportions [24]. So, we examine the number of probes that exhibit high variability, defined as Beta standard deviation greater than 0.05. Figure 3.10 shows that adult cell types have more within-population variable probes than cord across all shared cell types. Thus, we can eliminate variability as the source of low performance in cord sample estimation using cord references.

3.3.5 Comparison of Three Cord Blood References

Currently, there are three publicly available reference cell type profiles for human cord blood. Each reference has salient characteristics (Section 3.1) such as number of samples, method of purification, and cell types represented. In this section, we estimate CTP on the cord validation set with each reference to ensure that they all show similar patterns of low estimation accuracy.

Figure 3.11 shows that there are large performance differences between the different cord references for a few cell types. Gran, Mono, Bcell and CD4T show similar correlations across all three references. For CD8T and NK, JHU has markedly worse correlations. Oslo-ref does not contain nRBC references and so does not make any predictions for that cell type. Overall, Oslo and UBC perform similarly across shared cell types, while JHU stands apart.
Given the large performance differences, we next investigate the similarity of the selected signature probes. Figure 3.12 shows that the overlap between signature probes selected for UBC-ref and JHU-ref range from 24/100 for CD8T to 66/100 for NK. This was surprisingly, because the least overlapping cell type signature and the most overlapping cell type signature both correspond to cell types with markedly lower correlations. Figure 3.12 shows a similar pattern of overlap between UBC-ref and Oslo-ref signature probes, with NK with the most and CD8T with the least. Since the proportion of overlap between JHU/UBC and Oslo/UBC are similar, we cannot attribute JHU-ref’s large drop in correlation to changes in the selected signature probes.

Since the reference datasets differences in performance are not associated with signature similarity, we examine the estimates for individual samples. Figure 3.13 shows that, for shared cell types, Oslo-ref CTP estimates are reasonably correlated with measured CTP, and generally similar to UBC-ref estimates in Figure 3.6.
In contrast, JHU-ref shows large overestimation of CD8T proportions and mostly estimates of 0 for NK proportions.

The similar estimation accuracy between UBC-ref and Oslo-ref was surprising given the difference in sample size; seven and eleven respectively. UBC-ref performed best, despite having the smallest sample size. This suggests that further increasing sample size would be of limited benefit. We are mindful of the fact that UBC-ref might be performing better on cord-validation because they were both measured in the same facilities. However, we believe this effect to be quite small because the reference and validation data were measured in two completely separate experiments.

We proceeded with UBC-ref due to the results above. JHU-ref was eliminated because of poor prediction performance on NK and CD8T. Between the more similar Oslo-ref and UBC-ref, we chose UBC-ref because it included a reference profile

![Estimation Performance By Cord Reference](image_url)

**Figure 3.11:** Correlation (Estimated vs. Measured) by Cord Blood Reference Set
Figure 3.12: Comparison of Signature Probes Between Cord Reference Datasets
Figure 3.13: Estimation Performance by Alternative Cord Reference
for nRBC. Previous results, Figure 3.7 and Figure 3.4, suggest that the missing cell type can highly skew estimates and lead to decreased correlation with measured proportions.

3.4 Normalization

In the last section, we evaluated the performance of reference datasets using an existing estimation pipeline. Here, we begin our examination of the estimation pipeline’s components. The first step is to perform normalization on both the reference and validation datasets.

Preprocessing and normalization are essential to the analysis of 450k data. The intent is to remove sources of variation unrelated to biological phenomena. This variation originates from many sources like dye bias, background fluorescence, differing dynamic ranges between probe types and running samples on different arrays. Enumerations of these technical challenges and their removal can be found elsewhere [11, 14, 33]. In this section, we focus on optimizing the normalization procedure for accuracy of CTP estimation in cord blood samples.

We optimized the DNAm data normalization procedure for CTP using a series of intermediate diagnostics. Table 3.3 summarizes steps for normalizing DNAm microarray array data. Since normalization of DNAm requires multiple steps an exhaustive exploration leads to a combinatorial explosion of procedures. In addition, such a brute force approach would result in severe overfitting on our small validation set. Instead, we used SVD analysis, clustering and p-value inflation to independently improve each step of the normalization.

