Title
Capturing Hidden Signals From High-Dimensional Data and Applications to Genomics

Author
Rahmani, Elior

Publication Date
2020

UNIVERSITY OF CALIFORNIA

Los Angeles

Capturing Hidden Signals From High-Dimensional Data and Applications to Genomics

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Elior Rahmani

2020

ABSTRACT OF THE DISSERTATION

Capturing Hidden Signals From High-Dimensional Data and Applications to Genomics

by

Elior Rahmani
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2020
Professor Eran Halperin, Chair

The analysis of high-dimensional data, albeit challenging owing to various computational and statistical aspects, often provides opportunities to uncover hidden signals by leveraging inherent structure in the data. In the context of genomics, where molecular markers are probed at ever-increasing resolution and throughput, large sets of features that follow specific patterns, in conjunction with large sample sizes, allow us to implement richer and more sophisticated models than before in attempt to extract signal that is not immediately evident from the data. Particularly, genomic markers are often affected by multiple genetic and environmental factors, they may differ in their regulation and presentation in different tissues, cell types, conditions, or over time, and some markers may affect multiple biological processes; unveiling those signals is likely to be pivotal in advancing our understanding of complex biology and disease. This dissertation introduces novel computational methodologies and theory that address several key challenges faced in the analysis of high-dimensional genomic data coming from heterogeneous sources ("bulk" genomics) with a particular focus on DNA methylation data. Through a range of simulations and the analysis of multiple data sets, we demonstrate that our proposed methods provide opportunities to conduct powerful and statistically sound population-level studies at an unprecedented resolution and scale.

The dissertation of Elior Rahmani is approved.

Steve Horvath

Jason Ernst

Sriram Sankararaman

Eran Halperin, Committee Chair

University of California, Los Angeles

2020

To my mother,

for her selfless giving and endless dedication;

my late father,

for being my greatest teacher and for igniting my love for math;

my wife Dovrat and my son Roee,

for being my anchor that holds on a stormy day.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to acknowledge my academic advisor Eran Halperin. I have known Eran and worked with him since 2012 - first, as an undergraduate research student and later as a master's student at Tel Aviv University, and eventually as a PhD student at UCLA. In hindsight, choosing an advisor in a very early stage of your academic training is almost an arbitrary decision from a professional point of view - you lack the skills that would allow you to properly evaluate the qualities of your prospective advisor, let alone their approach for doing meaningful and rigorous science, a key determinant that affects your scientific development and the kind of scientist you become by the end of your training. Yet, I was very fortunate to have started working with Eran, a scientist of exceptional qualities.

From an academic perspective, Eran's guidance spanned the entire spectrum of research: from high-level conceptualization, through developing the details, to rigorously solving equations or engaging in hands-on coding when needed. Eran was involved in my research at the right times, and, sometimes more importantly, he knew when to be supportive of me taking my own path. In fact, Eran's unique combination of setting almost no restrictions or boundaries for his students on their work habits and style while setting a high bar of expectations and standards in science has proven to be tremendously productive for me.

From a personal perspective, Eran has been showing a genuine sense of responsibility towards me, well above and beyond what I could have expected. He has taught me a lot about work-life-family balance and markedly changed my way of thinking in many aspects of life. Altogether, Eran's mentorship provided me with the confidence and credibility to focus on research and navigate to success even in times of uncertainty and low productivity and I am grateful for that.

I owe a special gratitude to Eleazar Eskin, who made my transition to UCLA so much easier than it would have been otherwise. Beyond being an amazing collaborator and providing insightful guidance in my research, Eleazar has been a great mentor for me, and during my time at UCLA I have always thought of him as my unofficial second advisor. His continuous will to help and guide me through academic and personal matters has always given me the feeling that he is looking out for me. At a wider level, through his tireless efforts to enhance the computational biology community at UCLA in numerous academic, pedagogical, and social aspects, Eleazar has been a prominent driver behind transforming UCLA into a leading center for computational biology; being part of such a vibrant and stimulating environment was an amazingly enriching experience for me.

Throughout my academic training, I have had many great collaborators and mentors. Some of them were instrumental in their contribution to my work or in their effect on my development as a scientist by mentoring, inspiring, and giving me good advice along the way. In particular, I learned a lot from Sriram Sankararaman, whose vast knowledge and skills are truly admirable; I very much enjoyed having discussions, sharing ideas, and working with the extraordinarily creative Noah Zaitlen, whose passion for science is contagious; through our typically technical (and occasionally seemingly nitpicking) discussions, I learned a ton from Saharon Rosset, who impelled me to strive for scientific rigor; I was privileged to work and exchange fascinating ideas with Jonathan Flint, a remarkable scientist who inspired me to think big; and I had a great pleasure to closely collaborate with the energetic Päivi Pajukanta, who greatly broadened my understanding of genomics and improved my big-picture thinking in research. In addition, I would like to express my sincere gratitude to Steve Horvath, Jason Ernst, and Sriram Sankararaman for their support and guidance as my dissertation committee members.

I would also like to acknowledge the past members of Eran's lab and my former fellow students from Tel Aviv University, and, in particular, Regev Schweiger, Omer weissbrod, Eyal Fisher, Oren Avram, and Reut Yedidim, an elite group of people whom I closely collaborated with; we spent many days and hours on stimulating scientific discussions and developing ideas that enriched me as a scientist.

At UCLA, I very much enjoyed working with the highly talented Brandon Jew, Marcus Alvarez, Brian Hill, Johnson Chen, Liat Shenhav and Mike Thompson. On the administrative front, my deep appreciation is given to Sim-Lin Lau and Joseph Brown, who have always went above and beyond their duties to assist me in various administrative matters.

The journey for higher education begins much before you take your first class or start to work on your first research project. Successfully overcoming the various challenges I faced throughout my PhD training in a competitive environment required having the right mental state-of-mind, much of which is the embodiment of the skill set I acquired over time with the guidance and support of my family. For that, I am forever grateful to my parents and sisters.

Finally, Erich Fromm once said “The quest for certainty blocks the search for meaning. Uncertainty is the very condition to impel man to unfold his powers.” My wife Dovrat has been the key person that made it possible for me to take the uncertain path of pursuing a PhD degree abroad. While embarking on an uncertain voyage of her own by relocating to a different country, away from her family, friends, and career, Dovrat’s unconditional love, support, and endurance in the good and in the bad times allowed me to unlock the full potential of my time at UCLA; I am indebted for that. Lastly, I thank my son Roee, whose smile brings me an unparalleled sensation of uplifting joy.

Chapter Two in this dissertation is a version of Rahmani E, Shenhav L, Schweiger R, Yousefi P, Huen K, Eskenazi B, Eng C, Huntsman S, Hu D, Galanter J, Oh SS. “Genome-wide methylation data mirror ancestry information”. Epigenetics & chromatin. 2017 Dec;10(1):1-2. doi.org/10.1186/s13072-016-0108-y.

Chapter Three is a version of Rahmani E, Schweiger R, Shenhav L, Wingert T, Hofer I, Gabel E, Eskin E, Halperin E. “BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference”. Genome biology. 2018 Dec;19(1):1-8. doi.org/10.1186/s13059-018-1513-2.

Chapter Four is a version of Rahmani E, Schweiger R, Rhead B, Criswell LA, Barcellos LF, Eskin E, Rosset S, Sankararaman S, Halperin E. “Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology”. Nature communications. 2019 Jul

31;10(1):1-1. doi.org/10.1038/s41467-019-11052-9.

Chapter Five is a version of a work that was submitted for publication by Rahmani E, Jew B, Schweiger R, Rhead B, Criswell LA, Barcellos LF, Eskin E, Rosset S, Sankararaman S, Halperin E, tentatively titled “Calling differential DNA methylation at cell-type resolution: addressing misconceptions and best practices”.

VITA

Spring 2018 Teaching Assistant, Computational Genetics, Department of Computer Science, University of California, Los Angeles, CA, USA.

2016 – 2017 Lecturer, Bioinformatics Tools, Department of Computer Science, Tel Aviv University, Tel Aviv, Israel.

2014 – 2016 MSc., Computer Science, Tel Aviv University, Tel Aviv, Israel.

2010 – 2013 BSc., Computer Science and Biology (specialization in Bioinformatics), Tel Aviv University, Tel Aviv, Israel.

SELECTED JOURNAL PUBLICATIONS

Alvarez M*, **Rahmani E***, Jew B, Garske KM, Miao Z, Benhammou JN, Ye CJ, Pisegna JR, Pietiläinen KH, Halperin E, Pajukanta P. Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *Scientific reports.* 2020 Jul 3;10(1):1-6. (* - joint first authorship)

Jew B*, Alvarez M*, **Rahmani E**, Miao Z, Ko A, Garske KM, Sul JH, Pietiläinen KH, Pajukanta P, Halperin E. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications.* 2020 Apr 24;11(1):1-1. (* - joint first authorship)

Rahmani E, Schweiger R, Rhead B, Criswell LA, Barcellos LF, Eskin E, Rosset S, Sankararaman S, Halperin E. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature communications.* 2019 Jul 31;10(1):1-1.

Thompson M, Chen ZJ, **Rahmani E**, Halperin E. CONFINED: distinguishing biological from technical sources of variation by leveraging multiple methylation datasets. *Genome biology*. 2019 Dec;20(1):138.

Rahmani E, Schweiger R, Shenhav L, Wingert T, Hofer I, Gabel E, Eskin E, Halperin E. BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. *Genome biology*. 2018 Dec;19(1):1-8.

Schweiger R, Fisher E, Weissbrod O, **Rahmani E**, Mller-Nurasyid M, Kunze S, Gieger C, Waldenberger M, Rosset S, Halperin E. Detecting heritable phenotypes without a model using fast permutation testing for heritability and set-tests. *Nature communications*. 2018 Nov 21;9(1):1-9.

Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, Oh S, Burchard EG, Eskin E, Zou J, Halperin E. Correcting for cell-type heterogeneity in DNA methylation: a comprehensive evaluation. *Nature methods*. 2017 Mar;14(3):218-9.

Weissbrod O, **Rahmani E**, Schweiger R, Rosset S, Halperin E. Association testing of bisulfite-sequencing methylation data via a Laplace approximation. *Bioinformatics*. 2017 Jul 15;33(14):i325-32.

Rahmani E*, Yedidim R*, Shenhav L, Schweiger R, Weissbrod O, Zaitlen N, Halperin E. GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data. *Bioinformatics*. 2017 Jun 15;33(12):1870-2. (* - joint first authorship)

Rahmani E, Shenhav L, Schweiger R, Yousefi P, Huen K, Eskenazi B, Eng C, Huntsman S, Hu D, Galanter J, Oh SS. Genome-wide methylation data mirror ancestry information. *Epigenetics & chromatin*. 2017 Dec;10(1):1-2.

Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, Oh S, Burchard EG, Eskin E, Zou J, Halperin E. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature methods*. 2016 May;13(5):443.

CHAPTER 1

Introduction

1.1 Scope of research

Technologies for probing molecular genomic markers at ever-increasing resolution and throughput hold a promise to provide a bridge to understanding complex biology. Indeed, the study of genomics has already had a tremendous impact, ranging from our understanding of basic building blocks of biology through practical tools for enhancing healthcare. Yet, success in studying complex biological systems through genomics has been very limited thus far, presumably owing to the complex nature of genomics.

Genomic markers such as gene expression and DNA methylation may be affected by multiple genetic and environmental factors [1, 2, 3, 4], they may differ in their regulation and presentation in different tissues [5], cell types [6], conditions [7], or over time [8], and some markers may affect multiple biological processes [9]. Critically, the effect of a given marker on a complex system likely involves an interplay with multiple other genomic factors. The development of better models and methodologies that can account for those complexities is therefore of primary interest in the study of genomics.

The high-dimensional nature of high-throughput genomic data, albeit challenging, provides opportunities to uncover hidden signals by leveraging the structure of the data: large sets of features, in conjunction with large sample sizes that follow specific patterns, allow us to implement richer and more sophisticated models in attempt to eventually improve our understanding of complex biology. Here, we focus on genomic data generated from heterogeneous sources (”bulk” genomics). Particularly, we focus on DNA methylation data collected from heterogeneous tissues such as blood; such data are now ubiquitously available for large samples.

Bulk data represent convolution of signals coming from different sources (typically different cell types), thus introducing further challenges with data analysis and interpretation. More specifically, one of the most eminent challenges in the analysis of bulk genomics stems from the fact that different cell types demonstrate differences in their genomic patterns. The observed bulk data of a sample therefore reflect convolutions of signals coming from the different cell types in the tissue under study, weighted by the cell-type composition of the sample. Both the cell-type composition and the cell-type level genomic profiles may vary across individuals, conditions, and over time, thus further complicating data modeling and interpretation.

Notably, technologies for profiling genomic markers at single-cell resolution - most markedly single-cell RNAseq - allow us to probe genomics at an unprecedented level of granularity. Those technologies hold the promise to address several challenges faced in the analysis of bulk genomics; perhaps most prominently, the challenge of tissue heterogeneity, arising due to the differences in genomic patterns between different cell types. However, single-cell technologies have yet to allow the routine and reliable generation of data at large scale (i.e. large number of individuals) and at reasonable costs and effort; as long as these challenges remain in place, bulk data is likely to still be the predominant tool for studying population-level genomics. Moreover, those technologies are still in relatively early stages for some genomic data types; for instance, there have been recent significant advances in the developement of single-cell DNA methylation, however, coverage and throughput remain major challenges that need to be addressed [10].

Even if further advances will alleviate the scalability and reliability limitations of current single-cell technologies in the near future, the large number of existing bulk samples that have been collected by now are still an extremely valuable resource for genomic research (e.g., more than 100,000 DNA methylation bulk profiles to date in the Gene Expression Omnibus (GEO) alone [11]). These data reflect years of substantial community-wide effort of data collection from multiple organisms, tissues, and under different conditions, and it is therefore of great importance to develop better models and tools for the analysis of bulk data.

1.2 Contributions and Overview

This dissertation tackles several key challenges faced in the analysis of bulk genomic data, with a particular attention to DNA methylation data. We introduce novel models, theory, and practical tools for capturing hidden signals from high-dimensional data, which we have developed and applied primarily in the context of DNA methylation studies.

The correlation between genetics and genomic markers such as gene expression and DNA methylation has been repeatedly established in the genomics literature [2, 12]. However, in the context of methylation it was previously not clear to what extent genetics and population structure are reflected in genome-wide methylation. In Chapter 2, we tackle this question and further provide a tool for capturing ancestry information from methylation data without the need for genetics.

While the problem of estimating cell-type composition from genomic data has gained much attention in the computational biology literature [13, 14], there was previously no reliable way to estimate cell-type proportions from methylation without the use of reference methylation data collected from purified cells. In Chapter 3, we demonstrate both theoretically and empirically that previous reference-free methods do not address this problem adequately, and we propose a new semi-supervised method that meets the goal without the need for methylation reference.

In Chapter 4, we introduce Tensor Composition Analysis (TCA), which aims at learning three-dimensional hidden signals from two-dimensional input data. We demonstrate the utility of TCA for obtaining cell-type-specific resolution epigenetics from bulk methylation data, and further use it for the identification of differential methylation at cell-type level.

Finally, we further discuss TCA and put it in context with alternative models. Particularly, in Chapter 5, we provide theoretical results, showing the relation of TCA to other methods, as well as motivate the key idea behind TCA through asymptotic analysis. In addition, we demonstrate more experiments and further discuss the application of TCA under additional assumptions and statistical tests.

CHAPTER 2

Genome-wide methylation data mirror ancestry information

2.1 Background

The relation between ancestry and genetic variation has been repeatedly established over the last decade [15, 16]. Several methods now provide accurate estimates of ancestry information by leveraging genome-wide systematic difference in allele frequencies between subpopulations, commonly referred to as population structure [17, 18, 19, 20, 21]. These methods often apply Principal Component Analysis (PCA) or variants of PCA.

Inferring population structure across individuals provides a powerful source of information for various fields, including genetic epidemiology, pharmacogenomics and anthropology. For instance, in genetic and molecular epidemiology, in which identifying genetic associations with disease or exposure is of primary interest, it is essential to have ancestry information in order to distinguish effects of demographic processes from biological or environmental effects. Specifically, the importance of controlling for population structure in genome-wide association studies (GWAS) is now well appreciated. Unless appropriately accounted for, population structure in GWAS can lead to numerous spurious associations and might obscure true signals [18, 22].

Emerging epigenome-wide association studies (EWAS) revealed thousands of CpG methylation sites correlated with genetics and with ancestry [12, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. Not surprisingly, due to the genetic signal present in many CpGs, several studies have shown that the first several principal components (PCs) of methylation data can capture population structure in cohorts composed of European and African individuals [28, 34]. However, unlike the case of genotyping data in which global ancestry information is robustly reflected by the top PCs, the first several PCs of methylation data were also shown to capture other factors in different scenarios, mainly cell-type composition in case of data collected from heterogeneous tissues [35, 36], but also other factors, including technical variables, age and sex [28, 34]. Moreover, it is now appreciated that collecting methylation using probes with polymorphic CpGs is affected by hybridization sensitivity and does not necessarily reflect methylation variability but rather genetic variability [37]. Therefore, it is not clear to what extent global whole-genome DNA methylation states are affected by population structure when these artifacts are removed.

We introduce EPISTRUCTURE, a method for capturing ancestry information from DNA methylation data. EPISTRUCTURE is based on the observation that PCA computed from a set of methylation CpG sites that are highly correlated with SNPs efficiently captures population structure. Thus, we use a large reference data set that includes both genotypes and methylation in order to find correlations of CpGs with cis-located SNPs, and to generate a reference list of genetically-informative CpGs. Then, given new methylation data we compute the PCs of the methylation levels from the same sites included in the reference list. We validate the robustness of this method by assessing the correlation between the methylation-inferred ancestry and the genetically inferred ancestry on two additional large methylation data sets.

In order to shed light on the relation between genetic ancestry and methylation-based ancestry, we further explore the unsupervised detection of ancestry from methylation data. We show that genome-wide methylation mirrors ancestry information in admixed populations after properly adjusting for known variability in genome-wide methylation, and after properly removing technical artifacts, particularly probes that include SNPs that may confound the results. Thus, unlike previous studies that were potentially confounded by these artifacts, here we show that ancestry is indeed robustly mirrored by methylation data as one of the main variance components in the data.

EPISTRUCTURE can be used to infer ancestry information from methylation data in the absence of genetic data. Although genetic data are often collected in epigenetic studies of large cohorts, these are typically not made publicly available, making the application of EPISTRUCTURE especially useful for anyone working on public data.

2.2 Methods

2.2.1 Model and algorithm

Previous studies reported a large number of correlations between DNA methylation and genetics, mainly cis-correlations between CpGs and nearby SNPs [12, 23, 24, 25]. We therefore assume that cis-located SNPs can capture the genetic variability accounting for the methylation levels of a given CpGs. For a given CpG x we assume the following linear model:

$$x = \beta_0 + \sum_{s_j \in S_x} \beta_j s_j + \epsilon$$

where S_x is a group of w SNPs, cis-located with respect to x, $\{\beta_j\}_j$ are their corresponding effects on the methylation levels and ϵ represents an error term, assumed to be independent between different samples.

Given reference data of methylation levels and genotypes for the same individuals, we fit the above linear model for each CpG. We regard the CpGs for which the model fits well as linear combinations of SNPs. We define the set of these genetically-informative CpGs as the reference list. Given methylation data for new individuals, we can estimate the population structure in the data by applying a standard PCA on the sites in the reference list. The first several PCs of PCA are well-known to efficiently capture ancestry information when applied to genotypes data [18], therefore applying PCA on CpGs that are linear combinations of SNPs is expected to capture population structure as well. In the next subsection we further demonstrate this intuition mathematically.

Given reference methylation and genotypes data, our suggested algorithm can be summarized as follows:

1. For each CpG x, fit a linear model using w SNPs that are closest to x.
2. Define a reference list G of all the CpGs for which the linear model fits well. Evaluate model fit based on cross-validated squared linear correlation.
3. Given a new methylation data set, apply PCA on the sites defined by G and consider the first k principal components as the estimate of the population structure.

Note that creating a reference list, described in the first two steps of the algorithm, needs to be performed only once. Population structure can be then inferred in future data sets using this list of genetically-informative CpGs. In practice, an appropriate w may be relatively large (i.e. large number of predictors), while the sample size is typically limited. We therefore apply a regularized regression with ℓ_1 penalty, also known as LASSO regression [38]. For the same reason, we define a parameter p to limit the maximal number of predictors in each model. Furthermore, in order to avoid over-fitting of the model, we perform a k-fold cross-validation procedure for each CpG. The score of a CpG is defined as the median squared linear correlation of its predicted values with the real values across the k folds. Finally, a reference list of the CpGs is defined as the set of sites with highest scores.

In principle, one could use the same approach taken here in order to create and use a reference list of CpGs which explain SNPs well rather than CpGs which are captured well by SNPs. However, modeling methylation levels as a function of SNPs is more natural with respect to the causality relations assumed between SNPs and methylation. Moreover, many methylation sites are known to be affected by several factors (e.g. age [39], gender [40] and smoking [41]), and therefore considering a group of methylation sites explaining a given SNP may introduce into the data more, potentially unknown, variance in addition to genetic variance. This potential problem is expected to be less severe in the opposite direction of modeling methylation using SNPs. In this case, methylation sites that are well explained by genetics are less likely to be highly explained by more factors.

2.2.2 Interpreting EPISTRUCTURE

The EPISTRUCTURE algorithm can be divided into two main steps. First, a reference list of genetically-informative methylation sites is compiled from a group of CpGs, each found to be well approximated by its cis-located SNPs. Second, given a new methylation data set, the first several PCs of the data are calculated only from the sites that were included in the reference list. The reason for applying PCA in the second part of the algorithm is motivated by the success of PCA to capture ancestry information in genotyping data [18]. In the case of genotyping data coming from different populations, the first several PCs capture population structure by highlighting groups of individuals differing at the level of allele frequencies. Given an $n \times s$ centered genotyping data matrix G of s SNPs collected from n individuals, the generative model underlying PCA assumes:

$$\begin{gathered} G = ZW + \Sigma \\ \Sigma_j \sim MVN\left(0, \tau^2 I_n\right) \end{gathered} \tag{2.1}$$

where Z is an $n \times k$ matrix representing k-dimensional latent structure of the ancestry information for each individual and W is a $k \times s$ matrix representing ancestry-specific differences in allele frequencies for each SNP. Σ is an $n \times s$ error term, typically assumed to have independent entries (that is, no relatedness between the n individuals and independence between the SNPs).

Any methylation site can be modeled as a linear function of SNPs and additional error term, and therefore the methylation level of a specific site in a given individual can be approximated to some extent using merely the individual's SNPs. Formally, given an $n \times m$ centered methylation data matrix X of m methylation sites coming from the same n individuals in G, we can describe X_j, the j-th column of X as:

$$\begin{gathered} X_j = GB_j + E_j \\ E_j \sim MVN\left(0, \sigma_j^2 I_n\right) \end{gathered} \tag{2.2}$$

where B_j is an $s \times 1$ coefficients vector of the linear model and E_j is an $n \times 1$ error term. In particular, methylation site j that cannot be even partially explained by SNPs will have a corresponding B_j vector of only zeros. In the first step of the EPISTRUCTURE algorithm we find a group of methylation sites which can be well explained by their cis-located SNPs. Restricting the data matrix X to be consisted only of such methylation sites increases the signal-to-noise ratio in the data.

Plugging (2.1) into (2.2) we get

$$\begin{aligned} X_j &= (ZW + \Sigma)B_j + E_j &(2.3)\\ &= ZWBj + \Sigma B_j + E_j &(2.4)\end{aligned}$$

where ΣB_j and E_j are normally distributed as before. This model can be equivalently described as follows:

$$X_j \sim MVN\left(ZWB_j, \left(B_j^t B_j \tau^2 + \sigma_j^2\right) I_n\right) \tag{2.5}$$

Under this formulation there is a dependency between every two methylation sites. However, based on previous reports showing clear predominance of associations between CpGs and cis-located SNPs over trans-located SNPs [23, 24, 25], we assume that only cis-located SNPs are informative for explaining a given methylation site. As a result, B is expected to be very sparse with values concentrated around the diagonal, assuming the SNPs and CpGs are ordered by physical position. In particular, every two distant methylation sites are independent. In our case the matrix B was estimated from the KORA data for which both genotyping and methylation levels were available. We observed that the vast majority of the rows in the estimated matrix are sparse and only rarely have more than one non-zero entry (Figure 2.1). The main reason for this is the fact that we consider only a sparse set of methylation sites from the genome, resulting from the first step of the algorithm in which only sites that can be well approximated by SNPs are selected. Therefore, we neglect the theoretical dependency between close sites and assume no dependency between any of the

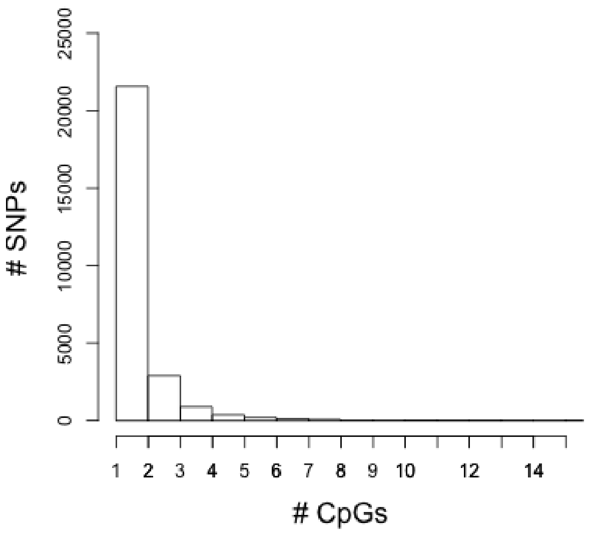

Figure 2.1: Most of the available SNPs used in creating the reference list of genetically-informative CpGs using the KORA data were found to be predictors of no more than one CpG in the reference list. Only 26,244 out of the available SNPs in KORA (657,103) were used in the prediction of the CpGs that were included in the reference list. Out of these sites 82.2% were found to be predictors of only one CpG and 93.3% were found to be predictors of at most two CpGs.

columns in B. Now, the model can be summarized as:

$$X_j \sim MVN(Z\tilde{W}_j, \psi_j^2 I_n) \tag{2.6}$$

where $\tilde{W}_j = WB_j$ and we are interested in extracting Z, the latent ancestry information structure of the individuals in the data. The maximum likelihood solution to the model in (2.6) is given by factor analysis, and the first k factors can be used as estimates of the latent population structure Z. In practice, factor analysis iteratively scales each site and the first iteration is equivalent to PCA after standardization of each of the sites. Empirically, applying more than one iteration did not improve the performance, therefore, in the second step of the EPISTRUCTURE algorithm we suggest to perform a standardized PCA and to consider the first k PCs as the estimate of the population structure.

2.2.3 Data and quality control

The longitudinal KORA study (Cooperative health research in the Region of Augsburg) consists of independent population-based subjects from the general population living in the region of Augsburg, southern Germany [42]. Whole-blood samples of the KORA F4 study were used ($n = 1,799$) as described elsewhere [43]. Briefly, DNA methylation levels were collected using the Infinium HumanMethylation450K BeadChip array (Illumina). Beta Mixture Quantile (BMIQ) [44] normalization was applied to the methylation levels using the R package wateRmelon, version 1.0.3 [45]. In total, 431,360 probes were available for the analysis. As described elsewhere [46], genotyping was performed with the Affymetrix 6.0 SNP Array (534,174 SNP markers after quality control), with further imputation using HapMap2 as a reference panel. A total of 657,103 probes remained for the analysis.

We used whole-genome DNA methylation levels and genotyping data from the Genes- environments & Admixture in Latino Americans (GALA II) data set, a pediatric Latino population study. Details of genotyping data including quality control procedures for single nucleotide polymorphisms (SNPs) and individuals have been described elsewhere [47]. Briefly, participants were genotyped at 818,154 SNPs on the Axiom Genome-Wide LAT 1, World Array 4 (Affymetrix, Santa Clara, CA) [48]. Non-autosomal SNPs and SNPs with missing data (> 0.05) and/or failing platform-specific SNP quality criteria ($n = 63,328$) were excluded as well as SNPs not in Hardy-Weinberg equilibrium ($n = 1,845$; $p < 10^{-6}$) within their respective populations (Puerto Rican, Mexican, and other Latino). Study participants were filtered based on 0.95 call rates and sex discrepancies, identity by descent and standard Affymetrix Axiom metrics. Finally, SNPs with low MAF (< 0.05; $n = 334,975$) were excluded. The total number of SNPs passing QC was 411,787. The data are available in dbGaP (accession ID phs000920.v1.p1).

Whole-blood methylation data for a subset of the GALA II participants ($n = 573$) are publicly available in the Gene Expression Omnibus (GEO) database (accession number GSE77716) and have been described elsewhere [26, 36]. Briefly, methylation levels were measured using the Infinium HumanMethylation450K BeadChip array and raw methylation data were processed using the R minfi package [49] and assessed for basic quality control metrics, including determination of poorly performing probes with insignificant detection P-values above background control probes and exclusion of probes on X and Y chromosomes. Finally, beta-normalized values of the the data were SWAN normalized [50], corrected for batch using COMBAT [51] and adjusted for age, gender and chip assignment information using linear regression. The number of participants with both methylation and genotyping data was 525. We further excluded 46 individuals collected in a separate batch since they were all Puerto-Ricans. A total of 479 individuals and 473,838 probes remained for the analysis.

In order to further evaluate and validate the performance of EPISTRUCTURE we used data from the CHAMACOS longitudinal birth cohort study [52]. For this analysis, we had a subset of subjects that had Infinium HumanMethylation450K BeadChip array data available at 9 years of age. Briefly, samples were retained only if 95% of the sites assayed had insignificant detection P-value and samples demonstrating extremes level in the first two PCs of the data were removed. Probes where 95% of the samples had insignificant detection P-value (> 0.01; $n = 460$) as well as cross-reactive probes ($n = 29,233$) identified by Chen et al. [37] were dropped. A total of 227 samples and 455,590 probes remained for the analysis. Color channel bias, batch effects and difference in Infinium chemistry were minimized by application of All Sample Mean Normalization (ASMN) algorithm [53], followed by BMIQ normalization [44]. The data were adjusted for gender and technical batch information using linear regression.

In line with a previous study showing that a panel of small size is sufficient to approximate genetic ancestry in Latino populations well [54], 106 SNPs were collected and used as ancestry informative markers for estimating genetic ancestry of the CHAMACOS individuals [55]. The panel of ancestry informative markers was selected according to previously reported studies of Latino populations [25, 55, 56, 57]. Briefly, only SNPs with large differences in allele frequencies between ancestries were selected.

2.2.4 450K Human Methylation array

This methodology allows for examination of $> 450,000$ CpG sites across the genome, representing 99% of RefSeq genes. Sites include promoters, gene bodies, and 96% of UCSC database CpG islands (dense concentrations of CpGs), many of which are known to be associated with transcriptional control [58, 59, 60, 61, 62, 63]. This platform has been especially amenable to population studies because of its relative cost effectiveness and low sample requirements. Several studies have identified CpG sites differentially methylated by environmental exposures [64, 65] (e.g. arsenic and tobacco smoke) and health outcomes including obesity [66], rheumatoid arthritis [67], and Crohn's disease [68] demonstrating its utility in environmental and molecular epidemiology studies. The relative methylation (beta-normalized values) for each CpG site is calculated as the ratio of methylated-probe signal to total (methylated + unmethylated) fluorescent signal intensity. The Infinium pipeline is streamlined with excellent reproducibility [69].

2.2.5 Compiling a reference list from the KORA cohort

The reference list of genetically-informative CpGs was created using the KORA cohort for which whole-blood methylation data as well as genotype data were available for 1,799 European individuals. Following the algorithm described above, a score was computed for each CpG using k-fold cross-validation with $k = 10$ and using the parameters $w = 50$ and $p = 10$ (available in [70]). A reference list was then compiled from CpGs with median correlation of $R^2 > 0.5$ in the cross-validated prediction procedure, resulting in a total of 4,913 CpGs (available in [70]), out of which 2,436 are polymorphic CpGs and additional 801 CpGs have at least on common SNP in their probe outside the CpG target. The number of these reference CpGs available in the GALA II data set and in the CHAMACOS data set were 4,912 and 4,450, respectively. Removing probes with polymorphic CpGs results in 2,476 and 2,229 CpGs, and further removing probes with common SNPs results in 1,676 and 1,554 CpGs for GALA II and CHAMACOS, respectively. Unless stated otherwise, polymorphic CpGs were not excluded from the reference of informative CpGs in the executions of EPISTRUC-

TURE, therefore highly informative polymorphic CpGs ($R^2 > 0.5$) were also included in the reference list. In most cases, polymorphic CpGs are excluded as a preprocessing step in epigenetic studies, however, here we leverages the true genetic signal underlying in these probes for capturing the ancestry information better.

2.2.6 Detecting 450K probes containing SNPs

Probes with a SNP in their CpG target (polymorphic CpGs) were shown to be biased with underlying genetic polymorphisms rather than capture methylation signals solely [37]. The authors reported a list of 70,889 such polymorphic CpGs in the 450k DNA methylation array, as well as a list of common SNPs residing in probes of the 450K array outside the CpG target (MAF > 0.01 according to at least one of the major continental groups in the 1000 Genome database [71]). The total number of probes containing SNPs reported is 167,738.

2.2.7 Estimating ancestry information

Proportions of European, Native-American and African ancestries were estimated for each individual in both the GALA II and the CHAMACOS cohorts using the software ADMIXTURE [19] and the default reference data provided by the software. For the GALA II individuals we used the 411,787 SNPs remained after QC as an input, and for the CHAMACOS individuals we used the 106 available ancestry informative markers. The genotype based PCs were computed by applying PCA on the standardized values of the available genotypes in each data set. For the CHAMACOS data set, prior to computing PCA, we excluded sites with more than 5% missing values and completed the remaining missing values by assigning the mean. This resulted in a total of 99 SNPs.

2.2.8 Adjusting methylation levels for tissue heterogeneity

Methylation levels of the GALA II and CHAMACOS data sets were adjusted for cell-type composition using ReFACTor, a reference-free method for the correction of cell type heterogeneity in EWAS [36]. Each data set was adjusted for cell composition by regressing out the first six ReFACTor components, resulting in adjusted beta values. ReFACTor was executed using the default parameters and $K = 6$. For one of the experiments in the GALA II data we used an alternative approach for cell-type composition correction. Similarly, as with the ReFACTor components, we generated beta adjusted values, only this time we used reference-based cell-proportion estimates of main leukocyte cell types. Specifically, we obtained cell proportion estimates of six cell types (granulocytes, monocytes, B cells, NK cells, CD8T and CD4T cells) using the default implementation available in the minfi package [49], defined and assembled for the 450K array [72] based on the approach suggested by Houseman et al. [73] and a 450K reference data set [74].

2.2.9 Feature selection based on proximity to SNPs

For evaluating our suggested method, we calculated alternative methylation-based PCs after applying a feature selection that was previously suggested as a method for capturing population structure [34]. Following the authors' recommendation, we considered a list of the CpGs residing within 50 base pairs from SNPs, as provided by the authors.

2.3 Results

2.3.1 Inferring ancestry information from methylation using EPISTRUCTURE

EPISTRUCTURE selects a set of CpGs that are highly correlated with genotype information and then performs PCA on these sites while taking into account the cell-type composition effects and possibly other dominant factors that may affect genome-wide methylation. In order to compile a list of genetically-informative CpGs, we used the KORA cohort of European adults as reference data; these data include whole-blood methylation and genotyping data for a set of 1,799 individuals [42]. We fitted a regularized linear regression model for each CpG from SNPs in cis, and evaluated it based on a cross-validated linear correlation. Since the vast majority of reported CpG-SNP associations are between CpGs and cis-located

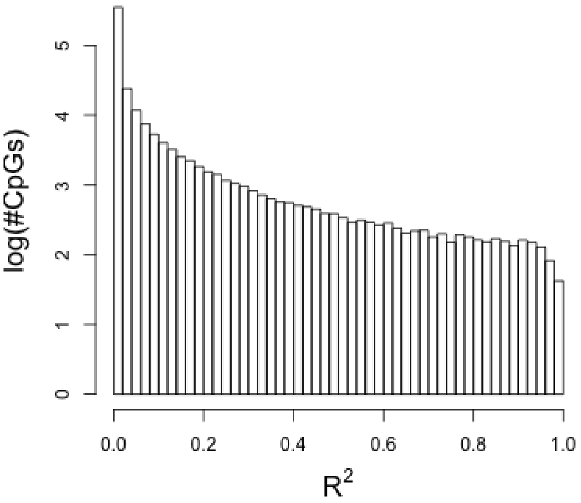

Figure 2.2: Correlation of methylation sites with cis-SNPs in the KORA data set. An R^2 score was calculated for each CpG available in the data from cis-SNPs. The results are presented in a log scaled histograms, showing that in most of the CpGs only a small portion of the variance can be explained by cis-SNPs.

SNPs [12, 23, 24], we only considered cis-located SNPs in capturing the genetic component of each CpG. We observed that for most CpGs only a small fraction of the variance can be explained by cis-SNPs ($R^2 < 0.1$ for 92.9% of the CpGs tested; Figure 2.2), thus motivating the use of only a relatively small subset of the CpGs for inferring ancestry information.

Considering only sites that most of their variance can be explained by cis-SNPs ($R^2 > 0.5$) resulted in a reference list of 4,913 genetically-informative CpGs (available in [70]). We note that polymorphic CpGs were not excluded from the KORA data set before learning the reference of informative CpGs, therefore polymorphic CpGs that can be well explained by cis-SNPs ($R^2 > 0.5$) were also included in the reference list. In most cases, polymorphic CpGs should be excluded before any data analysis, however, in our case, EPISTRUCTURE leverages the true genetic signal underlying in the probes of these CpGs. We later demonstrate the difference in performance when excluding these probes.

In order to test the performance of EPISTRUCTURE we applied it on the GALA II data set ($n = 479$), a pediatric Latino population study with Mexican and Puerto-Rican individ-

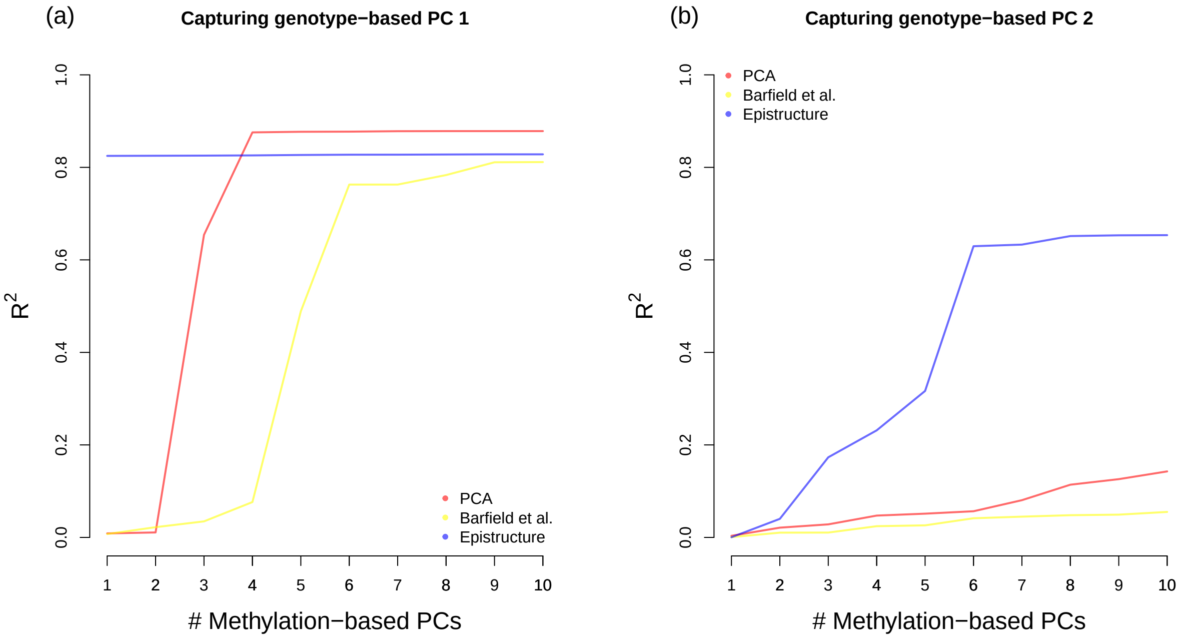

Figure 2.3: The fraction of variance explained in the first two genotype-based PCs of the GALA II data using several methods. Presented are linear predictors using increasing number of EPISTRUCTURE PCs (in blue), methylation-based PCs (in red) and methylation-based PCs after feature selection based on a previous study [34] (in yellow) for capturing (a) the first genotype-based PC and (b) the second genotype-based PC.

uals [75], for which both genotypes and 450K methylation array data (whole-blood) were available. First, we computed the largest (first) two PCs of the genotypes (genotype-based PCs), known to capture population structure [18]. We observed that EPISTRUCTURE provides substantially improved correlation with the first two genotype-based PCs as compared with the alternatives (Figure 2.3). Particularly, the first PC of EPISTRUCTURE captured the top genotype-based PC well ($R^2 = 0.82$), as compared to the first PC of the methylation data (methylation-based PC; $R^2 = 0.01$) and as compared to the methylation-based PC computed only from CpGs residing in close proximity to nearby SNPs ($R^2 = 0.01$), as was suggested in a recent study for capturing ancestry information in methylation data [34].

Next, as an alternative measure of population structure, we used the ADMIXTURE software [19] to estimate, for each individual, ancestry proportions of the three ancestries known to compose the Mexican and Puerto-Rican populations: European, Native-American and African. In this case, the top two principal components of EPISTRUCTURE capture well

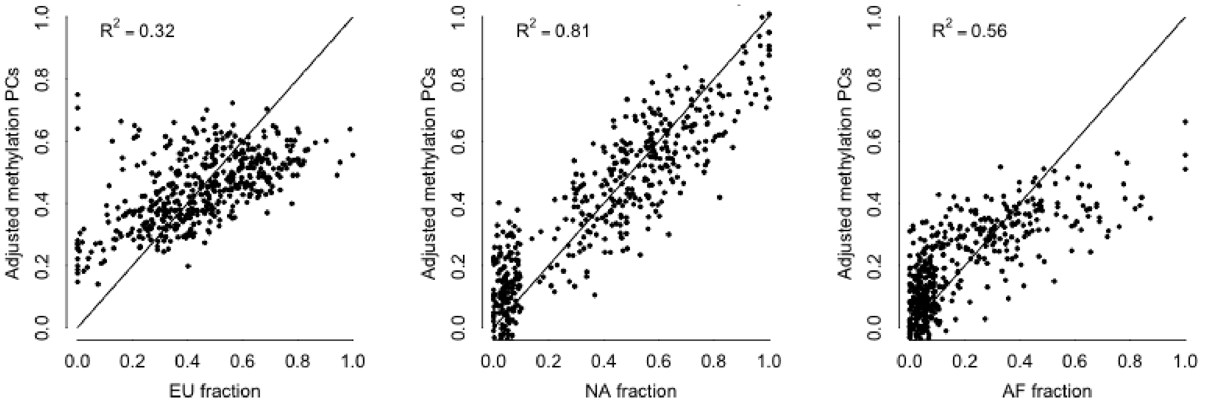

Figure 2.4: Capturing ancestry fraction estimates in the GALA II data using EPISTRUCTURE. Presented are linear predictors of European (EU), Native-American (NA) and African (AF) fraction estimates of the individuals in the data using the first two EPISTRUCTURE PCs.

both the Native American ancestry and the African Ancestry ($R^2 = 0.81$ and $R^2 = 0.56$ respectively), while the European ancestry was captured to a lesser extent ($R^2 = 0.32$; Figure 2.4).

We further tested whether ancestry information can be captured using EPISTRUCTURE in case there is a weaker population structure in the data. We observed that the first two PCs of EPISTRUCTURE could capture ancestry information well in both subpopulations of the GALA II data ($R^2 = 0.33$ in the Puerto-Rican group and $R^2 = 0.76$ in the Mexican group; Figure 2.5). These results suggest that EPISTRUCTURE can be used as an easy and efficient method for capturing ancestry information in methylation, even in data sets with relatively modest population structure.