3.4.1 Approach

Normalization of DNAm proceeds in several steps. First, probes with unreliable measurements are removed. Then within-array normalization corrects for noise unique to each sample. Finally, between-array normalization is used to adjust the measurements so that different arrays are comparable.
Table 3.3: Summary of Evaluated DNAm Normalization Methods

<table>
<thead>
<tr>
<th>Step</th>
<th>Method</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Remove Unreliable Probes</td>
<td>Nordland</td>
<td>Filter for probes with low detection p-value, located at a known SNP, or hybridizes with multiple genomic locations. [40]</td>
</tr>
<tr>
<td>2. Background Correction</td>
<td>ILBG</td>
<td>Background subtraction using the average value of the negative control probes. [4]</td>
</tr>
<tr>
<td></td>
<td>Noob</td>
<td>Background subtraction using out-of-band measurements, instead of just negative control probes. [51]</td>
</tr>
<tr>
<td>3. Probe Type Normalization</td>
<td>SWAN</td>
<td>Match Type 1 and Type 2 probes by CpG density stratification, quantile normalize, and interpolation [1]</td>
</tr>
<tr>
<td></td>
<td>BMIQ</td>
<td>Match Type 1 and Type 2 probe distributions by matching two beta-mixture; one for each probe type. [49]</td>
</tr>
<tr>
<td>4. Between-Array Normalization</td>
<td>Quantile</td>
<td>Match distribution between all arrays by averaging across quantiles. [9]</td>
</tr>
<tr>
<td></td>
<td>ComBat</td>
<td>Match distribution between all arrays and remove batch effects by using location and scale adjustment and empirical Bayes [27]</td>
</tr>
</tbody>
</table>

Baseline Normalization Procedure

Currently, there is no consensus on how to best normalize data for CTP. One common normalization procedure consists only of quantile normalization between the reference and validation datasets [50]. Specifically, the validation and reference samples are pooled together before quantile normalization. This make measurements comparable across samples, but may be removing signal that is desirable in CTP estimation. We will use this as our baseline normalization procedure.

Steps to Optimize

The first step in the normalization procedure, removing unreliable probes, has been well established [16, 42]. Therefore, we did not believe this step had a large potential for improving estimation performance. Instead, we applied a standard filtering protocol as described by Dedeurwaerder et al. [16]. First, we filter for probes with high detection p-value (> .01), low bead coverage (< 3), known to have a SNP at
the CpG, known to cross hybridize with multiple genome loci, or found on a sex chromosome. To identify probes on sex chromosomes, containing a SNP or known to cross-hybridize, we used the list published by Nordlund et al. [40]; other similar lists exist [42]. After this step, probes should have reliable measurements of DNA methylation.

After removing unreliable probes, we dealt with noise and technical artifacts unique to each sample. Two factors have been identified as the main sources of technical effects at the sample level: variability in background fluorescence and differing dynamic ranges between Type 1 and Type 2 probes [16, 49]. For each source, we tested two methods shown to have performed well in several benchmarks [16]. For background subtraction, we tested Illumina background correction (ILBG), and normal-exponential using out-of-band probes (NOOB) [51]. For Type 1 and Type 2 probe balancing, we tested subset within array normalization (SWAN) and beta mixture quantile normalization (BMIQ). To assess the efficacy of each method, we used SVD analysis to test for association of major components of variation with control probes. The Illumina 450k has several control probes designed to measure technical effects, independent of the sample used. We measured association of signal strength at these control probes to the major components of variation using SVD analysis [47]. We expected association with technical factors to negatively impact CTP estimation and thus seek to minimize such associations.

Next, we explored between-array normalizations to make measurements between samples comparable. Measurements between arrays are not directly comparable because they differ by a variety of factors unrelated to the biological signal. Between-array normalizations seek to minimize these differences, leaving only biological signal. We tested two of the recommended between-array normalization techniques recommended by Dedeurwaerder et al. [16]: Quantile Normalization and ComBat. After applying each technique, we used p-value inflation between biologically similar samples to diagnose these technical effects. Specifically, we compared reference cord blood samples measured before purification with cord blood samples in the validation dataset. We used probe-by-probe two-group t-tests and plotted the p-values with the expectation of a uniform distribution. Any deviations from uniform was taken to be indicative of batch effects. Once again, we
expected strong batch effects to negatively impact downstream estimates of CTP.

Finally, after normalizations, we clustered all samples. We used hierarchical clustering with Euclidian distances between all probes beta values. We expected a good normalization procedure to result in samples of the same cell type clustering together.

3.4.2 Within Array Normalization

For within array normalization, we used SVD analysis to assess the association between the principal components of the data and control probe signals designed to be independent of the measured sample to assess presence of strong batch effect. We tested four different within-array procedures: BMIQ, NOOB, SWAN, and ILBG. Only one of NOOB or ILBG can be applied since they both perform background subtraction. Similarly, only one of , SWAN or BMIQ can be applied because they both balance Type 1 and Type 2 probes (described in Section 1.1.2). We concluded that a combination of NOOB and BMIQ best mitigates within-array technical artefacts.

Control probes on the 450k are designed to test the efficiency of various steps involved in measure methylation. These probes are not dependent on the sample, and provide array diagnostics. For example, each negative control probe contains a randomly permuted sequence without CpGs. Thus, fluorescence at these probes measures the system-wide background signal strength. Any association of control probes measurements with major components of variation is indicative of confounding between biological and technical signals. Successful application of preprocessing methods should reduce the significance of associations.

Figure 3.14 shows the association of various control probes with the major components of variation in the raw sample data. We observed that the largest component of variation (PC1) is strongly associated with control probes. Specifically, linear modelling of the first principle component (PC) against many green channel control probes show p-value < $10^{-5}$ for many green channel control probes.

For balancing Type 1 and Type 2 probes, BMIQ out performed SWAN. SWAN erased the association of PC1 to control probes, observed in the raw data, but leaves PC2 strongly associated with control probes. Figure 3.15B shows that SWAN
leaves the green channel control probes strongly associated with PC2 (p-value $< 10^{-5}$). In contrast, Figure 3.15A shows that BMIQ also erases PC1 associations, but reduces the strength of associations between PC2 and control probes. Thus, we use BMIQ for probe type balancing.

For background subtraction methods, we tested ILBG and NOOB. ILBG only uses the negative control probes to estimate the background signal. Due to the small number of probes used, this can lead to poor estimates of the background. In contrast, NOOB uses the out-of-band measurements made on Type 1 probes. These out-of-band measurements are possible because of the probe design. Both methylated and unmethylated CpGs cause fluorescence on only one colour channel, ignoring the second channel. Thus, the second channel can be used to estimate non-specific fluorescence.

We found that NOOB performed better than ILBG. Compared to raw measurements, ILBG left weak associations between control probes and PC1 (p-value $< .05$). Also after ILBG, we observed some strong associations between PC2 and control probes (p-value $< 10^{-5}$). In contrast, Figure 3.15C shows that NOOB completely removed association of PC1 to control probes. There was residual association with PC2 after NOOB, but strength of associations were much lower than ILBG (p-value $< 10^{-1}$).

### 3.4.3 Between Array Normalization

After applying within-array normalizations, we evaluated between-array normalizations that bring the reference and validation sets on to a common scale. We compared the mixed cord blood samples in the reference set to the cord blood samples in the validation set. When these two sets were normalized together using the baseline quantile normalization procedure, we observed broad p-value inflation in probe-by-probe two-group t-tests. P-value inflation is unexpected because both reference and validation sets are biologically similar cord blood samples. Figure 3.16A shows the p-value inflation to have a dominant left skew, indicating a large number of probes with differing methylation distributions. Figure 3.16B shows the result of applying ComBat with each sample’s cell type explicitly modelled. Using ComBat results in over-correction, indicated by the right-skew of the resulting
In the context of CTP estimation, over-correction is preferable to under-correction. When comparing samples of the same type between reference and validation, quantile normalization results in many probes that are differentially methylated. The reference dataset contains measurements from mixed cord blood samples prior to purification. These reference samples should have similar methylation profiles to the validation samples, which are also mixed cord blood. Detection of many differentially methylated sites is a severe violation of the linear mixing assumption (see Section 2.1). Thus, quantile normalization is unfit for our application. On the other hand, over-correction may not have large negative downstream consequences. CTP estimation first builds a signature of discriminating probes used in estimation. Broad over-correction is tolerable if there are probes that retain their between cell type discriminability. Therefore, we conclude that ComBat is better suited to CTP estimation.
Figure 3.15: SVD Analysis of Background Subtraction and Probe Type Normalization Methods
Figure 3.16: P-value Inflation When Comparing Cord Blood Samples After (A) Quantile Normalization (B) ComBat
3.4.4 Cluster Analysis of Full Normalization Pipeline

We compared the baseline normalization against our final normalization pipeline using hierarchical clustering. From the results above, we settled on a final normalization consisting of removing bad probes, NOOB, BMIQ and ComBat. As described in 1.2, we expected samples of the same cell type to cluster together if technical noise was correctly removed. Figure 3.17 shows how the baseline normalization procedure results in imperfect separation of samples by cell type. Two particularly imperfect separations were Gran mixed with Mono and CD4T mixed with CD8T. In contrast, Figure 3.17 shows clustering of samples after applying the optimized normalization procedure. Here, all except two samples cluster cleanly with samples of the same type. The two samples are a swap between NK and cord blood. This shows that our optimized normalization procedure, while possibly removing some signal, reliably corrects for the data’s technical effects.

3.5 Signature Selection

After normalizations, we examined the construction of cell type signatures. First, we looked at the heuristic of balancing the number of probes that are highly and lowly methylated. Then we used a cross-validation framework to find the optimal number of signature probes. Finally, based on our previous pairwise cell type distance results (Figure 3.9), we investigated whether treating CD4T and CD8T as the same cell type improves estimation performance. Since the optimization procedure only considers probes in the signature, the proper selection is crucial to accurate estimation.