2.3.2 Unsupervised ancestry inference from methylation data

EPISTRUCTURE is a supervised approach since it uses a reference data set in which both methylation and genotype data are available. In order to shed light on the extent to which ancestry is reflected by methylation, we also explored unsupervised approaches for the inference of ancestry from methylation data. Consistent with a previous study of individuals

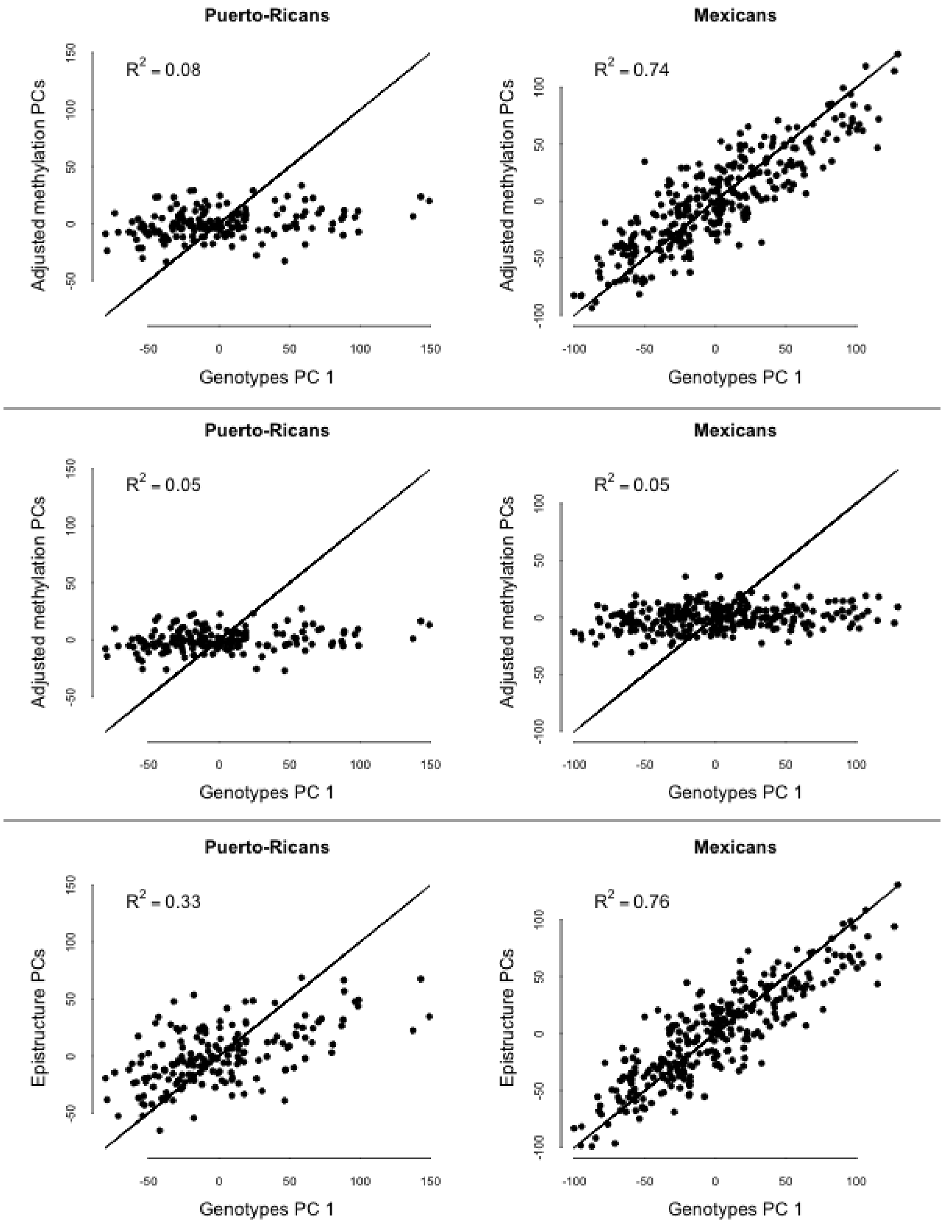

Figure 2.5: Capturing ancestry information in the GALA II data from Puerto-Rican individuals and from Mexican individuals separately. Presented are linear predictors of the first genotype-based PC using the first two methylation PCs computed from each subpopulation separately after adjusting the data for cell composition, before and after excluding probes containing SNPs from the data (top and middle panels, respectively) and using the first two EPISTRUCTURE PCs (bottom panel).

from the same population [76], the first two genotype-based PCs of the GALA II data clustered the samples into two groups, generally corresponding to Mexican and Puerto-Rican

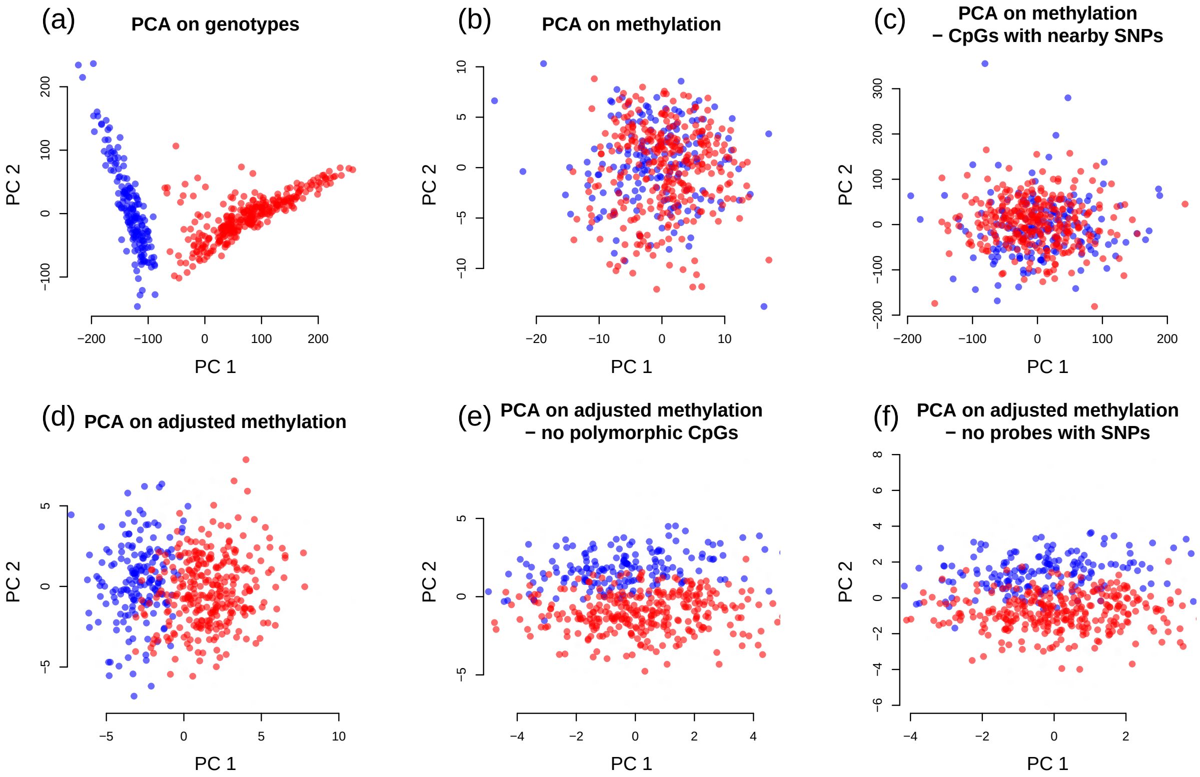

Figure 2.6: Capturing population structure in the GALA II data using an unsupervised approach. (a) The first two PCs of the genotypes, considered as the gold standard, separate the samples into two subpopulations: Puerto-Ricans (in blue) and Mexicans (in red). (b) The first two PCs of the methylation levels (methylation PCs) cannot reconstruct the separation found with the genotype data. (c) Recalculating the first two PCs after applying a feature selection based on proximity of CpGs to nearby SNPs as was proposed by Barfield et al. [34] (d) The first two PCs of the methylation after adjusting the data for cell-type composition (adjusted methylation PCs) can reconstruct most of the separation found in the genotypes. (e) Using adjusted methylation PCs after excluding the 70,889 polymorphic sites from data. (f) Using adjusted methylation PCs after excluding the 167,738 probes containing at least one common SNP.

subpopulations (Figure 2.6a). Since PCA has been shown to mirror ancestry very accurately in the case of genetic data [15], we first computed the top two methylation-based PCs while accounting for known technical factors as well as for age and sex, which are known to affect methylation genome-wide [39, 40, 77]. Considering the population structure characterized by the first two genotype-based PCs as the ”gold standard”, the first two methylation-based PCs could not sufficiently capture the population structure in the data (Figure 2.6b).

We then applied a few more sophisticated procedures, as follows. First, as before, we applied a feature selection step prior to calculating the methylation-based PCs according to a recent study, that suggested to consider only CpGs residing in close proximity to nearby SNPs in order to capture ancestry information in the first few PCs of the data [34]. We found that this procedure did not sufficiently reflect population structure in methylation data (Figure 2.6c). Next, since the first several PCs in methylation data coming from heterogeneous source such as blood are known to be dominated by cell-type composition variation [35, 36, 78], we adjusted the data for cell-type composition using ReFACTor [36] and recalculated the first two PCs. This approach effectively reconstructed most of the separation determined by the genotype-based PCs (Figure 2.6d).

These results show that 450K-probed methylation data indeed reflect genotype data well. Specifically, after accounting properly for known confounders, the top methylation-based PCs capture the genotype-based PCs. However, these results can potentially be driven by artifacts. Particularly, many probes in the 450K methylation array contain single nucleotide polymorphisms (SNPs) in their target CpGs. Such polymorphic CpGs were shown to bias measured methylation levels as a function of the individual's genotypes, apparently due to changes in probe binding specificity [37]. Thus, the results above might be biased by these probes. To address this possibility, we recalculated the first two methylation-based PCs after excluding 70,889 CpGs that are known to be polymorphic. We found that the new methylation-based PCs could still capture well the first genotype-based PC ($R^2 = 0.77$ as opposed to $R^2 = 0.83$ before removing the polymorphic CpGs), accounting for the separation evident from the first two genotype-based PCs (Figure 2.6e). In addition, we performed a more conservative analysis by repeating the last step, only this time we excluded all probes containing at least one common SNP anywhere on the probe (i.e. not only in the target CpG but anywhere on the prove; in total, 167,738 probes). We found that in this case as well the reconstruction using the top two methylation-based PCs provided almost the same separation determined by the genotype-based PCs (Figure 2.6f; $R^2 = 0.70$ with the first genotype-based PC).

We note that repeating the last two experiments while accounting for estimated cell proportions computed using a commonly applied reference-based method [73] as an alternative approach for correction of cell composition effects in methylation could not achieve the same results ($R^2 = 0.23$ and $R^2 = 0.14$ in the experiments without the polymorphic sites and in the experiment removing all probes with common SNPs, respectively). This can be explained by the additional cell-type composition signal captured by ReFACTor but not by the reference-based approach, as was previously demonstrated on the GALA II data [36]. Substantial difference in performance is especially expected in cases where the reference methylation data used by the reference-based method do not represent the target population well [36, 79]. Removing only part of the cell-type composition signal from the data results in PCs that are likely to be still dominated by tissue composition information rather than by population structure. Alternatively, it may be the case that ReFACTor also removed another sparse confounder, in addition to the cell-type composition signal.

We also compared the different approaches using the ancestry estimates of the ADMIXTURE software [19]. The results were consistent with our previous experiment - while the first two methylation-based PCs could not capture the ancestry estimates ($R^2 = 0.02$ with European, $R^2 = 0.01$ with Native-American and $R^2 = 0.02$ with African fractions), we found the first two methylation-based PCs after adjusting for cell composition to capture the ancestry estimates well, even after excluding from the data all probes containing common SNPs ($R^2 = 0.28$ with the European fraction, $R^2 = 0.69$ with Native-American and $R^2 = 0.47$ with African; Figure 2.7).

We further tested whether ancestry information can be captured in the same manner when applied to each of the two subpopulations in the data (Mexican and Puerto-Rican) separately. We found the methylation-based PCs to capture well only the first genotype-based PC of the Mexican group when not excluding probes containing common SNPs ($R^2 = 0.08$ for the Puerto-Rican cluster and $R^2 = 0.74$ for the Mexican cluster). After excluding the 167,738 probes containing at least one common SNP from the data, the methylation-based PCs could not capture a substantial fraction of the first genotype-based PC in either clusters ($R^2 = 0.05$ for the Puerto-Rican cluster and $R^2 = 0.05$ also for the Mexican cluster). Thus,

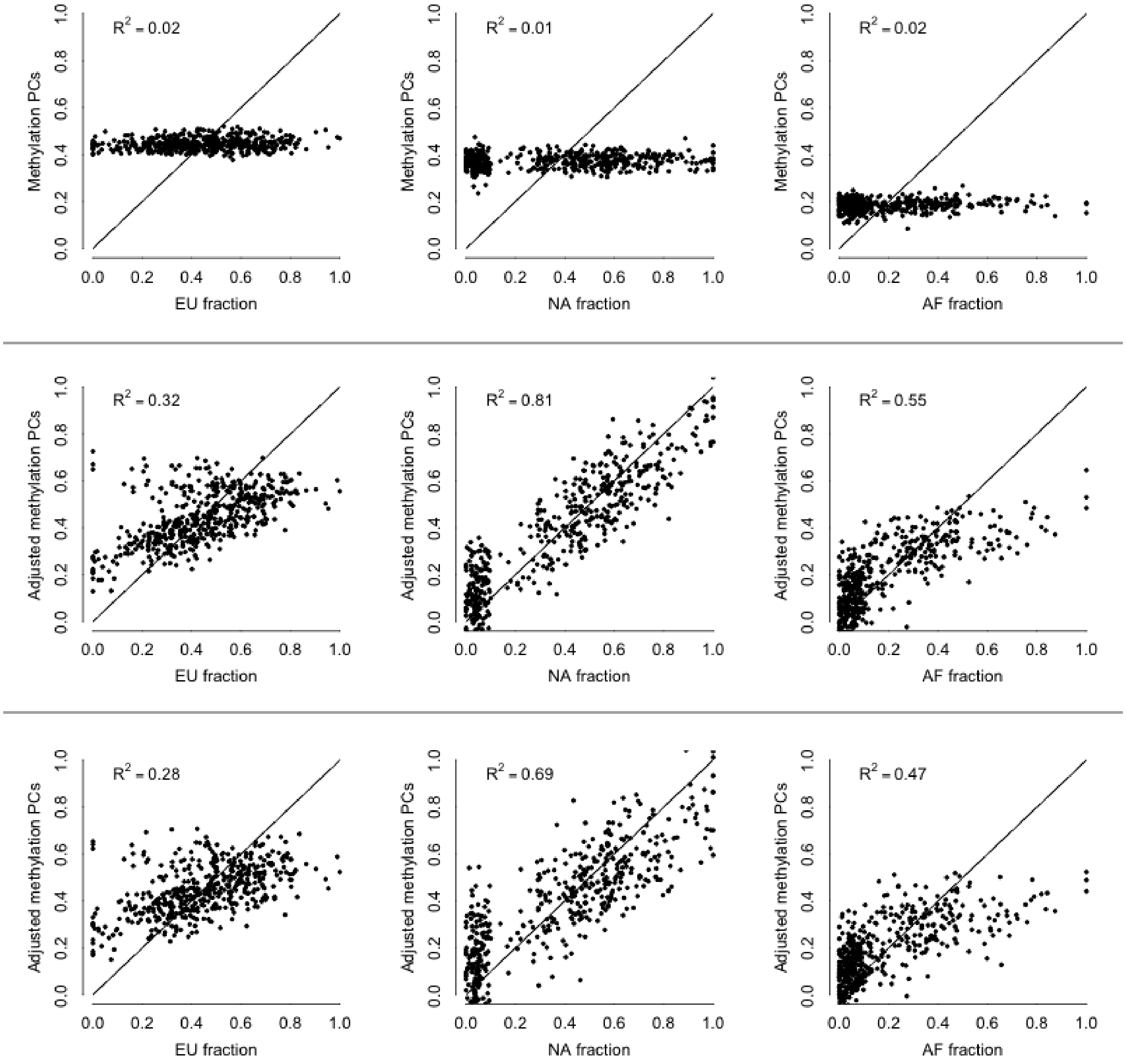

Figure 2.7: Capturing ancestry fraction estimates in the GALA II data. Presented are linear predictors of European (EU), Native-American (NA) and African (AF) fraction estimates of the individuals in the data using the first two methylation PCs of the data (top panel), the first two PCs after adjusting the data for cell composition (adjusted methylation PCs; middle panel) and using the adjusted methylation PCs after excluding from the data all probes containing SNPs (bottom panel).

we conclude that under weak population structure the current unsupervised approach does not mirror ancestry well. However, as we demonstrated earlier, the supervised approach of

EPISTRUCTURE, using only a relatively small subset of highly informative CpGs (including highly informative polymorphic CpGs), performed well in this case.

2.3.3 Validation using the CHAMACOS study data

We further validated the effectiveness of EPISTRUCTURE and the unsupervised approaches using data from the primarily Mexican-American CHAMACOS cohort [52, 80]. We used whole-blood methylation levels from nine years old participants (n=227) for which we had 106 ancestry informative markers [55], previously shown to approximate ancestry information well in another Hispanic admixed population [81].

We computed the first two PCs of the available ancestry informative markers (genotype-based PCs) in order to capture the ancestry information of the samples. Since the CHAMACOS cohort primarily consists of Mexican-American individuals, we observed no separation into distinct subpopulations in the first several genotype-based PCs. We then computed the first two methylation-based PCs, before and after adjusting the data for cell composition. In addition, we computed the first two EPISTRUCTURE PCs of the data, and measured how much of the variance of the first genotype-based PC can be explained by each of the approaches. As shown in Figure 2.8, the first two methylation-based PCs could capture only a small portion of the first genotype-based PC ($R^2 = 0.04$ before adjusting for cell composition and $R^2 = 0.16$ after adjusting for cell composition), as opposed with the first two EPISTRUCTURE PCs which could capture the first genotype-based PC substantially better ($R^2 = 0.38$). As in the GALA II data, applying feature selection based on proximity of CpGs to SNPs [34] could capture only a small portion of the ancestry information ($R^2 = 0.05$).

As before, we used the ADMIXTURE software [19] as an alternative measure of population structure. For each individual we estimated the ancestry proportions of the three ancestries known to compose Mexican individuals: European, Native-American and African. The first two EPISTRUCTURE PCs were found to explain a large portion of the European and Native-American fraction estimates ($R^2 = 0.46$ for European and $R^2 = 0.6$ for Native-American ancestry), as opposed with the first two methylation-based PCs ($R^2 = 0.11$ for

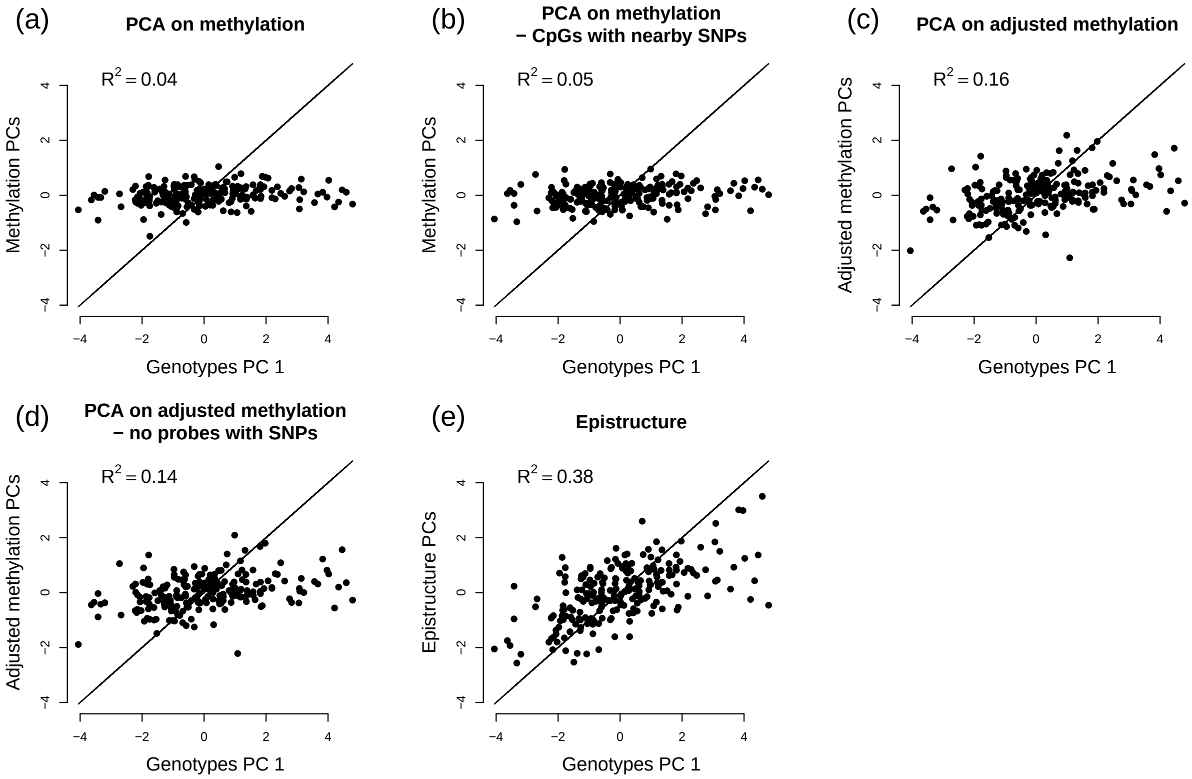

Figure 2.8: Capturing population structure in the CHAMACOS data. Presented are linear predictors of the first genotype-based PC using (a) the first two methylation PCs of the data, (b) the first two PCs calculated after applying a feature selection based on proximity of CpGs to nearby SNPs [34], (c) the first two PCs after adjusting the data for cell-type composition (adjusted methylation PCs), (d) the first two adjusted methylation PCs after excluding 167,738 probes containing SNPs from the data, and (e) the using the first two EPISTRUCTURE PCs.

European and $R^2 = 0.14$ for Native-American ancestry, after adjusting for cell-type composition; Figure 2.9). The estimates of African proportions, however, were not captured well by either approach. This result was expected due to the low average proportion of African ancestry in Mexican samples (less than 10%) [47]. All the results are summarized in Table 2.1.

2.3.4 Implications for the EPIC array

The recently introduced EPIC array by Illumina, which allows to probe a set of approximately 850K CpGs, is likely to be used in many future methylation data collection efforts. Since

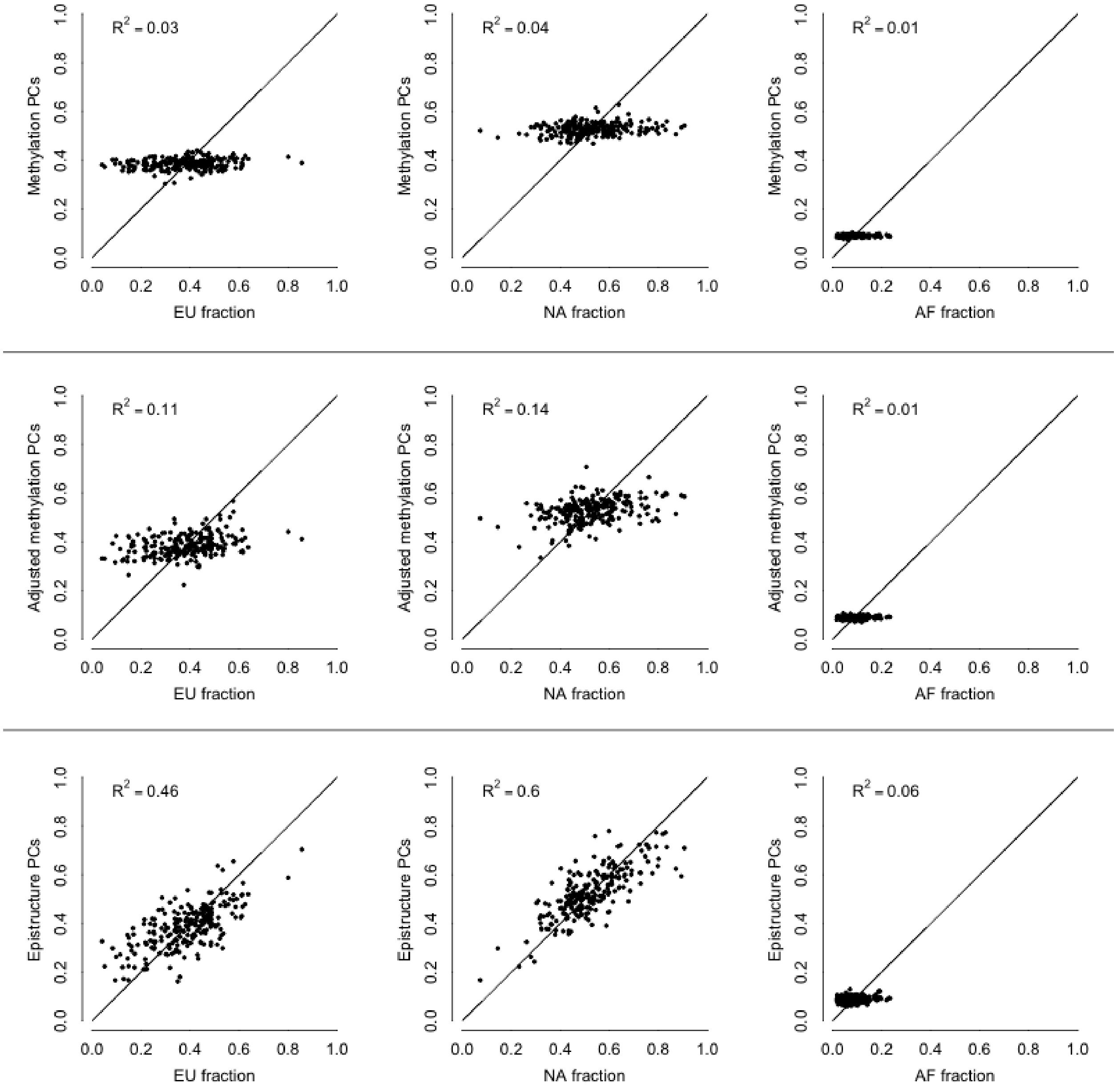

Figure 2.9: Capturing ancestry fraction estimates in the CHAMACOS data set. Presented are linear predictors of European (EU), Native-American (NA) and African (AF) fraction estimates of the individuals in the data using the first two methylation PCs (top panel), the first two PCs after adjusting the data for cell-type composition (adjusted methylation PCs; middle panel) and using the first two EPISTRUCTURE PCs (bottom panel). The methylation PCs in this experiment were computed without excluding probes containing SNPs from the data.

genotype data and corresponding EPIC array data for the same individuals were not publicly available at the time of this study, we were not able to compile a reference list of CpGs for the EPIC array. However, inspection of the probes available in each array reveals that only 32,425 of the probes in the 450K array were not included in the EPIC array. We further found that 94% of the CpGs in the 450K-based reference list we constructed (4,616 CpGs out of 4,913) were included in the EPIC array. Therefore, our suggested 450K-based reference list is expected to perform similarly on data generated from the EPIC array. In order to test that, we repeated all of the experiments we performed so far, only this time we removed from the data the set of 32,425 sites that were not included in the EPIC array. The results, summarized in Table 2.1, show that removing these sites leads to only a marginal decrease in the R^2 values. Clearly, as more EPIC array data will become available, EPIC-based reference list of CpGs is expected to further improve the performance of EPISTRUCTURE.

2.4 Discussion

We demonstrated that 450K DNA methylation data can capture population structure in admixed populations. Particularly, we observed that in the presence of a relatively strong population structure (GALA II) the dominant genome-wide signal of ancestry information could be revealed in an unsupervised manner once appropriately correcting for tissue heterogeneity. In contrast, we observed that in the presence of weaker population structure in the data (CHAMACOS) the genome-wide signal of ancestry methylation is only moderately reflected by the dominant axes of variation in the data, even after accounting for tissue heterogeneity.

Using KORA, a large data set for which both methylation levels and genotypes were available, we generated a reference list of genetically-informative CpGs and successfully used it to estimate ancestry information in new data sets by applying PCA on the reference sites. Polymorphic CpGs that were found to be highly correlated with genetics were also include in the reference list. Although these CpGs are generally treated as artifacts, they represent true genetic signal and therefore were used in order to further increase the signal captured by EPISTRUCTURE. As we showed, by taking this approach, EPISTRUCTURE was able to effectively isolate and capture ancestry information in methylation data.

While we observed strong correlations between the EPISTRUCTURE PCs and the genotype-based population structure estimates of the GALA II individuals, only moderate correlations were found in the CHAMACOS data set (though substantially better than alternative approaches, in which only negligible correlations with the true ancestry signal were found). These results can be explained in part by the fact that only 106 ancestry informative markers were available for us in the CHAMACOS for capturing ancestry information, as opposed with the dense genotype array information used in the GALA II analysis. Therefore, it is likely that our inference of population structure by methylation data is in fact more accurate than reflected in the experiments conducted on the CHAMACOS samples.

The reference-list of CpGs was generated using methylation states and genotypes collected from European individuals, therefore it may not be optimized for capturing ancestry information in non-European populations. However, since we successfully used this list for the inference of ancestry in the Latino GALA II and CHAMACOS individuals, we expect it to prove useful for some other non-European populations as well.

Finally, we note that when constructing the linear models for each CpG from its cis-SNPs in the whole-blood KORA data, we decided not to account for tissue heterogeneity. To the best of our knowledge, there is currently no evidence for dramatic genome-wide effects of genotypes on the cell-type composition. Therefore, in the vast majority of CpGs, the cell-type composition is expected to be orthogonal to the genetic signal they contain. As a result, accounting for tissue heterogeneity in this case is more likely to reduce the accuracy of the model due to inaccuracies of the cell-type composition estimates rather than to bias the selection of CpGs into the reference-list.

Data Set	Meth PCs	Adj PCs	Adj PCs II	Barfield et al.	EPISTRUCTURE	Measurement
GALA II	$R^2 = 0.01$	$R^2 = 0.83$	$R^2 = 0.70$	$R^2 = 0.02$	$R^2 = 0.83$	Genotype-based PC 1
	$R^2 = 0.02$	$R^2 = 0.32$	$R^2 = 0.27$	$R^2 = 0.03$	$R^2 = 0.32$	EU fraction
	$R^2 = 0.01$	$R^2 = 0.81$	$R^2 = 0.69$	$R^2 = 0.03$	$R^2 = 0.81$	NA fraction
	$R^2 < 0.01$	$R^2 = 0.79$	$R^2 = 0.67$	$R^2 < 0.01$	$R^2 = 0.78$	AF fraction
CHAMACOS	$R^2 = 0.04$	$R^2 = 0.15$	$R^2 = 0.14$	$R^2 = 0.05$	$R^2 = 0.38$	Genotype-based PC 1
	$R^2 = 0.03$	$R^2 = 0.11$	$R^2 = 0.08$	$R^2 = 0.01$	$R^2 = 0.46$	EU fraction
	$R^2 = 0.04$	$R^2 = 0.14$	$R^2 = 0.11$	$R^2 = 0.01$	$R^2 = 0.60$	NA fraction
	$R^2 = 0.01$	$R^2 = 0.01$	$R^2 = 0.01$	$R^2 = 0.01$	$R^2 = 0.06$	AF fraction
GALA II 450K-specific CpGs excluded	$R^2 = 0.01$	$R^2 = 0.83$	$R^2 = 0.70$	$R^2 = 0.04$	$R^2 = 0.82$	Genotype-based PC 1
	$R^2 = 0.02$	$R^2 = 0.32$	$R^2 = 0.28$	$R^2 = 0.03$	$R^2 = 0.31$	EU fraction
	$R^2 = 0.01$	$R^2 = 0.81$	$R^2 = 0.69$	$R^2 = 0.04$	$R^2 = 0.80$	NA fraction
	$R^2 = 0.01$	$R^2 = 0.55$	$R^2 = 0.46$	$R^2 = 0.03$	$R^2 = 0.55$	AF fraction
CHAMACOS 450K-specific CpGs excluded	$R^2 = 0.04$	$R^2 = 0.15$	$R^2 = 0.14$	$R^2 = 0.05$	$R^2 = 0.33$	Genotype-based PC 1
	$R^2 = 0.03$	$R^2 = 0.11$	$R^2 = 0.08$	$R^2 = 0.04$	$R^2 = 0.45$	EU fraction
	$R^2 = 0.04$	$R^2 = 0.14$	$R^2 = 0.11$	$R^2 = 0.04$	$R^2 = 0.58$	NA fraction
	$R^2 = 0.01$	$R^2 = 0.01$	$R^2 = 0.01$	$R^2 = 0.04$	$R^2 = 0.08$	AF fraction

Table 2.1: Summary of the results in the GALA II data set and in the CHAMACOS data set. In the first part of the table, squared linear correlations were measured between several measurements of ancestry information and linear predictors using the first two PCs of the data (Meth PCs), the first two PCs after adjusting the data for cell-type composition (Adj PCs), the first two PCs after adjusting the data for cell-type composition and excluding probes containing SNPs from the data (Adj PCs II), the first two PCs when considering only CpGs in close proximity to SNPs (Barfield et al.) and the first two EPISTRUCTURE PCs. The second part of the table presents the results of the same experiments, only after excluding all the CpGs of the 450K array that were not included in the EPIC methylation array.

CHAPTER 3

Refernce-free estimation of cell-type composition from DNA methylation

3.1 Background

DNA methylation status has become a prominent epigenetic marker in genomic studies, and genome-wide DNA methylation data have become ubiquitous in the last few years. Numerous recent studies provide evidence for the role of DNA methylation in cellular processes and in disease (e.g., in multiple sclerosis [82], schizophrenia [83], and type 2 diabetes [84]). Thus, DNA methylation status holds great potential for better understanding the role of epigenetics, potentially leading to better clinical tools for diagnosing and treating patients.

In a typical DNA methylation study, we obtain a large matrix in which each entry corresponds to a methylation level (a number between 0 and 1) at a specific genomic position for a specific individual. This level is the fraction of the probed DNA molecules that were found to have an additional methyl group at the specific position for the specific individual. Essentially, these methylation levels represent, for each individual and for each site, the probability of a given DNA molecule to be methylated. While simple in principle, methylation data are typically complicated owing to various biological and non-biological sources of variation. Particularly, methylation patterns are known to differ between different tissues and between different cell types. As a result, when methylation levels are collected from a complex tissue (e.g., blood), the observed methylation levels collected from an individual reflect a mixture of its methylation signals coming from different cell types, weighted according to mixing proportions that depend on the individual's cell-type composition. Thus, it is challenging to interpret methylation signals coming from heterogeneous sources.

One notable challenge in working with heterogeneous methylation levels has been highlighted in the context of Epigenome-Wide Association Studies (EWAS), where data are typically collected from heterogeneous samples. In such studies, we typically search for rows of the methylation matrix (each corresponding to one genomic position) that are significantly correlated with a phenotype of interest across the samples in the data. In this case, unless accounted for, correlation of the phenotype of interest with the cell-type composition of the samples may lead to numerous spurious associations and potentially mask true signal [72]. In addition to its importance for a correct statistical analysis, knowledge of the cell-type composition may provide novel biological insights by studying cell compositions across populations.

In principle, one can use high-resolution cell counting for obtaining knowledge about the cell composition of the samples in a study. However, unfortunately, such cell counting for a large cohort may be costly and often logistically impractical (e.g., in some tissues, such as blood, reliable cell counting can be obtained from fresh samples only). Due to the pressing need to overcome this limitation, development of computational methods for estimating cell-type composition from methylation data has become a key interest in epigenetic studies. Several such methods have been suggested in the past few years [36, 85, 86, 87, 88, 89], some of which aim at explicitly estimating cell-type composition, while others aim at a more specific goal of correcting methylation data for the potential cell-type composition confounder in association studies. These methods take either a supervised approach, in which reference data of methylation patterns from sorted cells (methylomes) are obtained and used for predicting cell compositions [85], or an unsupervised approach (reference-free) [36, 86, 87, 88, 89].

The main advantage of the reference-based method is that it provides direct (absolute) estimates of the cell counts, whereas, as we demonstrate here, current reference-free methods are only capable of inferring components that capture linear combinations of the cell counts. Yet, the reference-based method can only be applied when relevant reference data exist.

Currently, reference data only exist for blood [74], breast [90] and brain [91], for a small number of individuals (e.g., six samples in the blood reference [74]). Moreover, the individuals in most available data sets do not match the reference individuals in their methylation-altering factors, such as age [39], gender [40, 77], and genetics [70]. This problem was recently highlighted in a study in which the authors showed that available blood reference collected from adults failed to estimate cell proportions of newborns [79]. Furthermore, in a recent work, we showed evidence from multiple data sets that a reference-free approach can provide substantially better correction for cell composition when compared with the reference-based method [92]. It is therefore often the case that unsupervised methods are either the only option or are a better option for the analysis of EWAS.

As opposed to the reference-based approach, although can be applied for any tissue in principle, the reference-free methods do not provide direct estimates of the cell-type proportions. Previously proposed reference-free methods allow us to infer a set of components, or general axes, which were shown to compose linear combinations of the cell-type composition [36, 88]. Another more recent reference-free method was designed to infer cell-type proportions, however, as we show here, it only provides components that compose linear combinations of the cell-type composition rather than direct estimates [89]. Unlike cell proportions, while linearly correlated components are useful in linear analyses such as linear regression, they cannot be used in any nonlinear downstream analysis or for studying individual cell types (e.g., studying alterations in cell composition across conditions or populations). Cell proportions may provide novel biological insights and contribute to our understanding of disease biology, and we therefore need targeted methods that are practical and low in cost for estimating cell counts.

In attempt to address the limitations of previous reference-free methods and to provide cell count estimates rather than linear combinations of the cell counts, we propose an alternative Bayesian strategy that utilizes prior knowledge about the cell-type composition of the studied tissue. We present a semi-supervised method, BayesCCE (Bayesian Cell Count Estimation), which encodes experimentally obtained cell count information as a prior on the distribution of the cell-type composition in the data. As we demonstrate here, the required prior is substantially easier to obtain compared with standard reference data from sorted cells. We can estimate this prior from general cell counts collected in previous studies, without the need for corresponding methylation data or any other genomic data.

We evaluate our method using four large methylation data sets and simulated data, and show that our method produces a set of components that can be used as cell count estimates. We observe that each component of BayesCCE can be regarded as corresponding to scaled values of a single cell type (i.e. high absolute correlation with one cell type, but not necessarily good estimates in absolute terms). We find that BayesCCE provides a substantial improvement in correlation with the cell counts over existing reference free methods (in some cases a 50% improvement). We also consider the case where both methylation and cell count information are available for a small subset of the individuals in the sample, or for a group of individuals from external data. Notably, existing reference-based and reference-free methods for cell type estimation completely ignore this potential information. In contrast, our method is flexible and allows to incorporate such information. Specifically, we show that our proposed Bayesian model can leverage such additional information for imputing missing cell counts in absolute terms. Testing this scenario on both real and simulated data, we find that measuring cell counts for a small group of samples (a couple of dozens) can lead to a further significant increase in the correlation of BayesCCE's components with the cell counts.

3.2 Methods

3.2.1 Notations and related work

Let $X \in \mathbb{R}^{n \times m}$ be an n samples by m sites matrix of DNA methylation levels coming from a heterogeneous source consisting k cell types. For methylation levels, we consider what is commonly referred to as beta-normalized methylation levels, which are defined for each sample in each site as the proportion of methylated probes out of the total number of probes. Put differently, $X_{ij} \in [0, 1]$ for each site j and sample i. We denote $Z \in \mathbb{R}^{k \times m}$ as the cell-type-specific mean methylation levels for each site, $W \in \mathbb{R}^{k \times n}$ as the cell-type proportions of the samples in the data, and we denote a column vector of a matrix M by M_j for the j-th vector.

Given the above notations, a common model for heterogeneous DNA methylation (i.e. mixtures) is

$$\begin{align} & X_{ij} = W_i^T Z_j + \epsilon_{ij} \tag{3.1} \\ & \epsilon_{ij} \sim N(0, \sigma^2) \tag{3.2} \\ & \forall i \forall h : W_{hi} \geq 0 \tag{3.3} \\ & \forall i : \sum_{h=1}^{k} W_{hi} = 1 \tag{3.4} \\ & \forall j \forall h : 0 \leq Z_{hj} \leq 1 \tag{3.5} \end{align}$$

where the error term ϵ_{ij} models measurement noise and other possible unmodeled factors. The constraints in (3.3) and in (3.4) require the cell-type proportions to be positive and to sum up to one in each sample, and the constraints in (3.5) require the cell-type-specific mean levels to be in the range $[0, 1]$. This model was initially suggested for DNA methylation in the context of reference-based estimation of cell proportions by Houseman et al. [85]. We are interested in estimating W. Taking a standard maximum-likelihood approach for fitting the model results in the following optimization problem:

$$\begin{align} \hat{W}, \hat{Z} = \underset{W,Z}{\operatorname{argmin}} \quad & \|O - W^T Z\|_F^2 \tag{3.6} \\ \text{s.t} \quad & \forall i \forall h : W_{hi} \geq 0 \tag{3.7} \\ & \forall i : \sum_{h=1}^{k} W_{hi} = 1 \tag{3.8} \\ & \forall j \forall h : 0 \leq Z_{hj} \leq 1 \tag{3.9} \end{align}$$

where $\| \cdot \|_F^2$ is the squared Frobenius norm. The reference-based method [85] first obtains an estimate of Z from reference methylation data collected from sorted cells of the cell types composing the studied tissue. Once an estimate of Z is fixed, W can be estimated by solving a standard quadratic program.

If the matrix Z is unknown, which is a reference-free version of the problem, the above formulation of the problem can be regarded as a version of non-negative matrix factorization (NNMF) problem. NNMF has been suggested in several applications in biology; notably, the problem of inference of cell-type composition from methylation data has been recently formulated as an NNMF problem [88]. In order to optimize the model, the authors used an alternative optimization procedure in which Z or W are optimized while the other is kept fixed. However, as demonstrated by the authors [88], this solution results in the inference of a linear combination of the cell proportions W. Put differently, more than one component of the NNMF is required for explaining each cell type in the data.

The inability of NNMF to provide one component per cell type was recently highlighted and explained using geometric considerations [89], which nicely showed the non-identifiable nature of the NNMF model in (3.6) in case that a perfect factorization of X into Z, W exists (i.e. $X = W^T Z$). However, in practice, perfect factorization never exists in real biological data. Thus, in addition to empirical evidence from several data sets on which we apply the NNMF method (see Results), in the next subsection we provide a mathematical proof for the non-identifiability of the NNMF model in (3.6) under a more general case, where a perfect factorization does not necessarily exist.

In an attempt to overcome the non-identifiability of the model in (3.6) and to provide cell-type proportions when reference methylation data are not available, a recent modification of the NNMF model has been suggested [89]. The method, MeDeCom, solves the optimization of the NNMF model while including additional penalty term in the objective function. Derived from biological knowledge about mean methylation levels, the penalty negatively weights mean methylation levels diverging from a known bimodal behavior of methylation levels, wherein CpGs tend to be overall methylated or unmethylated [89]. While the modified objective suggested in MeDeCom overcomes the non-identifiability of the NNMF model for a given weight of the penalty (λ), it is not entirely clear how to select λ. To circumvent this problem, the authors proposed a cross-validation procedure for the selection of λ. However, our empirical results from four large whole-blood methylation data sets, as well as from simulated data, show sub-optimal performance for MeDeCom, similarly to the solutions of the simpler NNMF model. Our results suggest that the modification introduced by MeDeCom may not effectively avoid the non-identifiability nature of the NNMF model, possibly due to insufficient prior information or inability to effectively determine an appropriate value for λ.

Another recent reference-free method for estimating cell composition in methylation data, ReFACTor [36], performs an unsupervised feature selection step followed by a principal components analysis (PCA). Similarly to the NNMF solution, ReFACTor is an unsupervised method and it only finds principal components (PCs) that form linear combinations of the cell proportions rather than directly estimates the cell proportion values [36].

3.2.2 Non-identifiability of the NNMF model

We hereby show by construction the non-identifiability nature of the NNMF model in (3.6). For this proof, instead of the constraints in (3.9), we consider a slightly modified version of the constraints:

$$\forall j \forall h : 0 < Z_{hj} < 1 \tag{3.10}$$

While in theory we may have an equality (i.e. $Z_{hj} = 0$ or $Z_{hj} = 1$), in practice, such sites are typically not measured or excluded from the analysis, since they would not be demonstrating any variability.

Proposition: Let $\hat{W}, \hat{Z}$ be a solution to the problem in (3.6). There exist $\tilde{W} \neq \hat{W}, \tilde{Z} \neq \hat{Z}$ such that $\|X - \hat{W}^T \hat{Z}\|_F^2 = \|X - \tilde{W}^T \tilde{Z}\|_F^2$ and the constraints in (3.7), (3.8) and in (3.10) are satisfied.

Proof:

Let $0 < c < 1$, define $Q \in \mathbb{R}^{k \times k}$ to be the identity matrix up to two entries: $Q_{11} = 1-c, Q_{12} = c$. It follows that Q^{-1} is also the identity matrix up to two entries: $Q_{11}^{-1} = \frac{1}{1-c}, Q_{12}^{-1} = \frac{c}{c-1}$.

Denote $\tilde{W} = Q^T\hat{W}$ and denote $\tilde{Z} = Q^{-1}\hat{Z}$, we get that

$$\|X - \tilde{W}^T\tilde{Z}\|_F^2 = \|X - \hat{W}^T QQ^{-1}\hat{Z}\|_F^2 = \|X - \hat{W}^T\hat{Z}\|_F^2$$

The constraints in (3.7) hold since $\tilde{W}_{hi} \geq 0$ for each $1 \leq i \leq n, 1 \leq h \leq k$. The constraints in (3.8) hold since for each $1 \leq i \leq n$

$$\sum_{l=1}^{k} \tilde{W}_{li} = \sum_{l=1}^{k}\sum_{h=1}^{k} Q_{lh}^T \hat{W}_{hi} = (1-c)\hat{W}_{1i} + c\hat{W}_{1i} + \sum_{h=2}^{k} \hat{W}_{hi} = \sum_{h=1}^{k} \hat{W}_{hi} = 1$$

In addition, $\tilde{Z}_{hj} \in (0,1)$ for $2 \leq h \leq k, 1 \leq j \leq m$. In order to completely satisfy the constraints in (3.10), we also require these constraints to be satisfied for $h = 1, 1 \leq j \leq m$. It is easy to see that for each j the latter is satisfied if

$$0 < c < min\left\{\frac{1-\hat{Z}_{1j}}{1-\hat{Z}_{2j}}, \frac{\hat{Z}_{1j}}{\hat{Z}_{2j}}\right\}$$

Therefore, we can simply select a value of c in the range

$$0 < c < min_j\left\{min\left\{\frac{1-\hat{Z}_{1j}}{1-\hat{Z}_{2j}}, \frac{\hat{Z}_{1j}}{\hat{Z}_{2j}}\right\}\right\}$$

Note that we necessarily have either

$$0 < c < min_j\left\{min\left\{\frac{1-\hat{Z}_{1j}}{1-\hat{Z}_{j1}}, \frac{\hat{Z}_{1j}}{\hat{Z}_{2j}}\right\}\right\} < 1$$

or

$$0 < c < min_j\left\{min\left\{\frac{1-\hat{Z}_{2j}}{1-\hat{Z}_{1j}}, \frac{\hat{Z}_{2j}}{\hat{Z}_{1j}^T}\right\}\right\} < 1$$

In the latter case we can switch the positions of the first two columns in Z. Equality of the minimum to 1 in both cases would mean that the first two rows of Z are identical, which would mean that the problem is non-identifiable, as the first two cell types cannot be distinguished in this scenario. As a result of the above, the constraints in (3.10) can be satisfied for a range of values of c. □

3.2.3 The model

We suggest a more detailed model by adding a prior on W and taking into account potential covariates. Specifically, we assume that

$$W_i \sim Dirichlet(\alpha_1, ..., \alpha_k) \tag{3.11}$$

where $\alpha_1, ..., \alpha_k$ are assumed to be known. In practice, the parameters are estimated from external data in which cell-type proportions of the studied tissue are known. Such experimentally obtained cell-type proportions were used to test the appropriateness of the Dirichlet prior in describing cell composition distribution (data not shown). Also, we consider additional factors of variation affecting observed methylation levels, in addition to variation in cell-type composition. Specifically, denote $C \in \mathbb{R}^{n \times p}$ as a matrix of p covariates for each individual and $S \in \mathbb{R}^{m \times p}$ as a matrix of corresponding effects of the p covariates on each of the m sites. As before, we are interested in estimating W, the cell-type proportions of the k cell types. Deriving a maximum likelihood-based solution for this model and repeating the constraints for completeness results in the following optimization problem:

$$\begin{aligned}
\hat{W}, \hat{Z}, \hat{S} = \underset{W,Z,S}{\operatorname{argmin}} \quad & \frac{1}{2\sigma^2}\|X - W^T Z - CS^T\|_F^2 - \sum_{h=1}^{k}(\alpha_h - 1)\sum_{i=1}^{n}\log(W_{hi}) && (3.12)\\
\text{s.t} \quad & \forall i \forall h : W_{hi} \geq 0 && (3.13)\\
& \forall i : \sum_{h=1}^{k} W_{hi} = 1 && (3.14)\\
& \forall j \forall h : 0 \leq Z_{hj} \leq 1 && (3.15)
\end{aligned}$$

Our intuition in this model is that since the priors on W are estimated from real data, incorporating them will push the solution of the optimization to return estimates of W which are closer to the true values as opposed to a linear combination of them.

3.2.4 The BayesCCE algorithm

Our algorithm uses ReFACTor as a starting point, and we estimate W by finding an appropriate linear transformation of the ReFACTor principal components (ReFACTor components). In principle, any of the reference-free methods we examined (ReFACTor, NNMF and MeDeCom) could be used as the starting point for our method. However, as we later show, we found that ReFACTor captures a larger portion of the cell composition variance compared with the alternatives.