3.5.1 Approach

Constructing Cell Type Signature

Signature construction consists of ranking probes based on their ability to discriminate between cell types then deciding how many probes will be included for each cell type. This process relies exclusively on the reference dataset. For each cell type we identified probes that discriminate between that cell type and all others using a probe-by-probe two-group t-test comparing the target cell type with all others.
Figure 3.17: Hierarchical Clustering of Samples After Each Normalization
Probes with a p-value higher than $10^{-8}$ were removed. The remaining probes are ranked in terms of difference between mean beta value in target cell type versus all other cell types. Finally, we selected probes that are most discriminating, as measured by the difference in group mean beta values.

**Investigating the Balancing of Highly and Lowly Methylated Probes**

The adult estimation methodology constructs each cell type signatures from 100 probes balanced between high and lowly methylated measurements. That is, rank probes in terms of group mean beta differences. Then select 50 probes each from the top and bottom of the list. This corresponds to the top 50 probes where the target cell type is more methylated than the other cell types, and the top 50 probes where the target cell type is less methylated than the other cell types. The full signature appends all the individual cell type signatures together, so for 6 cell types the signature contains 600 probes.

We investigated this balancing heuristic by comparing probes ranked by difference in mean beta values, the measure of discriminability, in the cord and adult signatures. For both adult and cord, we constructed signatures for each cell type of size 100, balanced between high and low methylation as described above. Then we asked whether the magnitude of discrimination was different, and whether there were at least 50 discriminating probes in each category of highly and lowly methylated.

**Choosing A Signature Size**

We used a resampling approach to explore how signature size affects CTP estimation. To do so, we created a sequence of signatures, of varying sizes, from the reference samples and multiple validation sets by sampling from our 24 validation samples.

The same sequence of signatures was used for all measurements. We used the cord reference dataset and ranked probes in terms of discriminability as described above. For each cell type, we take the top $N$ most discriminable probes, where $N$ follows the sequence \{1, 3, 5, ... , 49\}. The resulting signature is a matrix with $N$ rows and columns equal to the number of cell types. These 25 signatures were
used for all subsequent measurements.

We created 40 validation sets to measure performance and variability of performance of different signature sizes. First, we randomly split the twenty-four validation samples into two smaller validation sets of twelve samples each. We do this 20 times for a total of 40 validation sets. We measured estimation performance on all 40 samples for each signature in the sequence.

### 3.5.2 Balancing of Highly and Lowly Methylated Probes

The heuristic of balancing high and low beta probes, while effective in adults, is unsuitable for cord blood. For each cell type, signature probes are identified by ranking probes in terms of discriminability, measured by mean absolute beta difference between target cell type and all other cell types. Figure 3.18 shows the discriminability of the top 50 probes for both high and low beta probes for each cell type. We observed that Bcells have similar levels of discriminability in cord and adult for both high and low beta probes. In adults, there were at least 50 probes.
that are capable of acting as signature probes for each cell type. However, this is not true for cord. For high beta probes, Mono does not have 50 high beta probes with an absolute beta difference greater than zero. Thus, the balancing heuristic for cord blood results in discarding highly discriminating probes in the Low Beta category. This result suggests that we should use the top 50 discriminating probes, irrespective of their high or low methylation values.

### 3.5.3 Finding an Optimal Signature Size

We measured the estimation performance of a sequence of signatures of sizes \{1, 3, 5, ..., 49\} on 40 datasets subsampled from our validation samples. Figure 3.19 shows that both mean and median estimation performance peak and plateau around signature sizes between 10 and 15 probes. Performance for nRBC, Bcell, NK and CD8T, all of which have low abundance, tend to have low performance at smaller signature sizes, between 1 and 10 probes, relative to their peak. CD4T and Gran
have relatively good performance even at the extreme case of a signature with 1 probe per cell type. Looking at the 10th and 90th percentiles of performance, variability of estimation performance decreases as N increases and stabilizes around a signature size of 20 probes. Thus, we conclude that in the current prediction pipeline, performance is robust to signature size above 20.

### 3.5.4 Treating T-cells as Indistinguishable

Our previous results on pairwise cell type differences, Figure 3.9, suggests that CD4T and CD8T may be difficult to distinguish using DNAm. So, we investigated whether treating CD4T and CD8T as the same cell type improves estimation performance. Figure 3.20 compares the mean performance over all subsampled validation sets at varying signature sizes. We observed that merging CD4T and CD8T cells did markedly improve performance at signature sizes below 20. However, at 20 probes estimation performance for nRBC, Bcell and nRBC have not yet peaked. At 50 probes, the merged and separated predictions are very similar for all
cell types. Thus, we proceed with a signature size of 50 and the T-cells are kept separate.