Applying ReFACTor on our input matrix X we get a list of t sites that are expected to be most informative with respect to the cell composition in X. Let $\tilde{X} \in \mathbb{R}^{n\times t}$ be a truncated version of X containing only the t sites selected by ReFACTor. We apply PCA on $\tilde{X}$ to get $L \in \mathbb{R}^{t\times d}, P \in \mathbb{R}^{n\times d}$, the loadings and scores of the first d ReFACTor components. Then, we reformulate the original optimization problem in terms of linear transformations of L and P as follows:

$$\hat{A}, \hat{V}, \hat{B} = \underset{A,V,B}{\operatorname{argmin}} \quad \frac{1}{2\sigma^2}||\tilde{X} - PVA^TL^T - CB^TL^T||_F^2 \tag{3.16}$$

$$- \sum_{h=1}^{k}(\alpha_h - 1)\sum_{i=1}^{n}\log\left(\sum_{l=1}^{d} P_{il}V_{lh}\right)$$

$$\text{s.t} \quad \forall i \forall h : \sum_{l=1}^{d} P_{il}V_{lh} \geq 0 \tag{3.17}$$

$$\forall i : \sum_{h=1}^{k}\sum_{l=1}^{d} P_{il}V_{lh} = 1 \tag{3.18}$$

$$\forall j \forall k : 0 \leq \sum_{l=1}^{d} L_{jl}A_{lh} \leq 1 \tag{3.19}$$

where $A \in \mathbb{R}^{d\times k}$ is a transformation matrix such that $\tilde{Z} = A^TL^T$ ($\tilde{Z}$ being a truncated version of Z with the t sites selected by ReFACTor), $V \in \mathbb{R}^{d\times k}$ is a transformation matrix such that $W^T = PV$ and $B \in \mathbb{R}^{d\times p}$ is a transformation matrix such that LB corresponds to the effects of each covariate on the methylation levels in each site. The constraints in (3.17) and in (3.18) correspond to the constraints in (3.13) and in (3.14), and the constraints in (3.19) correspond to the constraints in (3.15).

Given $\hat{V}$, we simply return $\hat{W} = \hat{V^T}P^T$ as the estimated cell proportions. Note that in the new formulation we are now required to learn only $d(2k+p)$ parameters - d, k and p being small constants - a dramatically decreased number of parameters compared with the original problem which requires $nk + m(k+p)$ parameters. By taking this approach, we make an assumption that $\tilde{X}$ consists of a low rank structure that captures the cell composition using d orthogonal vectors. While a natural value for d would be k, d is not bounded to be k. Particularly, in cases where substantial additional cell composition signal is expected to be captured by latter ReFACTor components (i.e. components beyond the first k), we would expect to benefit from increasing d. Clearly, overly increasing d is expected to result in overfitting and thus a decrease in performance. Finally, taking into account covariates with potentially dominant effects in the data should alleviate the risk of introducing noise into $\hat{W}$ in case of mixed low-rank structure of cell-composition signal and other unwanted variation in the data. We note, however, that similarly to the case of correlated explaining variables in regression, considering covariates that are expected to be correlated with the cell-type composition may result in underestimation of A, V and therefore lead to a decrease in the quality of $\hat{W}$.

3.2.5 Imputing cell counts using a subset of samples with measured cell counts

In practice, we observe that each of BayesCCE's components corresponds to a linear transformation of one cell type rather than to an estimate of that cell type in absolute terms. That is, it still lacks the right scaling (multiplication by a constant and addition of a constant) for transforming it into cell-type proportions. Furthermore, we would like the ith BayesCCE component to correspond to the ith cell type described by the prior using the α_i parameter.

Empirically, this is not necessarily the case, especially in scenarios where some of the α_i values are similar. In order to address these two caveats, we suggest incorporating measured cell counts for a subset of the samples in the data.

Assume we have n_0 reference samples in the data with known cell counts $W^{(0)}$ and n_1 samples with unknown cell counts $W^{(1)}$ $(n = n_0 + n_1)$. This problem can be regarded as an imputation problem, in which we aim at imputing cell counts for samples with unknown cell counts. We can find $\hat{Z}$ by solving the problem in (3.12) under the constraints in (3.15) for the n_0 reference samples while replacing W with $W^{(0)}$ and keeping it fixed. Then, given $\hat{Z}$, we can now solve the problem in (3.16), after replacing $A^T L^T$ with $\hat{Z}$ (i.e. we find only V, B now), under the following constraints

$$\forall (1 \leq i \leq n_0) \forall h : \sum_{l=1}^{d} P_{il}^{(0)} V_{lh} = W_{hi}^{(0)} \tag{3.20}$$

$$\forall (1 \leq i \leq n_1) \forall h : \sum_{l=1}^{d} P_{il}^{(1)} V_{lh} \geq 0 \tag{3.21}$$

$$\forall (1 \leq i \leq n_1) : \sum_{h=1}^{k} \sum_{l=1}^{d} P_{il}^{(1)} V_{lh} = 1 \tag{3.22}$$

where $P^{(0)}$ contains n_0 rows corresponding to the reference samples in P, and $P^{(1)}$ contains n_1 rows corresponding to the remaining samples in P. In this case, both problems of estimating Z and solving (3.16) while keeping $\hat{Z}$ fixed are convex - the first problem takes the form of a standard quadratic problem and the latter results in an optimization problem of the sum of two convex terms under linear constraints. Using $\hat{Z}$, estimated from cell counts and corresponding methylation levels of a group of samples, as well as adding the constraints in (3.20), are expected to direct the inference of W towards a set of components such that each one corresponds to one known cell type with a proper scale.

We note that given an estimate $\hat{Z}$ as described above, we can also solve directly the problem in (3.12) rather than the problem in (3.16). This approach may be more desired in cases where P does not effectively capture the cell composition variation in the data. In the context of our study, however, it is not possible to reliably evaluate the approach of solving directly the problem in (3.12), owing to the fact the the ground truth we set for evaluation is based on the same matrix Z. Specifically, in this case, the cell proportions of the reference individuals are expected to recover the same matrix Z that was used for computing the ground truth proportions of the non-reference individuals. As a result, the estimated proportions of the non-reference individuals will be exactly the ground truth that is used in the evaluation (up to a statistical error arising from the estimation of Z), regardless of the true accuracy of the estimate $\hat{Z}$ with respect to the true Z and regardless of the true accuracy of the cell proportion estimates.

3.2.6 Evaluation of performance

The fraction of cell composition variation (R^2) captured by each of the reference-free methods, ReFACTor, NNMF and MeDeCom, was computed for each cell type using a linear predictor fitted with the first k components provided by each method. In order to evaluate the performance of BayesCCE, for each component i we calculated its absolute correlation with the i-th cell type, and reported the mean absolute correlation (MAC) across the k estimated cell types. While the Dirichlet prior assigns a specific parameter α_h for each cell type h, empirically, we observed that in the case of $k = 6$ with no known cell counts for a subset of the samples, the i-th BayesCCE component did not necessarily correspond to the i-th cell type. Put differently, the labels of the k cell types had to be permuted before calculating the MAC. In this case we considered the permutation of the labels which resulted with the highest MAC as the correct permutation. In the rest of the cases, we did not apply such permutation (all the experiments using $k = 3$ and all the experiments using $k = 6$ with known cell counts for a subset of the samples).

For evaluating ReFACTor, NNMF and MeDeCom, reference-free methods which do not attribute their components to specific cell types in any scenario, we considered for each method the permutation of its components leading to the highest MAC in all experiments when compared with BayesCCE. In addition, we considered absolute error of the estimates from the ground truth as an additional quality measurement. We calculated the mean absolute error (MAE) across the k estimated cell types. When calculating absolute errors for the ReFACTor components, we scaled each ReFACTor component to be in the range $[0, 1]$.

3.2.7 Implementation and application of the reference-free and reference-based methods

We calculated the ReFACTor components for each data set using the parameters $k = 6$ and $t = 500$ and according to the default implementation and recommended guidelines of ReFACTor as described in the GLINT tool [93] and in a recent work [92], while accounting for known covariates in each data set. More specifically, in the Hannum et al. data [94] we accounted for age, sex, ethnicity and batch information, in the Liu et al. data [67] we accounted for age, sex, smoking status and batch information, and in the two Hannon et al. data sets [95] we accounted for age, sex and case/control state. We used the first six ReFACTor components ($d = 6$) for simulated data in order to accommodate with the number of simulated cell types, and the first ten components ($d = 10$) for real data, as real data are typically more complex and are therefore more likely to contain substantial signal in latter components.

The NNMF components were computed for each data set using the default setup of the RefFreeEWAS R package from the subset of 10,000 most variable sites in the data set, as performed in the NNMF paper by the authors [88]. Similarly, the MeDeCom components were computed for each data set using the default setup of the MeDeCom R package [89] from the subset of 10,000 most variable sites in the data set, as repeatedly running the method on the entire set of CpGs was revealed to be computationally prohibitive. The regularization parameter λ was selected according to a minimum cross-validation error criterion, as instructed in the MeDeCom package.

We used the GLINT tool [93] for estimating blood cell-type proportions for each one of the data sets, according to the Houseman et al. method [85], using 300 highly informative methylation sites defined in a recent study [96] and using reference data collected from sorted blood cells [74].

3.2.8 Data sets

We evaluated the performance of BayesCCE using a total of six data sets, as described bellow. For the real data experiments we downloaded four publicly available Illumina 450K DNA methylation array data sets from the Gene Expression Omnibus (GEO) database: a data set by Hannum et al. (accession GSE40279) from a study of aging rate [94], a data set by Liu et al. (accession GSE42861) from a recent association study of DNA methylation with rheumatoid arthritis [67], and two data sets by Hannon et al. (accessions GSE80417 and GSE84727; denote Hannon et al. I and Hannon et al. II) from a recent association study of DNA methylation with schizophrenia).

We preprocessed the data according to a recently suggested normalization pipeline [97]. Specifically, we retrieved and processed raw IDAT methylation files using R and the minfi R package [49] as follows. We removed 65 single nucleotide polymorphism (SNP) markers and applied the Illumina background correction to all intensity values, while separately analyzing probes coming from autosomal and non-autosomal chromosomes. We used a detection P-value threshold of P-value $< 10^{-16}$ for intensity values, setting probes with P-values higher than this threshold to be missing values. Based on these missing values, we excluded samples with call rates $< 95\%$. Since IDAT files were not made available for the Hannum et al. data set, we used the methylation intensity levels published by the authors.

As for data normalization, following the same suggested pipeline [97], we performed a quantile normalization of the methylation intensity values, subdivided by probe type, probe sub-type and color channel. Beta-normalized methylation levels were eventually calculated based on intensities levels (according to the recommendation by Illumina). On top of that, we excluded probes with over 10% missing values and used the ”impute” R package for imputing remaining missing values. Additionally, using GLINT [93], we excluded from each data set all CpGs coming from the non-autosomal chromosomes, as well as polymorphic and cross-reactive sites, as was previously suggested [37].

We further removed outlier samples and samples with missing covariates. In more details, we removed six samples from the Hannum et al. data set and two samples from the Liu et al. data set, which demonstrated extreme values in their first two principal components (over four empirical standard deviations). Furthermore, we removed from the Liu et al. data set two additional remaining samples that were regarded as outliers in the original study of Liu et al., and we removed from the Hannon et al. data sets samples with missing age information. The final number of samples remained for analysis were $n = 650, n = 658, n = 638$ and $n = 656$, and the numbers of CpGs remained were 382,158, 376,021, 381,338 and 382,158, for the Hannum et al. data set, Liu et al. data set, and the Hannon et al. I and Hannon et al. II data sets, respectively.

For learning prior information about the distribution of blood cell-type proportions we used electronic medical record (EMR) based study data that were acquired via the previously published Department of Anesthesiology and Perioperative Medicine at UCLA's perioperative data warehouse (PDW) [98]. The PDW is a structured reporting schema that contains all the relevant clinical data entered into an EMR via the use of Clarity, the relational database created by EPIC (EPIC Systems, Verona, WI) for data analytics and reporting. We used high-resolution cell count measurements from adult individuals ($n = 595$) for fitting a Dirichlet distribution. The resulted parameters of the prior were 15.0727, 1.8439, 2.5392, 1.7934, 0.7240 and 0.7404 for granulocytes, monocytes , CD4+, CD8+, B cells, and NK cells, respectively. The parameters of the prior calculated for the case of three assumed cell types ($k = 3$) were 7.7681, 0.9503, and 2.9876 for granulocytes, monocytes, and lymphocytes, respectively. Finally, for generating simulated data sets and for generating correlation maps of cell-type-specific methylomes, we used publicly available data of methylation reference of sorted cell types collected in six individuals from whole-blood tissue (GEO accession GSE35069) [74].

3.2.9 Data simulation

We simulated data following a model that was previously described in details elsewhere [36]. Briefly, we used methylation levels from sorted blood cells [74] and, assuming normality, estimated maximum likelihood parameters for each site in each cell type. cell-type-specific DNA methylation data were then generated for each simulated individual from normal distributions with the estimated parameters, conditional on the range [0,1], for six cell types and for each site. Cell proportions for each individual were generated using a Dirichlet distribution with the same parameters used in the real data analysis. Eventually, observed DNA methylation levels were composed from the cell-type-specific methylation levels and cell proportions for each individual, and a random normal noise was added to every data entry to simulate technical noise ($\sigma = 0.01$). To simulate inaccuracies of the prior, the Dirichlet parameters required by BayesCCE were learned from cell-type proportions of 50 samples generated at random from a Dirichlet distribution using the parameters learned from real data.

3.3 Results

3.3.1 Benchmarking existing reference-free methods for capturing cell-type composition

We first demonstrate that existing reference-free methods can infer components that are correlated with the tissue composition of DNA methylation data collected from heterogeneous sources. For this experiment, as well as for the rest of the experiments that follow, we used four large publicly available whole-blood methylation data sets: a data set by Hannum et al. [94] ($n = 650$), a data set by Liu et al. [67] ($n = 658$), and two data sets by Hannon et al. [95] ($n = 638$ and $n = 665$; denote Hannon et al. I and Hannon et al. II, respectively). In addition, we simulated data based on a reference data set of methylation levels from sorted leukocytes cells [74] (see section 3.2). While cell counts were known for each sample in the simulated data, cell counts were not available for the real data sets. We therefore estimated the cell-type proportions of six major blood cell types (granulocytes, monocytes and four subtypes of lymphocytes: CD4+, CD8+, B cells and natural killer cells) based on a reference-based method [85], which was shown to reasonably estimate leukocyte cell proportions from whole-blood methylation data collected from adult individuals [35, 79, 96]. Due to the absence of large publicly available data sets with measured cell counts, these estimates were considered as the ground truth for evaluating the performance of the different methods.

For benchmarking performance of existing methods, we considered three reference-free methods, all of which were shown to generate components that capture cell-type composition information from methylation: ReFACTor [36], Non-Negative Matrix Factorization (NNMF) [88] and MeDeCom [89]. Although the reference-free methods can potentially allow the detection of more cell types than the set of predefined cell types in the reference-based approach, we evaluated six components of each of the reference-free methods - six being the number of estimated cell types composing the ground truth. We found all methods to capture a large portion of the cell composition information in all data sets; particularly, we observed that ReFACTor performed considerably better than NNMF and MeDeCom in all occasions (Figure 3.1).

In spite of the fact that all three methods can capture a large portion of the cell composition variation, each component provided by these methods is a linear combination of the cell types in the data rather than an estimate of the proportions of a single cell type. As a result, as we show next, in general, these methods perform poorly when their components are considered as estimates of cell-type proportions. Of note, ReFACTor was not designed for estimating cell proportions but rather for providing orthogonal principal components of the data that together capture variation in cell compositions, however, NNMF and MeDeCom, which extends the underlying model in NNMF, were designed to provide estimates of cell-type proportions.

3.3.2 Evaluation of BayesCCE

Every method that has been developed so far for capturing cell composition signal from methylation can be classified as either reference-based, wherein a reference of methylation

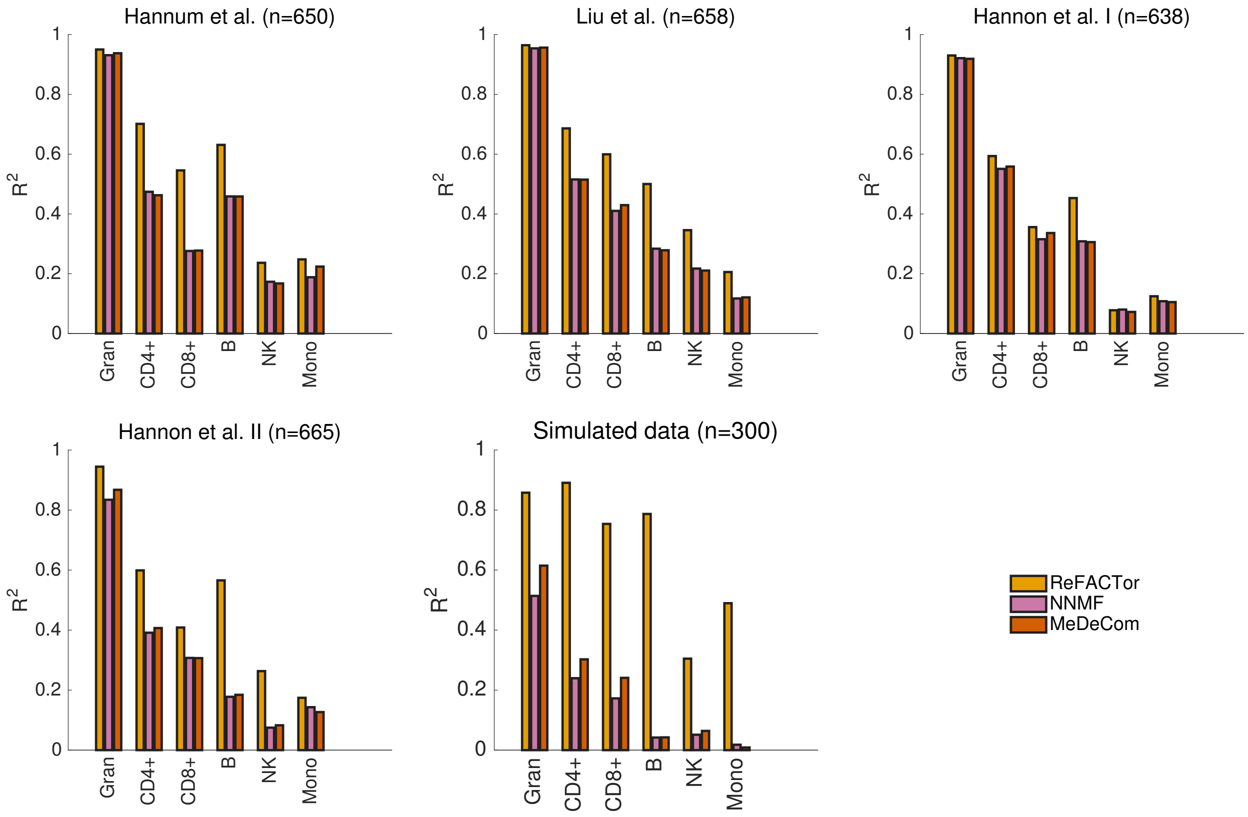

Figure 3.1: The fraction of cell-type composition variance explained (R^2) by several reference-free methods. For each of the different methods, ReFACTor, NNMF and MeDeCom, a linear model was fitted for each of the six cell types using six components. The results presented for the simulated data were averaged across ten different simulated data sets.

patterns of sorted cells is used, or reference-free, wherein cell composition information is inferred in an unsupervised manner. Our proposed method, BayesCCE, combines elements from the underlying models of previous reference-free methods with further assumptions. BayesCCE does not use standard reference data of sorted methylation levels, but rather it leverages relatively weak prior information about the distribution of cell-type composition in the studied tissue. This allows BayesCCE to direct the solution towards the inference of one component for each cell type that is encoded in the prior information.

In order to evaluate BayesCCE, we obtained prior information about the distribution of leukocyte cell-type proportions in blood using high resolution blood cell counts that were previously measured in 595 adult individuals (see section 3.2). In concordance with the estimated cell-type proportions used as the ground truth, we first considered the assumption of six constituting cell types in blood tissue ($k = 6$). We applied BayesCCE on each of the four data sets, and evaluated the resulted components. We observed that each time BayesCCE produced a set of six components such that each component was correlated with one of the cell types, as desired (Figure 3.2 and Tables 3.1 and 3.2). Specifically, we found the mean absolute correlation values across all six cell types to be 0.58, 0.63, 0.45 and 0.45 in the Hannum et al., Liu et al., Hannon et al. I and Hannon et al. II data sets, respectively. We note, however, that the assignment of components into corresponding cell types could not be automatically determined by BayesCCE. In addition, in general, the BayesCCE components were not in the right scale of their corresponding cell types (i.e. each component represented the proportions of one cell type up to a multiplicative constant and addition of a constant). These symptoms are expected due to the nature of the prior information used by BayesCCE. For more details about the assignment of components into cell types and evaluation measurements see section 3.2.

We next considered a simplifying assumption of only three constituting cell types in blood tissue ($k = 3$): granulocytes, lymphocytes and monocytes. We applied BayesCCE on each of the four data sets, and observed high correlations between the estimated components of granulocytes and the granulocytes levels ($r \geq 0.91$ in all data sets) and between the estimated components of lymphocytes and the lymphocytes levels ($r \geq 0.87$ in all data sets), yet much lower correlations for monocytes ($r \leq 0.27$ in all data sets; Figure 3.3 and Tables 3.1 and 3.2). We note that poor performance in capturing some cell type may be partially derived by inaccuracies introduced by the reference-based estimates, which are used as the ground truth in our experiments. Notably, three recent studies, which consisted of samples for which both methylation levels and cell count measurements were available, demonstrated that while the reference-based estimates of the overall lymphocyte and granulocyte levels were found to be highly correlated with the true levels, the accuracy of estimated monocytes was found to be substantially lower [36, 79, 99]. This may explain in part the low correlations we report for monocytes in our experiments. Low correlations with some of the cell types may be driven by various reasons, such as utilizing inappropriate reference or failing to perform a

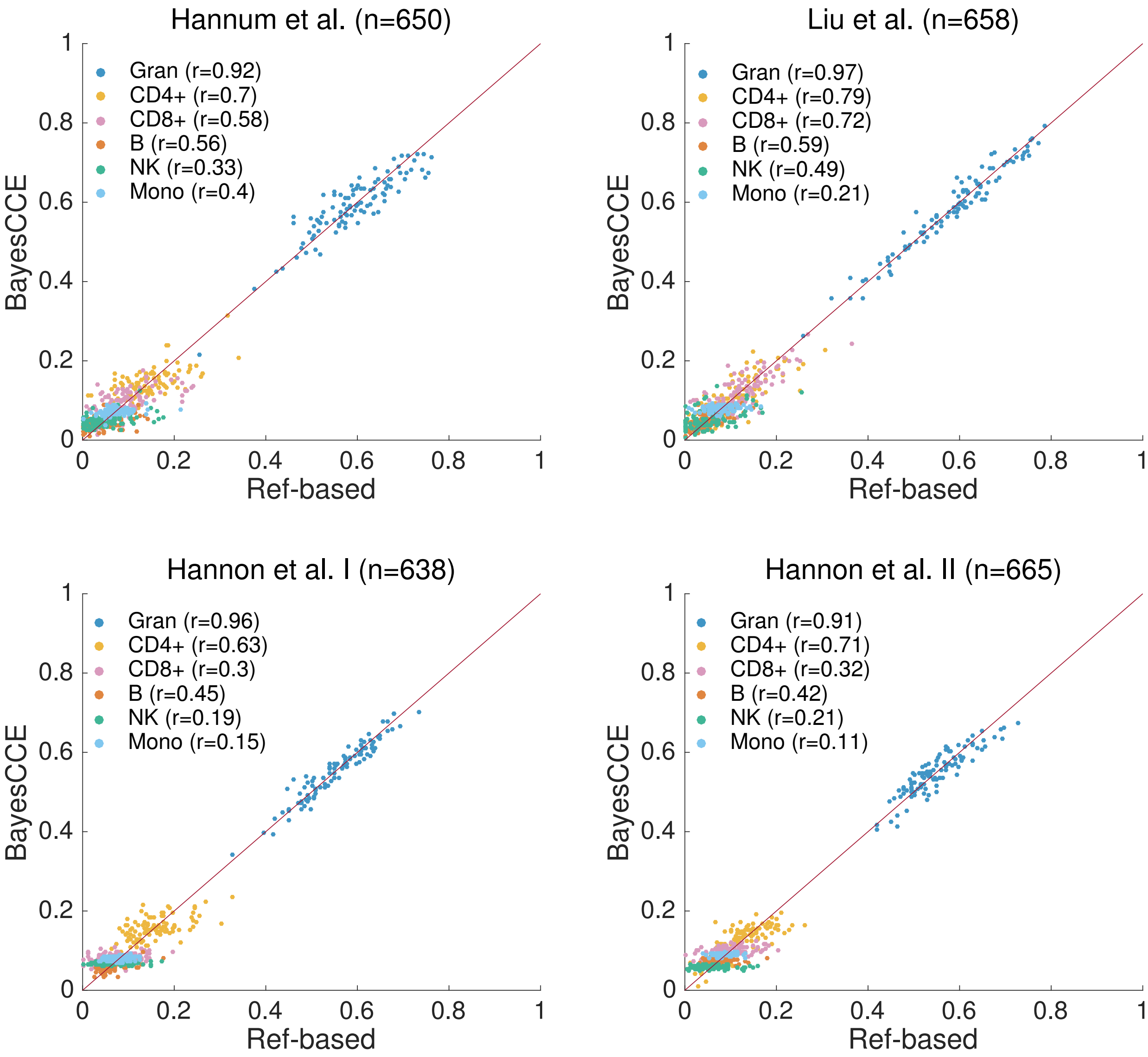

Figure 3.2: BayesCCE captures cell-type proportions in four data sets under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells). The BayesCCE estimated components were linearly transformed to match their corresponding cell types in scale (see section 3.2). For convenience of visualization, we only plot the results of 100 randomly selected samples for each data set.

good feature selection. We later provide a more detailed discussion about these issues.

For assessing the performance of BayesCCE in light of previous reference-free methods, we sub-sampled the data and generated ten data sets of 300 randomly selected samples from each one of the four data sets. In addition, we simulated ten data sets of similar size ($n =$

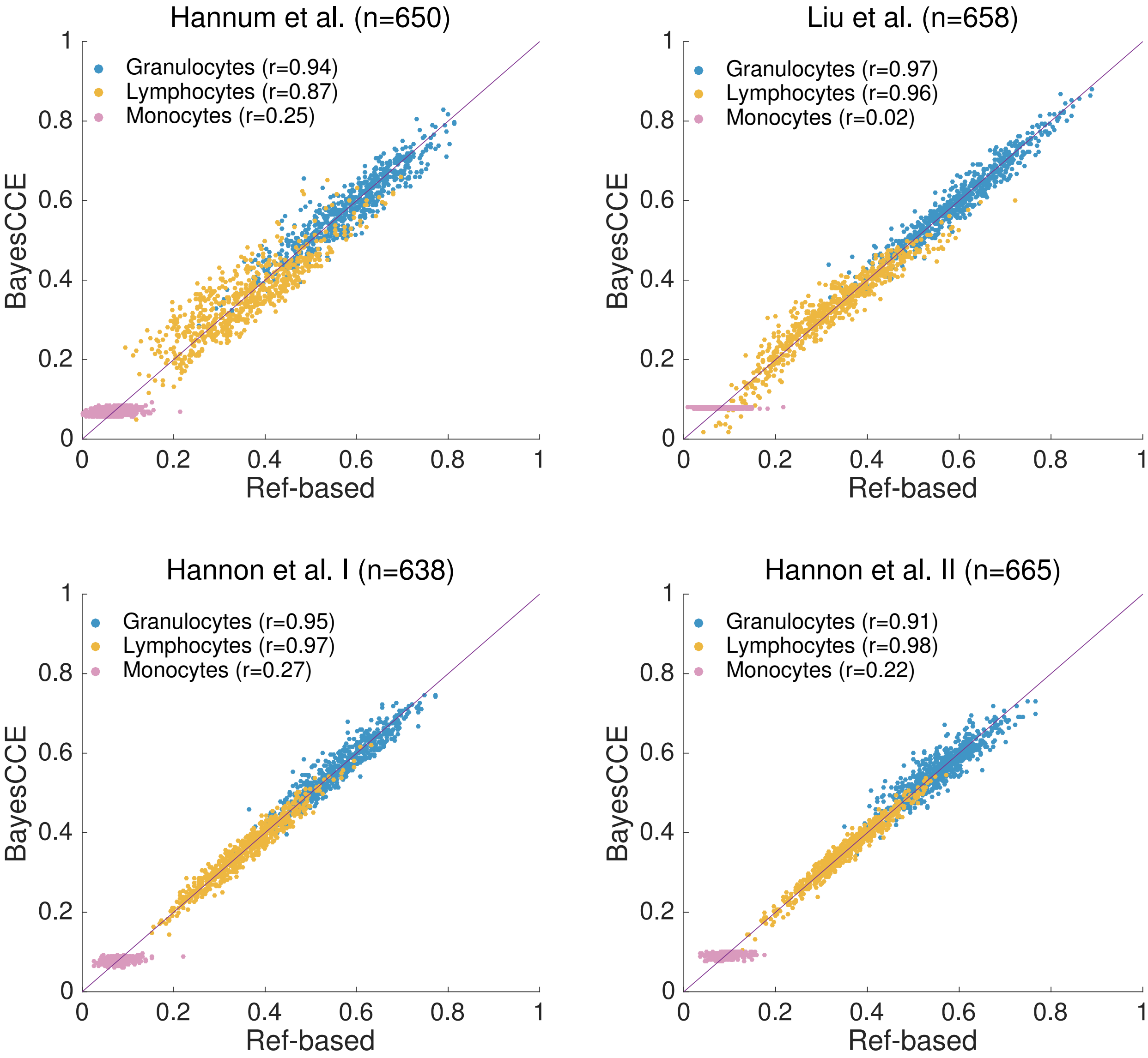

Figure 3.3: BayesCCE captures cell-type proportions in four data sets under the assumption of three constituting cell types in blood ($k = 3$): granulocytes, monocytes and lymphocytes. The BayesCCE estimated components were linearly transformed to match their corresponding cell types in scale (see section 3.2).

300; see section 3.2). Figure 3.4 demonstrates a significant and substantial improvement in performance for BayesCCE upon existing methods under the assumption of six constituting cell types ($k = 6$). Repeating the same set of experiments while assuming three constituting cell types ($k = 3$) revealed similar results (Figure 3.5).

		Absolute Correlation								
		$k = 3$			$k = 6$					
data set	Method	Gran	Lymph	Mono	Gran	CD4+	CD8+	B	NK	Mono
Hannum et al. [94]	ReFACTor	0.166	0.975	0.051	0.95	0.389	0.335	0.34	0.129	0.201
	NNMF	0.952	0.938	0.172	0.85	0.3	0.184	0.644	0.077	0.118
	MeDeCom	0.448	0.923	0.285	0.631	0.505	0.351	0.433	0.015	0.258
	BayesCCE	0.936	0.872	0.251	0.921	0.703	0.575	0.559	0.326	0.405
	BayesCCE imp	0.965	0.988	0.516	0.951	0.851	0.626	0.899	0.636	0.403
	BayesCCE imp ext	0.959	0.985	0.214	0.957	0.804	0.513	0.744	0.474	0.103
Liu et al. [67]	ReFACTor	0.164	0.982	0.105	0.961	0.089	0.495	0.338	0.137	0.309
	NNMF	0.936	0.98	0.092	0.902	0.269	0.588	0.023	0.089	0.328
	MeDeCom	0.97	0.773	0.054	0.73	0.563	0.24	0.16	0.283	0.293
	BayesCCE	0.971	0.956	0.021	0.973	0.785	0.719	0.59	0.487	0.209
	BayesCCE imp	0.977	0.986	0.561	0.982	0.792	0.675	0.609	0.554	0.496
	BayesCCE imp ext	0.988	0.986	0.529	0.971	0.726	0.66	0.646	0.516	0.483
Hannon et al. I [95]	ReFACTor	0.387	0.919	0.025	0.883	0.013	0.403	0.358	0.043	0.147
	NNMF	0.916	0.959	0.157	0.682	0.597	0.401	0.159	0.074	0.193
	MeDeCom	0.934	0.7	0.027	0.801	0.342	0.285	0.297	0.16	0.135
	BayesCCE	0.947	0.973	0.266	0.956	0.628	0.297	0.451	0.186	0.153
	BayesCCE imp	0.938	0.977	0.305	0.944	0.738	0.467	0.643	0.366	0.35
	BayesCCE imp ext	0.971	0.973	0.528	0.967	0.665	0.355	0.687	0.384	0.419
Hannon et al. II [95]	ReFACTor	0.106	0.977	0.072	0.952	0.011	0.214	0.427	0.429	0.05
	NNMF	0.833	0.805	0.14	0.598	0.416	0.245	0.234	0.038	0.143
	MeDeCom	0.829	0.724	0.018	0.482	0.329	0.222	0.107	0.124	0.102
	BayesCCE	0.91	0.981	0.217	0.914	0.713	0.316	0.425	0.206	0.107
	BayesCCE imp	0.973	0.983	0.441	0.965	0.756	0.62	0.823	0.641	0.519
	BayesCCE imp ext	0.957	0.98	0.299	0.972	0.751	0.563	0.775	0.618	0.604

Table 3.1: A summary of the correlation of existing reference-free methods and BayesCCE with each cell type in four whole-blood data sets (considering reference-based estimates as the ground truth), under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells), and under the assumption of three constituting cell types in blood ($k = 3$): granulocytes, monocytes and lymphocytes. For each of the methods, ReFACTor, NNMF, MeDeCom and BayesCCE, we considered a single component per cell type (see section 3.2). In addition, we considered the scenario wherein cell counts are known for 5% of the samples (BayesCCE imp), and the scenario wherein samples from external data with both methylation levels and cell counts are available (5% of the sample size; BayesCCE imp ext). For BayesCCE imp and BayesCCE imp ext, correlations were calculated after excluding the samples with assumed known cell counts.

		Mean Absolute Error								
		$k = 3$			$k = 6$					
data set	Method	Gran	Lymph	Mono	Gran	CD4+	CD8+	B	NK	Mono
Hannum et al. [94]	ReFACTor	0.187	0.104	0.587	0.233	0.498	0.335	0.627	0.593	0.161
	NNMF	0.113	0.11	0.067	0.141	0.121	0.062	0.272	0.051	0.046
	MeDeCom	0.232	0.064	0.276	0.445	0.072	0.125	0.148	0.134	0.106
	BayesCCE	0.237	0.186	0.114	0.501	0.097	0.166	0.041	0.043	0.422
	BayesCCE imp	0.022	0.022	0.021	0.022	0.027	0.029	0.015	0.023	0.021
	BayesCCE imp ext	0.044	0.018	0.046	0.032	0.042	0.032	0.031	0.037	0.027
Liu et al. [67]	ReFACTor	0.183	0.14	0.54	0.233	0.356	0.497	0.41	0.423	0.317
	NNMF	0.197	0.196	0.046	0.223	0.082	0.276	0.042	0.049	0.058
	MeDeCom	0.284	0.193	0.202	0.398	0.071	0.079	0.1	0.165	0.108
	BayesCCE	0.23	0.214	0.043	0.094	0.034	0.038	0.049	0.076	0.038
	BayesCCE imp	0.023	0.015	0.017	0.02	0.033	0.034	0.016	0.027	0.018
	BayesCCE imp ext	0.013	0.016	0.021	0.019	0.032	0.045	0.021	0.03	0.019
Hannon et al. I [95]	ReFACTor	0.222	0.131	0.383	0.201	0.318	0.284	0.445	0.404	0.437
	NNMF	0.218	0.221	0.045	0.463	0.221	0.305	0.05	0.046	0.043
	MeDeCom	0.215	0.151	0.246	0.408	0.062	0.083	0.117	0.122	0.115
	BayesCCE	0.27	0.23	0.084	0.311	0.159	0.053	0.054	0.066	0.042
	BayesCCE imp	0.022	0.014	0.023	0.034	0.027	0.028	0.014	0.026	0.016
	BayesCCE imp ext	0.014	0.03	0.017	0.017	0.03	0.03	0.027	0.026	0.016
Hannon et al. II [95]	ReFACTor	0.231	0.199	0.368	0.185	0.363	0.272	0.39	0.223	0.28
	NNMF	0.468	0.47	0.048	0.502	0.086	0.624	0.039	0.048	0.061
	MeDeCom	0.207	0.08	0.277	0.413	0.082	0.097	0.131	0.123	0.125
	BayesCCE	0.205	0.191	0.064	0.31	0.192	0.034	0.07	0.07	0.038
	BayesCCE imp	0.013	0.012	0.015	0.027	0.025	0.025	0.011	0.023	0.015
	BayesCCE imp ext	0.017	0.015	0.035	0.014	0.026	0.027	0.015	0.022	0.016

Table 3.2: A summary of the mean absolute error of existing reference-free methods and BayesCCE with each cell type in four whole-blood data sets (considering reference-based estimates as the ground truth), under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells), and under the assumption of three constituting cell types in blood ($k = 3$): granulocytes, monocytes and lymphocytes. For each of the methods, ReFACTor, NNMF, MeDeCom and BayesCCE, we considered a single component per cell type (see section 3.2). In addition, we considered the scenario wherein cell counts are known for 5% of the samples (BayesCCE imp), and the scenario wherein samples from external data with both methylation levels and cell counts are available (5% of the sample size; BayesCCE imp ext). For BayesCCE imp and BayesCCE imp ext, absolute errors were calculated after excluding the samples with assumed known cell counts.

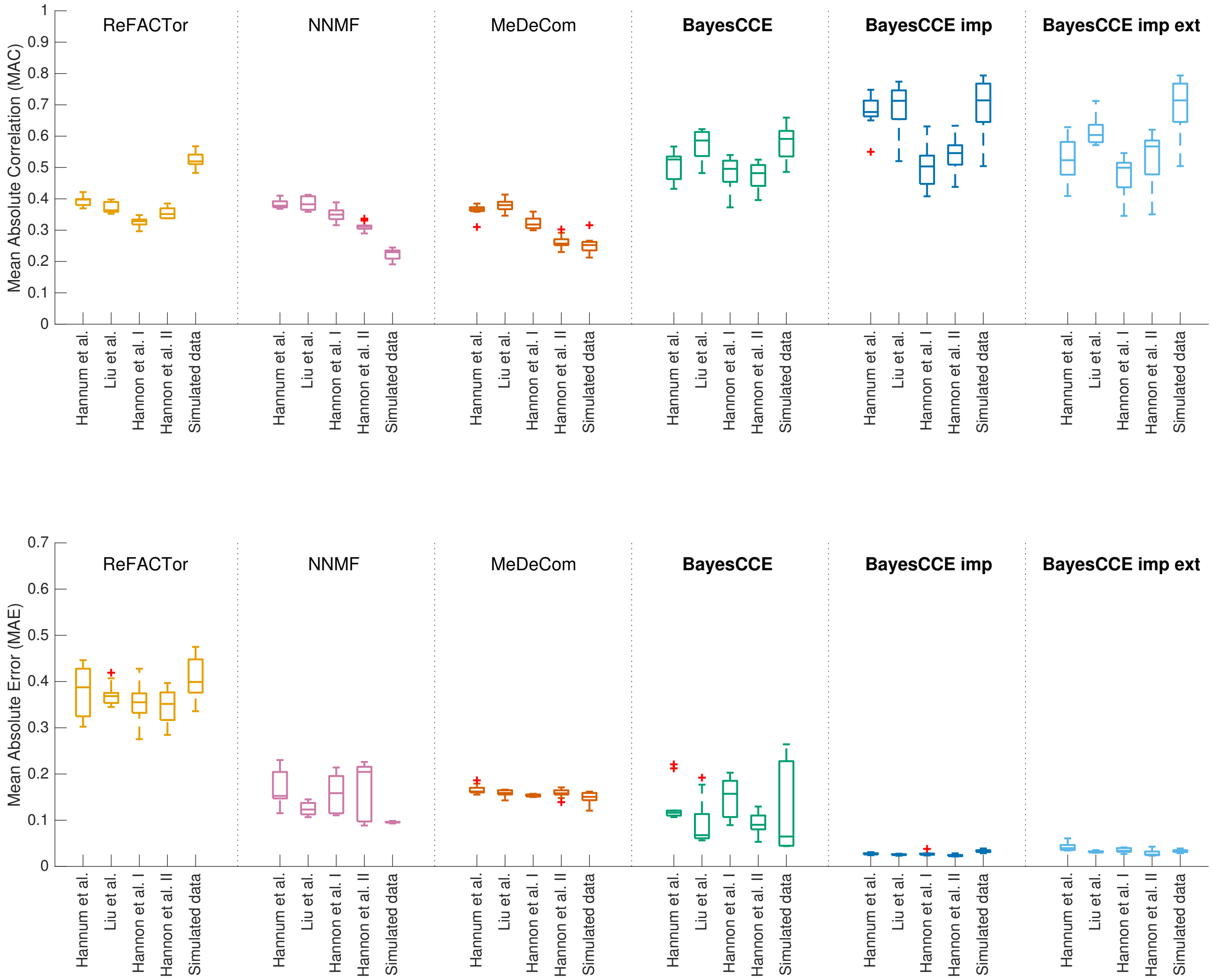

Figure 3.4: The performance of existing reference-free methods and BayesCCE under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells). For each method, box plots show for each data set the performance across ten sub-sampled data sets ($n = 300$), with the median indicated by a horizontal line. For each of the methods, ReFACTor, NNMF, MeDeCom and BayesCCE, we considered a single component per cell type (see section 3.2). Additionally, we considered the scenario of cell counts imputation wherein cell counts were known for 5% of the samples ($n = 15$; BayesCCE imp), and the scenario wherein samples from external data with both methylation levels and cell counts were used in the analysis ($n = 15$; BayesCCE imp ext). Top panel: mean absolute correlation (MAC) across all cell types. Bottom panel: mean absolute error (MAE) across all cell types. For BayesCCE imp and BayesCCE imp ext, the MAC and MAE values were calculated while excluding the samples with assumed known cell counts.

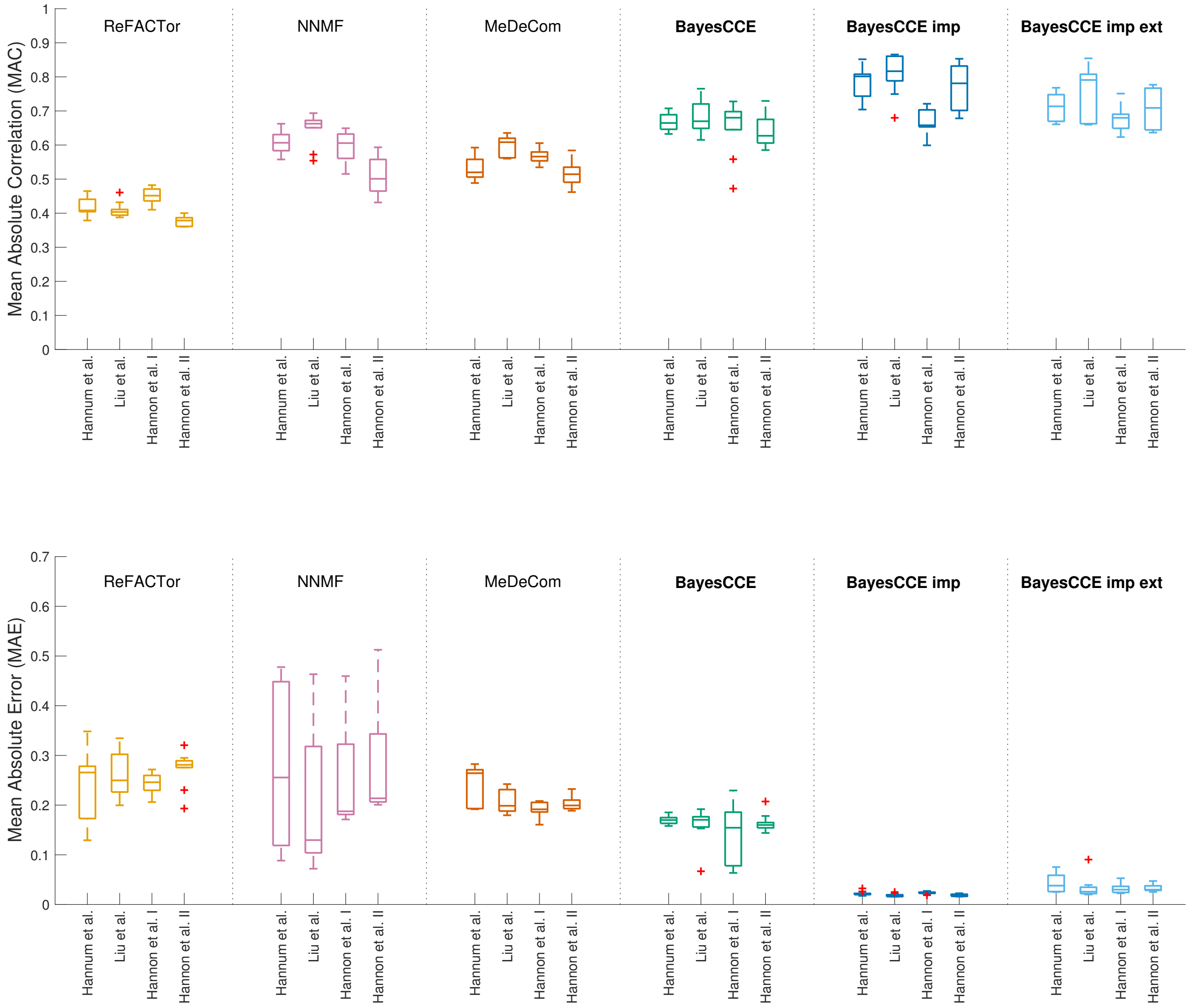

Figure 3.5: The performance of existing reference-free methods and BayesCCE under the assumption of three constituting cell types in blood ($k = 3$): granulocytes, monocytes and lymphocytes. For each method, box plots show for each data set the performance across ten sub-sampled data sets ($n = 300$), with the median indicated by a horizontal line. For each of the methods, ReFACTor, NNMF, MeDeCom and BayesCCE, we considered a single component per cell type (see section 3.2). Additionally, we considered the scenario of cell counts imputation wherein cell counts were known for 5% of the samples ($n = 15$; BayesCCE imp), and the scenario wherein samples from external data with both methylation levels and cell counts were used in the analysis ($n = 15$; BayesCCE imp ext). Top panel: mean absolute correlation (MAC) across all cell types. Bottom panel: mean absolute error (MAE) across all cell types. For BayesCCE imp and BayesCCE imp ext, the MAC and MAE values were calculated while excluding the samples with assumed known cell counts.

3.3.3 BayesCCE impute: cell counts imputation

We next considered a scenario in which cell counts are known for a small subset of the samples in the data. This problem can be viewed as a problem of imputing missing cell count values (see section 3.2). We repeated all previous experiments, only this time we assumed that cell counts are known for randomly selected 5% of the samples in each data set. As opposed to the previous experiments, in which each one of the BayesCCE components constituted a scaled estimate of the proportions of one of the cell types, incorporating samples with known cell counts allowed BayesCCE to produce components that form absolute estimates of the cell-type proportions (i.e. not scaled components, but components with low absolute error compared with the true proportions). Moreover, in contrast to previous experiments, each component was now automatically assigned to its corresponding cell type.

Under the assumption of six constituting cell types in blood tissue ($k = 6$), we observed a substantial improvement of up to 58% in mean absolute correlation values compared with our previous experiments (Figure 3.6 and Tables 3.1 and 3.2). Specifically, we found the mean absolute correlation values across all six cell types to be 0.71, 0.66, 0.56 and 0.71 in the Hannum et al., Liu et al., Hannon et al. I and Hannon et al. II data sets, respectively. In addition, in contrast to our previous experiments, inclusion of some cell counts resulted in low mean absolute error, which reflects a correct scale for the components. We observed similar results when assuming three constituting cell types ($k = 3$), providing an improvement of up to 28% in correlation and a substantial decrease in absolute errors compared with the previous experiments (Figure 3.7 and Tables 3.1 and 3.2).