3.6 Evaluation

Taking the previous results, we created a CTP estimation pipeline optimized for cord blood samples. We mitigated the risk of overfitting our normalization pipeline by using diagnostics unrelated to CTP estimates. While there is still some chance that we chose a pipeline effective only on our dataset, we believe this is unlikely. Several other benchmarks have suggested that NOOB, BMIQ and ComBat results in high reproducibility of DMP detection and minimizes differences between technical replicates [16, 37].

In this section, we validated the performance of our estimation pipeline. First, we measured the pipeline’s performance on a set of 24 cord blood samples with matched cell counts (Section 3.1). Then we explored how the new normalization procedure affected the construction of cell type signatures during estimation. Finally, we compared our reference-based estimates against reference-free techniques for ability to explain variance in mixed samples.

3.6.1 Approach

Validating Estimation Accuracy Of A Cord Optimized Pipeline

We used our pipeline to estimate CTP in our 24 cord blood validation samples. First, we filtered for unreliable probe measurements from probes with SNPs, cross-hybridization or on a sex chromosome. Second, we performed within-array normalizations on both the reference and validation datasets separately. Within-array normalization consisted of background subtraction using NOOB and probe balancing with BMIQ. nRBCs references are not subject to BMIQ because they do not satisfy the expected bimodal distribution. Third, the reference and validation samples are made comparable using ComBat for between-array normalizations. We used ComBat to correct sample batch with the sample’s cell type explicitly modelled. Once normalization is done, we constructed cell type signatures from the reference by finding the 50 most discriminating probes irrespective of high or
low methylation status. Finally, we pass the signature and validation samples to the optimization procedure to obtain the final CTP estimates.

After benchmarking our pipeline in totality, we broke down the estimation improvements by step. To do this, we applied the changes in a step-wise fashion. Since we do not expect changes to be additive, we instead measured cumulative performance. We applied filtering for reliable probes, background subtraction, probe type normalization, and ComBat in a cumulative fashion. The probe type normalization was applied twice, once to all cell types and once to all cell types except nRBC since it doesn’t exhibit the bimodal distribution of beta values that is expected. For comparison, we applied the baseline normalization with the probe filtering and modified signature selection.

As a sanity check, we tested for association between major components of variation and our estimated CTPs. PCA was used to find the major directions of variation. The top 5 PCs were then tested for association with estimated proportions using a spearman two-sided correlation test.

Changes To The Cell Type Signatures
Since normalization affects the resulting cell type signatures, we compared the resulting signatures from our normalization scheme and the baseline. We constructed two signatures, one after each normalization procedure, by selecting the top 50 discriminating probes by absolute difference in group means.

The two signatures were compared visually with Venn diagrams and analytically with the Jaccard Index. Then, we looked for patterns in the relative rankings of probes. We did this by partitioning the probes into three groups: intersecting probes, probes unique to baseline normalization and probes unique to optimized normalization. For probes in the intersection, we compared their relative rankings, based on discriminability, within each signature. For probes unique to each signature, we examined where they appear in their respective relative rankings.

Comparison To Reference-Free Techniques
Reference-free techniques are an alternative class of methods used to correct for CTH in association studies like genome wide association study (GWAS) and EWAS.
We will collectively refer to these association studies under the more generic xWAS. See Chapter 2 for a summary of reference free techniques relevant to DNAm. We compared reference-based estimation to ReFACTOR, a reference-free technique based on sparse PCA [44]. We chose ReFACTOR because it was developed and previously validated on DNA methylation data. Also, it doesn’t require specification of case and control status, a common requirement for reference-free techniques.

We carried out this comparison from a variance explained perspective because it closely aligns with the task of correction for CTP in xWAS studies. We used the cord validation set with true CTP. To quantify true variance explained by CTP, we fit a linear model to each probe’s beta value as the response and the measured CTP as the explanatory variables. Once fit, the multiple $R^2$ is taken as the true proportion of variability due to CTP. In xWAS, CTP correction is achieved by including either the estimates or surrogate variables in a probe-by-probe regression against case/control status. Any significant variability that is associated with case/control status, when accounting for changes in the CTP, is considered an association. Since we have the “true” amount of variance attributable to CTP, any deviations when regressing against estimated CTP or ReFACTOR variables can be considered over or under correction.

3.6.2 Validating Estimation Accuracy in Cord Blood

We validated our pipeline’s performance by estimating CTP for our cord validation samples. Figure 3.21 shows that all estimates, except CD8T, are moderate or highly correlated with measured CTP: Bcell (Rho=.73), CD4T (Rho=.84), NK (Rho=.66), nRBC (Rho=.68), Gran (Rho=.64), Mono(Rho=.54), CD8T(Rho=.42).

Next, we compared our pipeline's performance to existing estimation methods. Figure 3.22 shows the estimation performance of the standard pipeline with an adult-ref, the standard pipeline with ubc-ref, and our optimized pipeline with ubc-ref. The adult reference is clearly unsuitable for estimation of cord blood CTP. It poorly estimates Bcells, CD8T cell proportions, and provides no estimates for nRBC. Compared to the standard method with cord blood reference, our method markedly improves correlation of estimates for Bcells (+0.15), and Gran(+0.09).
Other cell types showed a slight decrease in correlation: CD4T(-.03), CD8T(-.04), Mono(-.03), NK(-.02), nRBC(-.04).

To understand the incremental gains of each modification, we apply the optimized pipeline step-by-step. Figure 3.23 shows how optimizations have an uneven impact across cell types. Background subtraction using NOOB greatly improved estimation of Bcell and nRBC, but negatively impacts CD8T and Mono. Signature selection without balancing for high and low methylation improves Mono and nRBC, but slightly degrades CD8T. Inter-array normalization using ComBat greatly improves estimation of Bcell and Gran but mostly leaves other cell types alone. BMIQ, which forces methylation measurements to be bimodal, strongly degrades estimation performance when applied to nRBC because their methylation profile do not satisfy the method’s bimodal assumption.

As a sanity check for our CTP estimates, independent of the measured cell counts, we tested for association with the principle components of the validation
data. Figure 3.24 shows that most CTP estimates are strongly associated with PC1: Bcell (p-value < .01), CD4T (p-value ≤ .001), CD8T (p-value ≤ .001), Gran (p-value ≤ .001), and NK (p-value ≤ .001). Further, nRBC is strongly associated with PC2 (p-value ≤ .01) and PC4 (p-value ≤ .01). These results are reassuring since we expect CTP to account for a large fraction of the variation observed in EWAS. Furthermore, this suggests that our estimates could effectively correct for CTH.

### 3.6.3 Changes to the Cell Type Signatures

To investigate how normalization modifies the estimation procedure, we compared the signature probes selected between the baseline and optimized normalization procedures. Figure 3.25 shows that most cell types share between 60% and 74% of their signature probes. nRBC, with 16% shared probes, is an outlier with an almost completely different signature. Using the Jaccard Index as a measure of similarity, Table 3.4 leads to the same conclusion. Once again, most cell types have moderate to high signature similarity, except nRBCs which have low signature

---

**Figure 3.22:** Estimation Performance By Reference And Normalization
**Figure 3.23:** Performance Improvements Step by Step.

**Table 3.4:** Jaccard Index of Signature Probes for Standard and Optimized Normalization

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>Jaccard Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gran</td>
<td>0.471</td>
</tr>
<tr>
<td>Mono</td>
<td>0.515</td>
</tr>
<tr>
<td>Bcell</td>
<td>0.429</td>
</tr>
<tr>
<td>CD4T</td>
<td>0.587</td>
</tr>
<tr>
<td>CD8T</td>
<td>0.493</td>
</tr>
<tr>
<td>nRBC</td>
<td>0.0870</td>
</tr>
<tr>
<td>NK</td>
<td>0.471</td>
</tr>
</tbody>
</table>
Figure 3.24: Association of PCs to Estimated Cell Types
Figure 3.25: Effect of Normalizations on Signature Probe Selection

Table 3.5: Pearson Correlations of Signature Probe Ranks Between Standard and Optimized Pipeline

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bcell</td>
<td>0.65</td>
</tr>
<tr>
<td>CD4T</td>
<td>0.45</td>
</tr>
<tr>
<td>CD8T</td>
<td>0.69</td>
</tr>
<tr>
<td>Gran</td>
<td>0.67</td>
</tr>
<tr>
<td>Mono</td>
<td>0.62</td>
</tr>
<tr>
<td>NK</td>
<td>0.72</td>
</tr>
<tr>
<td>nRBC</td>
<td>0.27</td>
</tr>
</tbody>
</table>

similarity. It was surprising that Bcell and Gran, the two cell types with large performance improvements, did not show larger than average changes to selected signature probes.

Ranks of the shared probes between baseline and optimized pipelines are well correlated (Table 3.5). However, Figure 3.26 shows that most shared probes ex-
Figure 3.26: Changes to Rank Of Signature Probes Between Normalizations
experienced a change in rank. Contrary to our intuition, the shared probes were not enriched for higher discriminatory power (i.e lower ranks). Instead, we observed instances, for example in Bcell, where probes that were ranked as poorly discriminating in the baseline procedure were promoted to be highly discriminating in our optimized procedure.