In the absence of cell counts for a subset of the individuals in the data, we can incorporate into the analysis external data of samples for which both cell counts and methylation levels (from the same tissue) are available. We repeated again all previous experiments ($k = 3$ and $k = 6$), only this time for each data set we added a randomly selected subset of samples from one of the other data sets (5% of the original sample size), and used both their methylation levels and cell-type proportions in the analysis. Specifically, we used randomly selected samples and corresponding estimates of cell-type proportions from the Hannon et al. I data

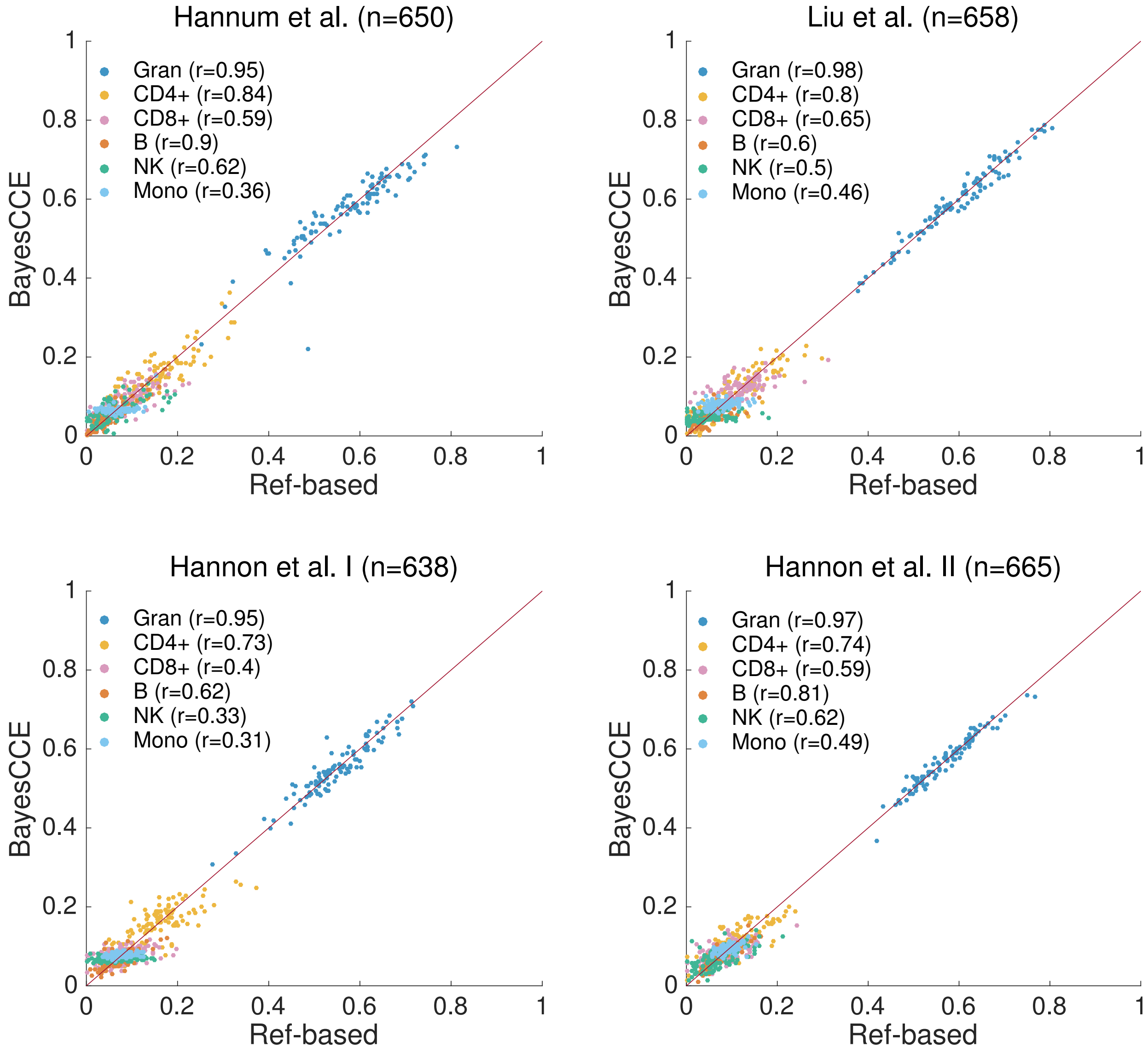

Figure 3.6: BayesCCE captures cell-type proportions in four data sets under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells), and assuming known cell counts for randomly selected 5% of the samples in the data. All correlations were calculated while excluding the samples with assumed known cell counts. For convenience of visualization, we only plot the results of 100 randomly selected samples for each data set.

set for the experiments in all three other data sets, and samples from the Hannon et al. II data set for the experiment with the Hannon et al. I data set. In order to pool samples from two data sets together, we considered only the intersection of CpG sites that were available for analysis in the two data sets. In addition, unlike in the previous experiments,

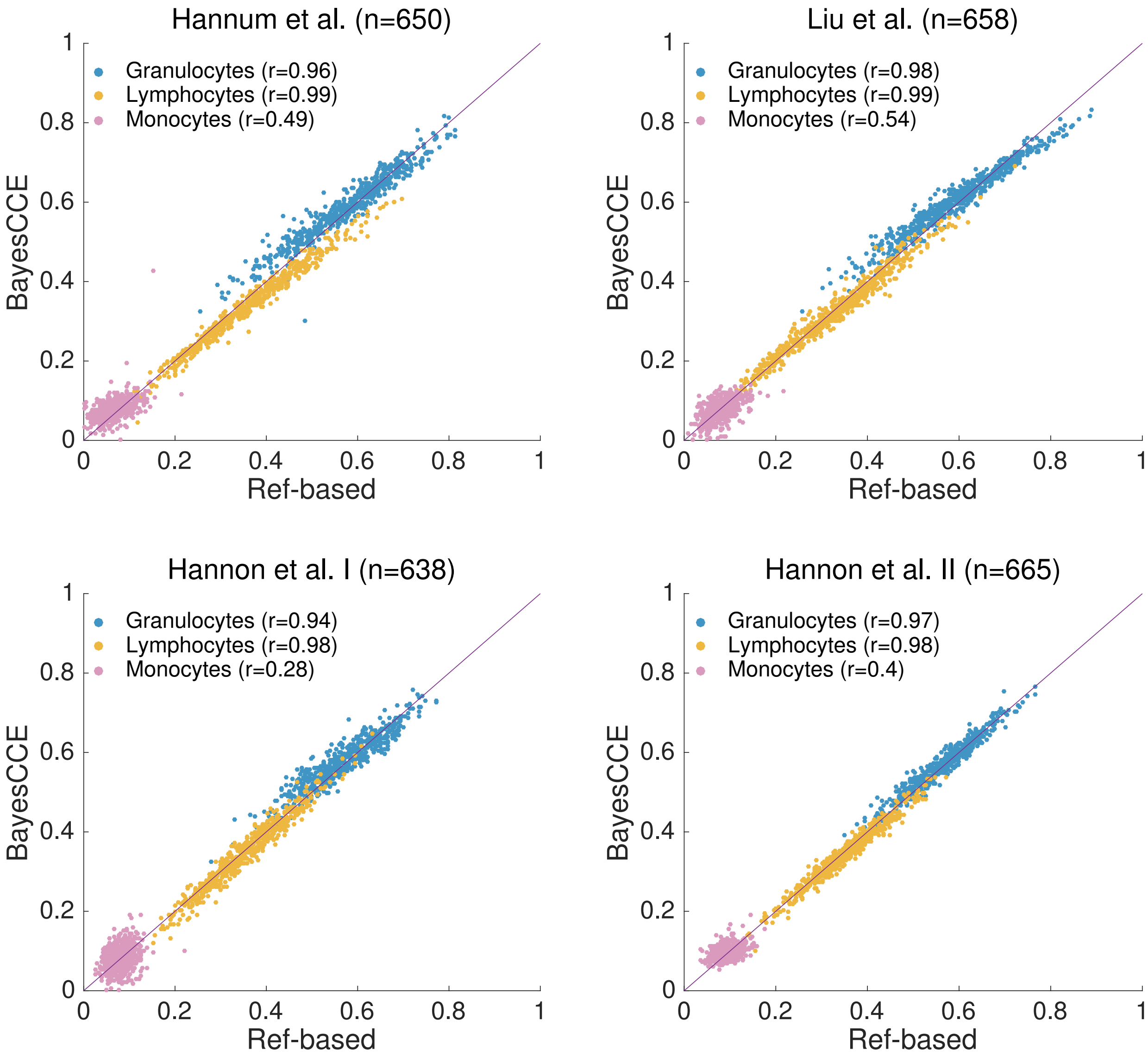

Figure 3.7: BayesCCE captures cell-type proportions in four data sets under the assumption of three constituting cell types in blood ($k = 3$): granulocytes, monocytes and lymphocytes, and assuming known cell counts for randomly selected 5% of the samples in the data. All correlations were calculated while excluding the samples with assumed known cell counts.

here we potentially introduce new batch effects into the analysis, as in each experiment the original sample is combined with external data. We therefore accounted for the new batch information by adding it as a new covariate into BayesCCE. As in the case of known cell counts for a subset of the samples, we found that the inclusion of external samples with both methylation and cell counts substantially improved the performance in terms of correlation

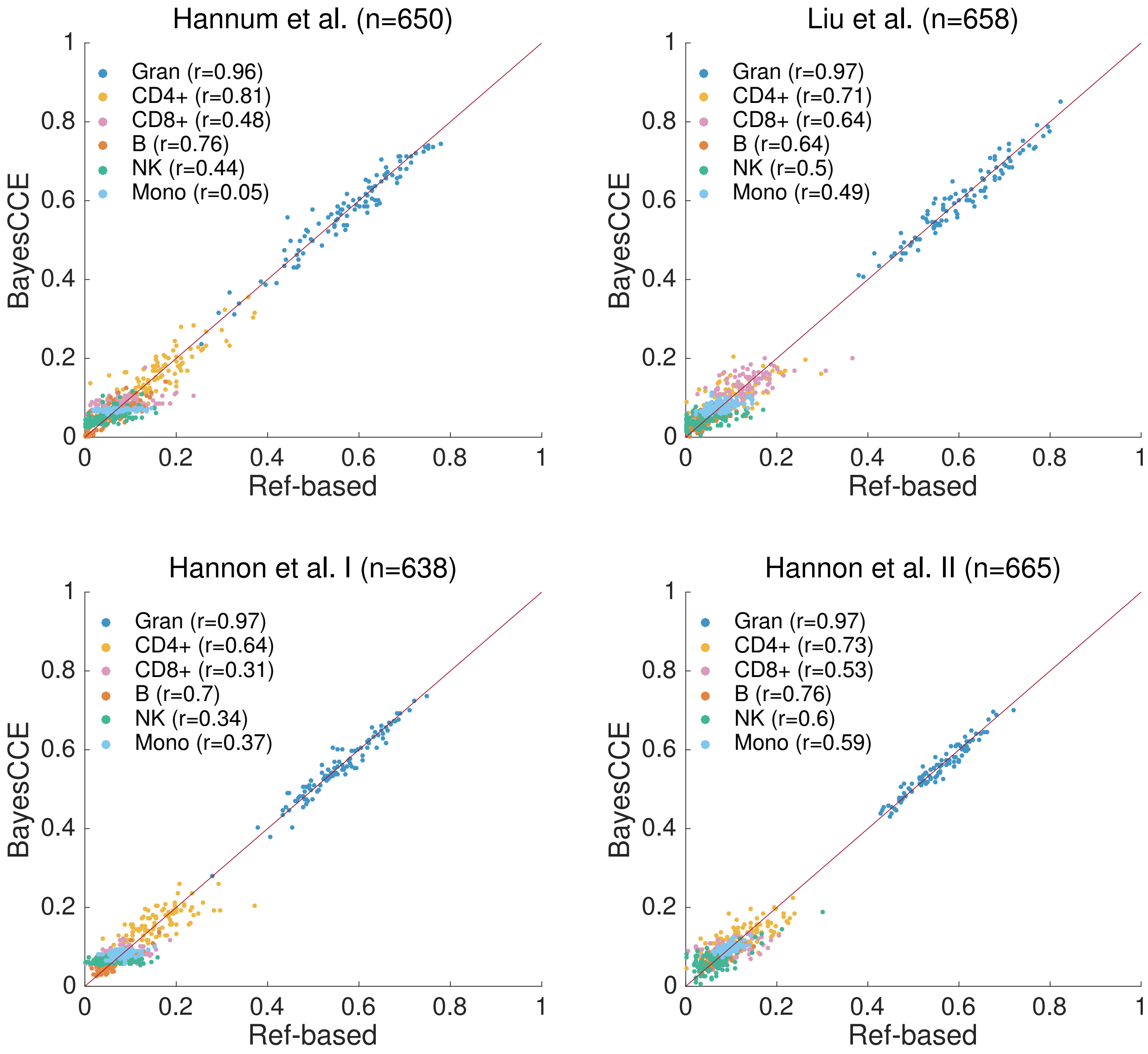

Figure 3.8: BayesCCE captures cell-type proportions in four data sets under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells), and including a group of samples with known cell counts from external data. For each data set, samples from one of the other data sets were included in the analysis (5% of the sample size), while assuming that both their methylation levels and cell counts are known. All correlations were calculated while excluding the samples with assumed known cell counts. For convenience of visualization, we only plot the results of 100 randomly selected samples for each data set.

and absolute errors (Figures 3.8 and 3.9 and Tables 3.1 and 3.2). These results clearly show that estimates can be dramatically more accurate given measured cell counts for as few as a couple of dozens of samples in the data (or such samples from external data).

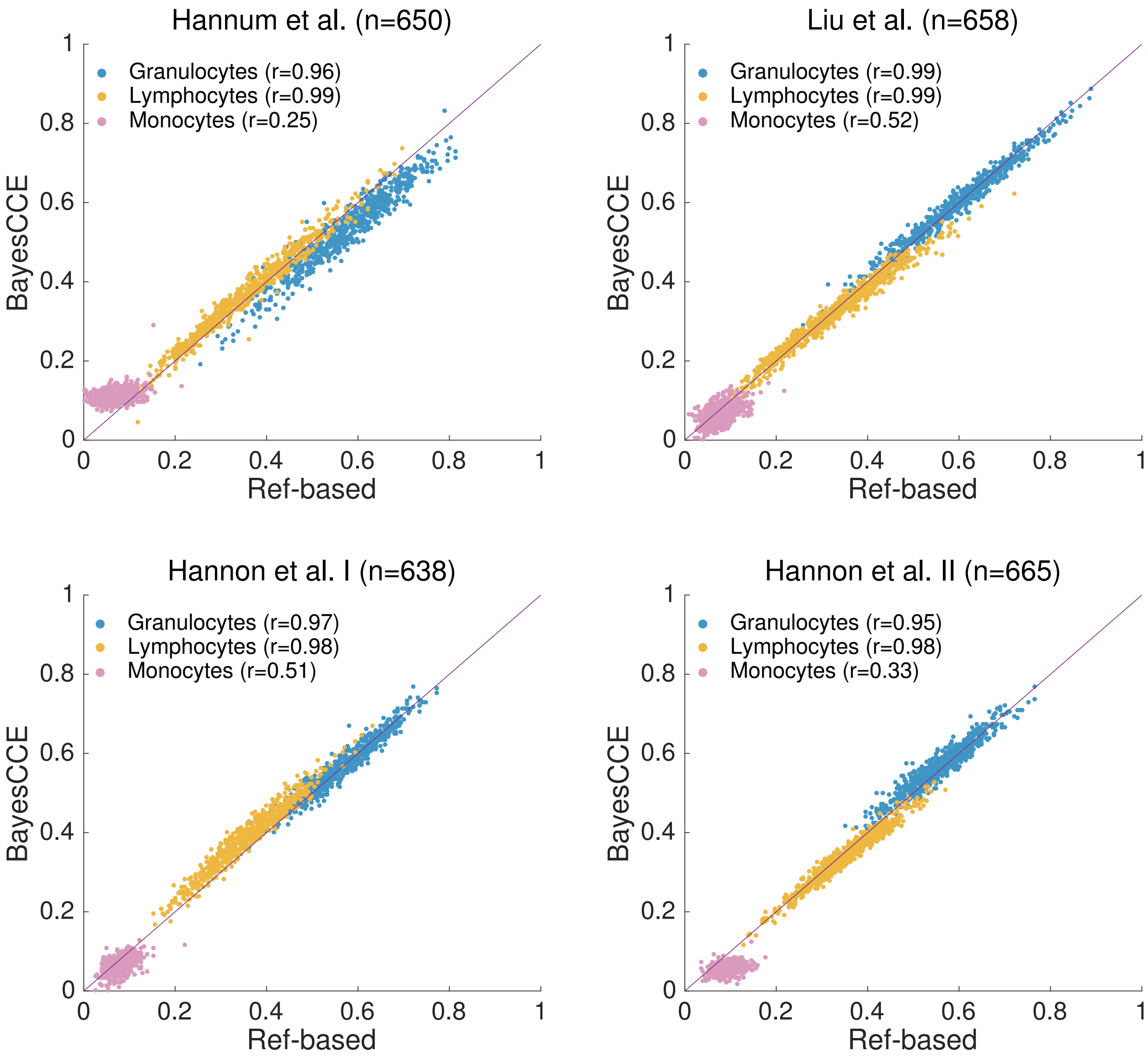

Figure 3.9: BayesCCE captures cell-type proportions in four data sets under the assumption of three constituting cell types in blood ($k = 3$): granulocytes, monocytes and lymphocytes, and including a group of samples with known cell counts from external data. For each data set, samples from one of the other data sets were included in the analysis (5% of the sample size), while assuming that both their methylation levels and cell counts are known. All correlations were calculated while excluding the samples with assumed known cell counts.

As before, for assessing performance more thoroughly, we applied BayesCCE on the same subsampled data sets we used before ($n = 300$), while assuming known cell counts for a subset of the samples. In one scenario we assumed cell counts are known for 5% of the samples in each data set ($n = 15$), and in a second scenario we included into the analysis methylation levels and cell-type proportions of 15 samples from external data. These experiments revealed in most cases a substantial improvement in correlation over a standard execution of BayesCCE (i.e. without inclusion of cell counts), and revealed in all cases a substantial improvement in mean absolute error. The results are summarized in Figure 3.4 for the case of six constituting cell types ($k = 6$) and in Figure 3.5 for the case of three constituting cell types ($k = 3$).

We further tested the performance of BayesCCE as a function of the number of samples for which cell counts are available. Remarkably, we found that known cell counts for only a couple of dozens of the samples are needed in order to achieve the maximal improvement in performance; including more samples with known cell counts did not provide a further improvement (Figure 3.10). In addition, we evaluated the performance of BayesCCE as a function of the sample size. Interestingly, while performance did not improve by increasing the sample over a few hundred of samples in the case of unknown cell counts, we found that knowledge of cell counts for as few as 15 samples in the data allowed a monotonic improvement in performance in larger sample sizes (Figure 3.11).

Finally, we considered an alternative approach for verifying the results of BayesCCE. Although our study aims at estimating cell-type proportions without the need for reference methylation data, BayesCCE jointly learns cell-type composition and cell-type-specific mean methylation levels (methylomes). Hence, as a by-product of the BayesCCE algorithm, we also obtain cell-type-specific methylomes across the CpG sites selected by BayesCCE as part of its feature selection process (see section 3.2). Our experiments found BayesCCE to provide one component per cell type; however, these components are not necessarily appropriately scaled, which implies that estimated cell-type-specific methylation profiles are also not necessarily calibrated. Nevertheless, in the scenario where cell counts were known even for a small subset of the individuals in the study, BayesCCE provided calibrated cell count estimates. In such cases, we therefore expect BayesCCE to provide calibrated cell-type-specific methylation profiles. Using correlation maps, for each of the four whole-blood methylation data sets we analyzed, we verified high similarity between the cell-type-specific methylomes obtained by BayesCCE to those estimated by a reference methylation data collected from

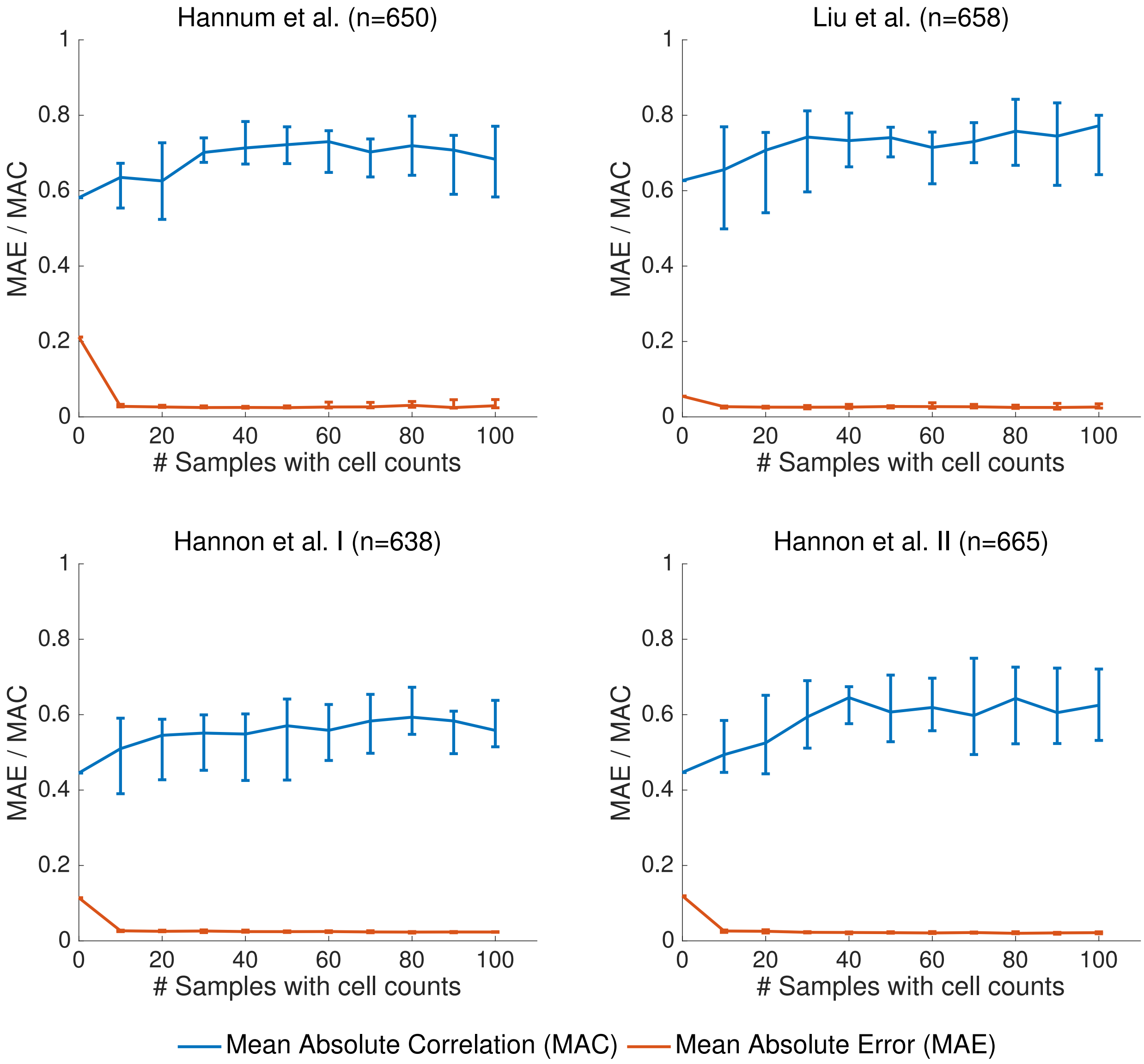

Figure 3.10: Performance of BayesCCE as a function of the number of samples for which cell counts are known, under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells). Presented are the medians of the mean absolute correlation values (MAC; in blue) and the medians of the mean absolute error values (MAE; in red) across the six cell types. Error bars indicate the range of MAC and MAE values across ten different executions for each number of samples with known cell counts. In every execution samples with known cell counts were randomly selected, and all MAC and MAE values were calculated while excluding the samples with assumed known cell counts.

sorted blood cells [74] (Figures 3.12 and 3.13).