The optimized pipeline promoted previously low ranked signature probes to improve estimation performance in Bcell and Gran. Looking at Figure 3.27, we observed that only 2 out of the top 20 discriminating probes for Bcell were unique to the optimized pipeline. The other 18 most discriminating probes were promoted from their previously low ranking in the baseline signature. A similar pattern is observed in Gran, with the top 20 discriminating probes originating from previously lower ranked positions in the baseline signature. This suggests that the better normalization scheme is removing probes that appear to be discriminating, but are actually just strongly affected by technical artefacts. For example, all reference samples for one cell type could be highly correlated at one probe due concurrent experimental processing. This probe would be selected during signature construction as being highly discriminatory. However, once the technical artefacts is corrected removed, the probe no longer conveys as much information about cell type identity.

For CTP estimation, measurement at the signature probes appears to be at least as important as their identity. There appears to be some redundancy of information across potential signature probes. This redundancy would manifest as high correlation of methylation in signature probes. Together, this analysis underlines the importance of value calibration when constructing signatures for CTP estimation.

### 3.6.4 Comparison to Reference-Free Techniques

In this section, we compare our reference-based estimation to ReFACTOR, a reference-free correction method, in terms of CTP correction efficacy. So far, we have discussed the performance of CTP estimation in terms of correlation to true measured proportions. However, the end goal of CTP estimation is usually to correct for confounding in genome wide association studies. Here, we explore the correction
Figure 3.27: Rank of Probes Unique to a Normalization
Table 3.6: sPCA Explained Variance at 500 Representative Probes

<table>
<thead>
<tr>
<th>PC</th>
<th>Prop. Variance Explained</th>
<th>$R^2$ to Cell Type Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.88</td>
<td>0.74</td>
</tr>
<tr>
<td>2</td>
<td>0.03</td>
<td>0.80</td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
<td>0.49</td>
</tr>
<tr>
<td>4</td>
<td>0.02</td>
<td>0.56</td>
</tr>
<tr>
<td>5</td>
<td>0.01</td>
<td>0.17</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>0.08</td>
</tr>
</tbody>
</table>

of confounding due to CTP with variance explained as the measure of efficacy (See Section 1.1). Since we can regress against measured CTP to quantify the true amount of variation explained at each probe, we consider any deviations to be spurious.

First, we examine the sparse principle components (sPCs), generated by ReFACTOR, and how they correlate with our measured CTPs. The top 6 sparse PCs account for 97% (Table 3.6) of the variance at the 500 sites used to estimate them (2.3). Within the 6 PCs, the variance explained is heavily skewed towards with first PC at 88%. The first two are highly correlated (Pearson) with measured CTP at $R=0.74$ and $R=0.80$ respectively. The third and fourth PCs show moderate correlation at .49 and .56 respectively. The last two PCs show almost no correlation to measured CTP. This suggests that using the first four PCs would accurately correct for CTP, while using the last two would result in removal of other sources of variation.

To understand whether sPCs explains variance differentially across probes, we measure the amount of variance explained by the first six PCs at different beta standard deviation (SD) value cutoffs. Table 3.7 shows that the amount of variance explained is stable across beta SD cutoffs. Since the results are so stable across all levels of variation, subsequent analyses are based on only probes with beta SD greater than 0.04.

Since sPC1 correlates well with Gran CTP, we examined how well it acts as a surrogate for Gran CTP in explaining variation. Figure 3.28 shows that across all probes, variance explained by sPC1 is highly correlated ($R=0.92$) with variance.
Table 3.7: sPCA Explained Variance Across All Probes

<table>
<thead>
<tr>
<th>$\beta$ SD Cutoff</th>
<th>N Probes</th>
<th>Variance Explained By sPCs</th>
<th>Variance Explained By Cell Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.04</td>
<td>39187</td>
<td>0.51</td>
<td>0.39</td>
</tr>
<tr>
<td>0.022</td>
<td>112744</td>
<td>0.52</td>
<td>0.38</td>
</tr>
<tr>
<td>0.009</td>
<td>256467</td>
<td>0.53</td>
<td>0.38</td>
</tr>
<tr>
<td>0.004</td>
<td>360530</td>
<td>0.52</td>
<td>0.38</td>
</tr>
<tr>
<td>0.002</td>
<td>381286</td>
<td>0.52</td>
<td>0.38</td>
</tr>
</tbody>
</table>

65

Figure 3.28: Variance Explained by sPC1 Versus Measured Gran Proportions

explained by measured Gran CTP. This suggests that sPC1 would be a reliable proxy for Gran CTP during correction in association studies. However, cord blood contains many other cell types besides Gran, so we next examine how well sPC1 represents all cell types.

Regressing against sPC1 does not correctly account for variation due to all cell types. Figure 3.29 shows a large number of probes that have variation attributable to cell type, but not sPC1; seen as the large blob spreading across the x-axis.
Figure 3.29: Variance Explained by sPC1 Versus All Measured CTP

Increasing the number of top sPCs used for correction from top 1 to 6 results in evidence of overcorrection. Figure 3.30 (bottom) shows that sparsePCA has a dense cloud of points on the upper-left quadrant, where each point represents a higher amount of variance explained by sPCs than expected. Amount of variance explained by predicted cell type, Figure 3.30 (top), also shows evidence of over-correction to a lesser extent, seen as a slight skew above the line of unity.

As shown in Figure 3.31, using only the top sPC results in under-correction and using the top six PCs results in overcorrection. Relative to sparse principle components, using estimated cell counts from reference based techniques, while slightly over-correcting, is closest to our measured ground truth. Therefore, we recommend using reference based techniques when available.
Figure 3.30: Variance Explained by Top 6 sPCs Versus All Measured CTPs. Red is line of unity.
Figure 3.31: Difference to Variance Explained by Measured CTP
Chapter 4

Conclusions

The number of EWAS studies on human infant cord blood continues to grow, but these results are difficult to interpret due to a lack of adequate CTH correction techniques. Currently, reference-based CTP estimation techniques have been effectively correcting for CTH in adult whole blood EWAS studies. However, these methods do not generalize to studies of cord blood, a proxy for prenatal conditions. In this thesis, we explore why current techniques are inadequate and propose a better CTP estimation technique for cord blood.

First, we validated that current reference-based estimation techniques perform poorly on a cord blood samples. Treating the existing methodology as a black box, we compared estimates to experimentally generated ground truth CTP on a cord blood validation set. The most widely used reference-based method for adult whole blood does not generalize to infant cord blood. We observed a large difference in estimation performance between adult and cord samples. One possible reason for this degradation is the distinct methylation patterns between adult and cord cell types in blood. Another possible cause is the presence of nRBCs, a cell type unique to cord blood, which is unaccounted for by the adult method.

Next, we compared how well the existing methodology performs using three different publicly available cord blood reference data sets. These datasets were generated specifically to improve estimation accuracy. Our results showed that the same few cell types are poorly estimated by all three references. However, updating the references does result in performance gains and so partially resolves
the problem. We also showed that two out of the three references provide similar estimation performance, with the third under-performing. We recommended one particular reference based on validated performance and more comprehensive representation of cord blood cell types.

Then, we explored how estimates can be improved by opening up the methodology black box. Our analyses suggested several causes for the loss of accuracy: insufficient preprocessing of batch effects, improper data normalization, and poor representation of cell types. We resolved the identified issues and demonstrate improved CTP estimation in cord blood.

Finally, we compared our estimates to a reference-free method using variance explained. This metric mimics correction for CTH without requiring case-control status. We observed that reference-based techniques can circumvent over-correction, a drawback of reference-free techniques.

This thesis is a case study in extending reference-based CTP estimation techniques for DNAm to a novel tissue. We demonstrated how extending a reference-based technique requires more than just a new reference dataset. Proper extension required careful consideration of both normalization and optimization in order to achieve high performance.

Despite their high performance, reference-base techniques are not without limitations. For example, one difficulty is the delineation of cell types. In our study, granulocytes could have been further subdivided into eosinophils, basophils and neutrophils. Another troublesome subdivision was the separation of T-cells into CD4T and CD8T. Another limitation of reference-based techniques are their scope of application. Reference-based techniques can only be used on tissues for which we can generate references.

In addition to the described results, we explored several other, ultimately unsuccessful, avenues. One promising avenue was to borrow information from the adult reference dataset for prediction in cord blood. This idea fits under the umbrella of transfer learning and we thought it could limit the search space for cell type markers. We didn’t observe any performance improvements, possibly because adult and cord cell types have very distinct methylation profiles. Another avenue was to perform estimation in a hierarchical fashion. We attempted to estimate cell types in groups that are iteratively refined, mirroring the process of cell type dif-
differentiation. This approach resulted in compounding errors and poor fine grained estimates of CTP. Finally, we assessed the effects of reference sample size by pooling together the different cord blood references. This was also unhelpful, possibly due to noise added by the diversity of methodology and experimental designs. Even though these paths did not improve estimation accuracy, they were instructive and paved the way to our successful method.

In the future, we’d like to study hybrid approaches that combine experimental and blind source separation techniques. This would extend the applicability of reference-based techniques to tissues that currently defy granular experimental characterization.
Bibliography


[26] R. Jiang, M. J. Jones, F. Sava, M. S. Kobor, and C. Carlsten. Short-term diesel exhaust inhalation in a controlled human crossover study is associated with changes in DNA methylation of circulating mononuclear cells in


URL http://www.nature.com/doifinder/10.1038/nmeth.2815