In spite of an overall high similarity between these two approaches, the correlation patterns

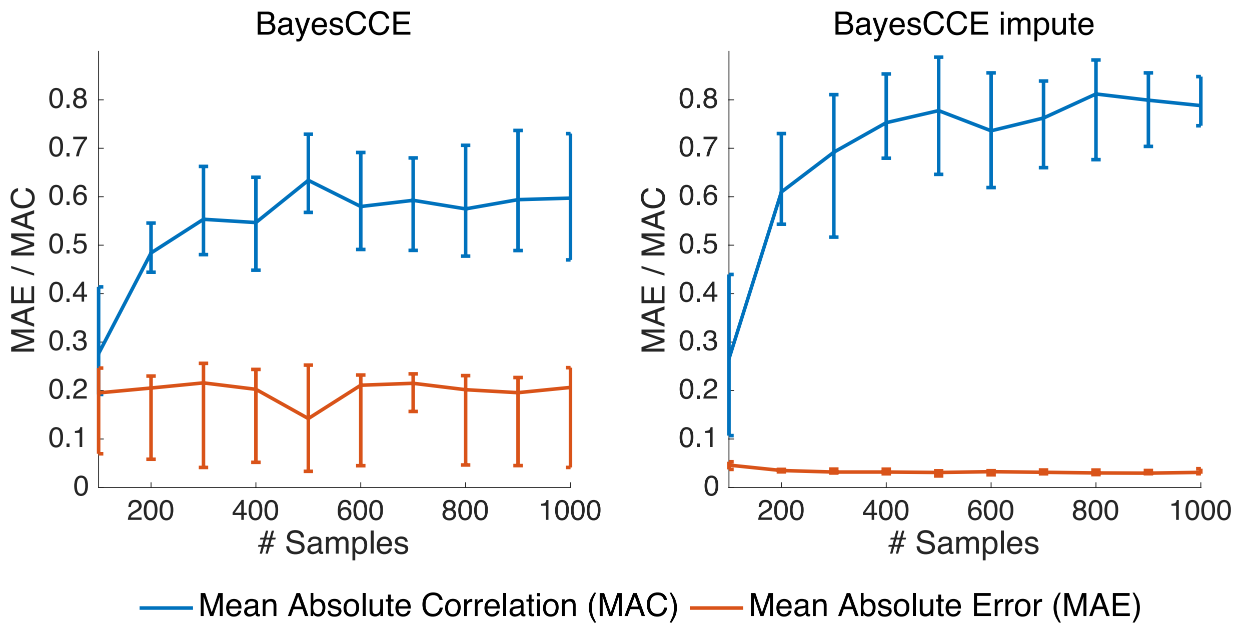

Figure 3.11: Performance of BayesCCE without known cell counts and BayesCCE with known cell counts (BayesCCE imp) for 15 of the samples as a function of the number of samples in simulated data ($k = 6$). Presented are the medians of the mean absolute correlation values (MAC; in blue) and the medians of the mean absolute error values (MAE; in red) across the six cell types. Error bars indicate the range of MAC and MAE values across ten different executions for each sample size. In BayesCCE imp, all MAC and MAE values were calculated while excluding the samples with assumed known cell counts.

detected by BayesCCE did not perfectly match those estimated using the reference data. While this may demonstrate the expected accuracy limitations of BayesCCE to some extent, we also attribute these imperfect matches, at least in part, to inaccuracies introduced by the reference data set, owing to the fact that it was constructed only from a small group of individuals ($n = 6$), which do not represent well all the individuals in other data sets in terms of methylome altering factors such as age [39], gender [40, 77], and genetics [70].

3.3.4 Robustness of BayesCCE to biases introduced by the cell composition prior

BayesCCE relies on prior information about the distribution of the cell-type composition in the studied tissue. In practice, the available prior information may not always precisely reflect the cell composition distribution of the individuals in the study. For instance, in a

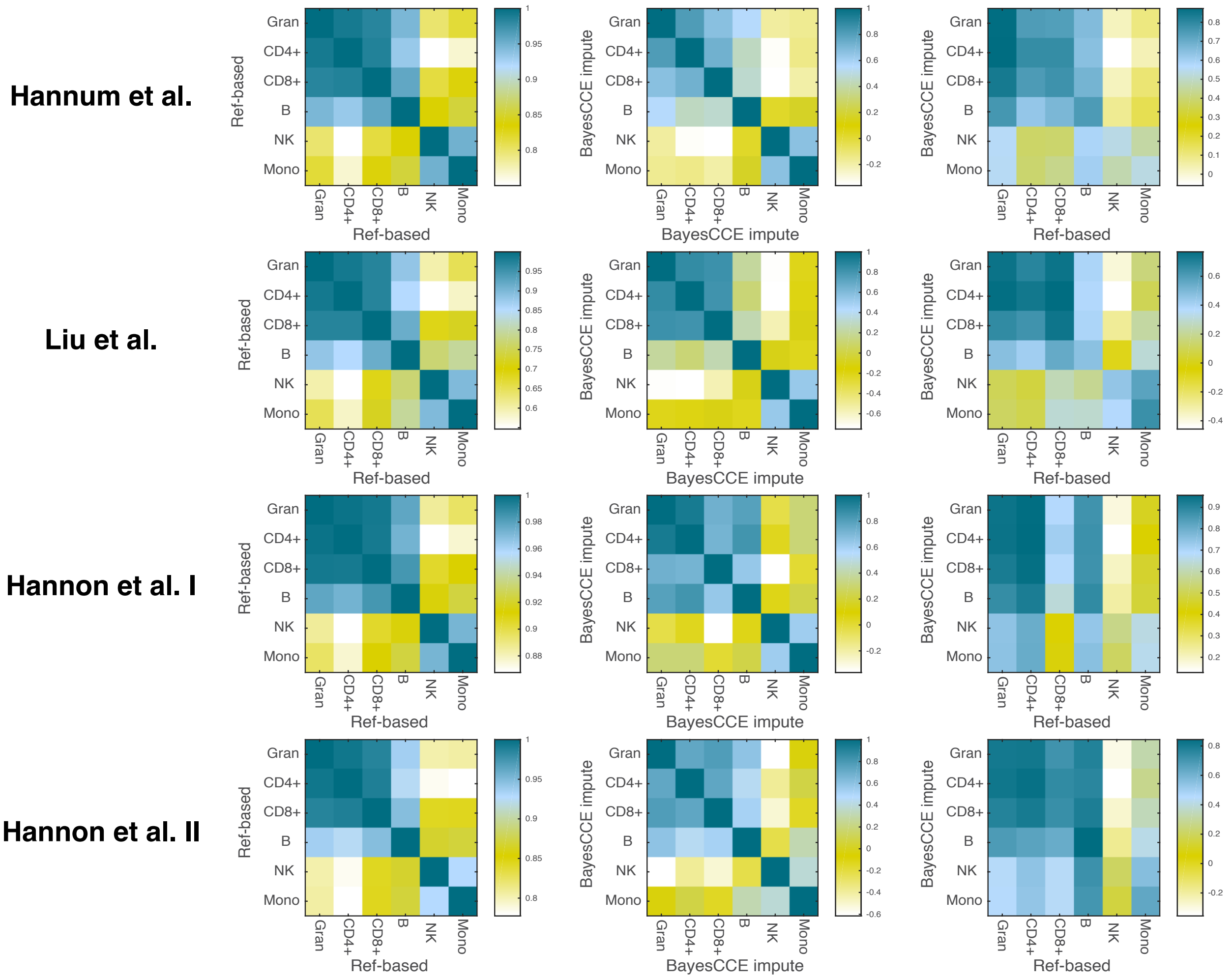

Figure 3.12: Correlation maps of the estimated cell-type-specific methylomes using BayesCCE impute, under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells). For each of four data sets, correlation maps were calculated using cell-type-specific mean methylation levels estimated from a reference data set of methylation levels collected from sorted blood cell types by Reinius et al. (left column), using the estimates obtained by BayesCCE under the assumption of known cell counts for 5% of the samples (BayesCCE impute; middle column), and using the reference-based estimates versus the BayesCCE impute estimates (right column).

case/control study design, cases may demonstrate altered cell compositions compared with healthy individuals. Therefore, in this scenario, a prior estimated from a healthy population (or a sick population) is expected to deviate from the actual distribution in the sample. This potential problem is clearly not limited to case/control studies, but also applies to studies with quantitative phenotypes, in case these are correlated with changes in cell composition of

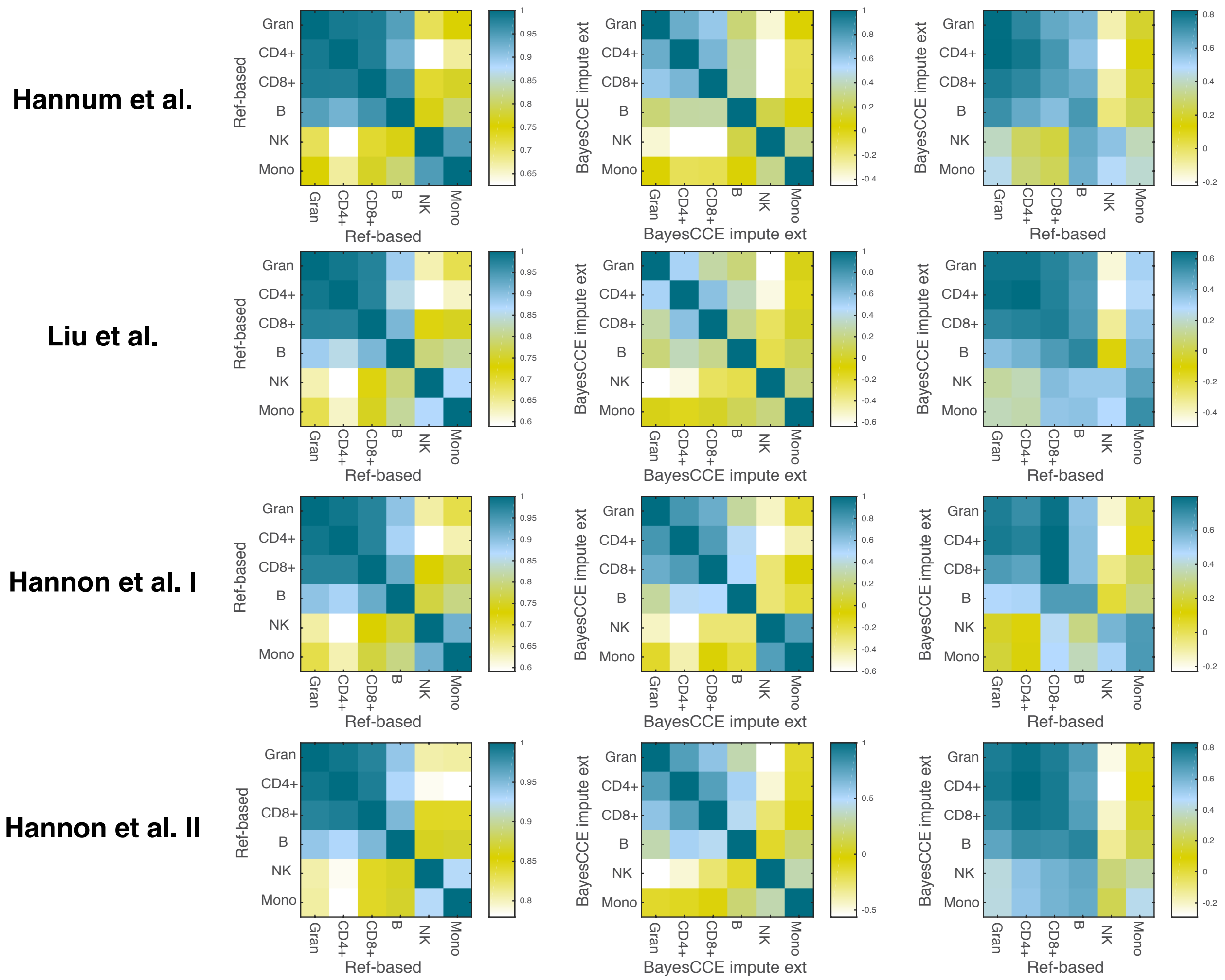

Figure 3.13: Correlation maps of the estimated cell-type-specific methylomes using BayesCCE impute with external data, under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells). For each of four data sets, correlation maps were calculated using cell-type-specific mean methylation levels estimated from a reference data set of methylation levels collected from sorted blood cell types by Reinius et al. (left column), using the estimates obtained by BayesCCE in a scenario wherein samples from external data with both methylation levels and cell counts were available (5% of the sample size; BayesCCE impute ext, middle column), and using the reference-based estimates versus the BayesCCE impute ext estimates (right column).

the studied tissue. In principle, we can address this issue by incorporating several appropriate priors and assigning different priors to different individuals in the study. However, in practice, population-specific priors may be hard to obtain, mainly owing to the fact that numerous known and unknown factors can affect cell composition.

We revisited our analysis from the previous subsections in attempt to assess the robustness of BayesCCE to non-informative or misspecified priors. A desired behavior would allow BayesCCE to overcome a bias introduced by a prior which does not accurately represent all the individuals in the sample. Particularly, we considered three whole-blood case/control data sets, two schizophrenia data sets by Hannon et al. and a rheumatoid arthritis data set by Liu et al., all of which are expected to demonstrate differences in blood cell composition between cases and controls [100, 101].

In fact, in our analysis we had an inherently misspecified prior since we learned the prior from hospital patients (outpatients), which are overall expected to represent a sick population better than a more general population. Specifically, out of the 595 individuals used for learning the prior, 64% are known to have taken at least one medication at the time of blood draw for cell counting and 24% were admitted to the hospital due to various conditions within two months before or after the time of their blood draw (70.4% were either admitted or took medications). We expect these conditions to be correlated with alterations in blood cell composition, and therefore the prior information we used is expected to represent deviation from a healthy population and, as a result, to misrepresent at least the control individuals in the case/control data sets we analyzed.

We further considered an additional fourth data set by Hannum et al., which was originally studied in the context of aging (age range: 19-101, mean: 64.03, SD: 14.73). Our prior was calculated using sample with a different distribution of ages (range: 20-88, mean: 49.19, SD: 16.69), thus misrepresenting the cell composition distribution in the Hannum et al. data to some extent.

Remarkably, we found the cell composition estimates given by BayesCCE to effectively detect differences between populations in the data sets, in spite of using a single prior estimated from one particular population. Specifically, we found that BayesCCE correctly detected the cell types which differentiate between cases and controls and between young and older populations; notably, in some of the data sets we found BayesCCE to demonstrate some differences between cases and controls which were not captured by the reference-based estimates (Figure 3.14). For example, NK cells abundance is known to change in aging in a process known as NK cell immunosenescence [102, 103], and monocyte levels are known to increase in RA patients compared with healthy individuals. [104, 105, 106]. These differences in cell populations were detected by BayesCCE but not by the reference-based method, thus suggesting that BayesCCE could uncover signal which was undetected by the reference-based method (Figure 3.14). That said, some other cell composition differences that were reported by BayesCCE but not by the reference-based method or vice versa may be the result of inaccuracies introduced by BayesCCE. Quantifying more accurately and reliably to what extent each method can detect cell composition differences would require several large data sets with known cell counts.

In addition, for each data set, we estimated the distribution of white blood cells based on the BayesCCE cell count estimates, and verified the ability of BayesCCE to correctly capture two distinct distributions (cases and controls or young and older individuals), regardless of the single distribution encoded by the prior information (Figure 3.14). While BayesCCE provides one component per cell type, these components are not necessarily appropriately scaled to provide cell count estimates in absolute terms. Therefore, for the latter analysis, we considered only the scenarios in which cell counts are known for a small number of individuals.

We further evaluated the scenario in which two different population-specific prior distributions are available. Specifically, one prior for cases and another one for controls in the case/control studies, and one for young and another one for older individuals in the aging study. For the purpose of this experiment, we estimated the priors using the reference-based estimates of a subset of the individuals (5% of the sample size) that were then excluded from the rest of the analysis. Interestingly, we found the inclusion of two prior distributions to provide no clear improvement over using a single general prior (Table 3.3). Thus, further confirming the robustness of BayesCCE to inaccuracies introduced by the prior information due to cell composition differences between populations.

Finally, we evaluated the effect of incorporating noisy priors on the performance of BayesCCE

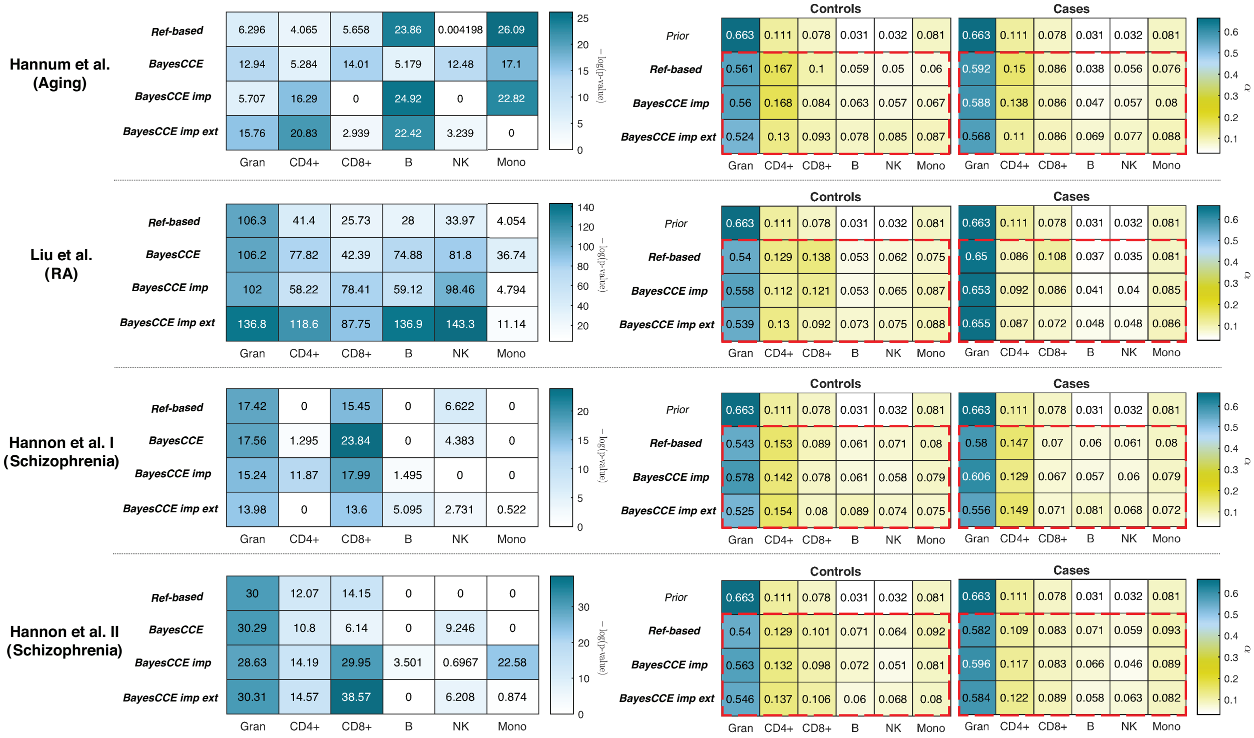

Figure 3.14: The robustness of BayesCCE to prior misspecification and its ability to capture population-specific variability in cell-type composition, under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells). Left side: t-test results (presented by the negative log of the Bonferroni-adjusted p-values) for the difference in proportions of each cell type between cases and controls. Right side: the Dirichlet parameters of estimated cell counts stratified by cases and controls; red dashed rectangles emphasize the high similarity in the estimated case/control-specific cell composition distributions yielded by the different methods, regardless of the prior used ("prior"). Results are presented for four different data sets and using cell count estimates obtained by four approaches: the reference-based method, BayesCCE, BayesCCE with known cell counts for 5% of the samples (BayesCCE imp), and BayesCCE with 5% additional samples with both known cell counts and methylation from external data (BayesCCE imp ext). For the Hannum et al. data set, for the purpose of presentation, cases were defined as individuals with age above the median age in the study. In the evaluation of BayesCCE imp and BayesCCE imp ext, samples with assumed known cell counts were excluded before calculating p-values and fitting the Dirichlet parameters.

by considering a range of possible priors with different levels of inaccuracies, including a non-informative prior (Figure 3.15). Not surprisingly, we observed that given cell counts for a small subset of samples, BayesCCE was overall robust to prior misspecification, which did not result in a substantially reduced performance even given a non-informative prior. In the absence of known cell counts, the performance of BayesCCE was somewhat decreased, however,

			Single prior		Stratified prior	
	data set	Method	MAC	MAE	MAC	MAE
$k = 3$	Hannum et al. [94] (Aging)	BayesCCE	0.661	0.102	0.667	0.105
		BayesCCE imp	0.829	0.022	0.830	0.021
	Liu et al. [67] (Rheumatoid arthritis)	BayesCCE	0.685	0.094	0.681	0.040
		BayesCCE imp	0.893	0.014	0.894	0.014
	Hannon et al. I [95] (Schizophrenia)	BayesCCE	0.632	0.111	0.633	0.111
		BayesCCE imp	0.784	0.017	0.785	0.016
	Hannon et al. II [95] (Schizophrenia)	BayesCCE	0.490	0.252	0.492	0.206
		BayesCCE imp	0.815	0.012	0.816	0.012
$k = 6$	Hannum et al. [94] (Aging)	BayesCCE	0.497	0.113	0.510	0.114
		BayesCCE imp	0.718	0.026	0.654	0.027
	Liu et al. [67] (Rheumatoid arthritis)	BayesCCE	0.537	0.041	0.557	0.058
		BayesCCE imp	0.711	0.024	0.697	0.023
	Hannon et al. I [95] (Schizophrenia)	BayesCCE	0.463	0.172	0.436	0.164
		BayesCCE imp	0.601	0.022	0.602	0.022
	Hannon et al. II [95] (Schizophrenia)	BayesCCE	0.485	0.086	0.471	0.075
		BayesCCE imp	0.603	0.023	0.613	0.024

Table 3.3: A summary of the performance of BayesCCE using a single prior versus using a separate prior for cases and controls (stratified prior). Mean absolute correlation (MAC) and mean absolute error (MAE) values are presented under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells), and under the assumption of three constituting cell types in blood ($k = 3$): granulocytes, monocytes and lymphocytes. A standard application of BayesCCE was compared with the scenario wherein cell counts are known for 5% of the samples (BayesCCE imp). In the later case, correlations were calculated after excluding the samples with assumed known cell counts. For the Hannum et al. data set, cases were defined as individuals with age above the median age in the study. For each data set, each of the calculated priors (the single general prior, the cases only prior and the controls only prior) was estimated using 5% of the samples in the data, which were then excluded from the subsequent analysis.

remained reasonable even in the scenario of a non-informative prior. Particularly, overall, BayesCCE with a non-informative prior performed better than the competing reference-free methods (ReFACTor, NNMF, and MeDeCom). We attribute this result to the combination of the constraints defined in BayesCCE with the sparse low-rank assumption it takes, which seems to handle more efficiently with the high-dimension nature of the computational problem (see section 3.2).

We note that in the presence of a non-informative prior, BayesCCE conceptually reduces to the performance of ReFACTor, and therefore it captures the same cell composition variability in the data. Yet, owing to the additional constrains, BayesCCE allows to overcome ReFACTor in capturing a set of components such that each component corresponds to one cell type.

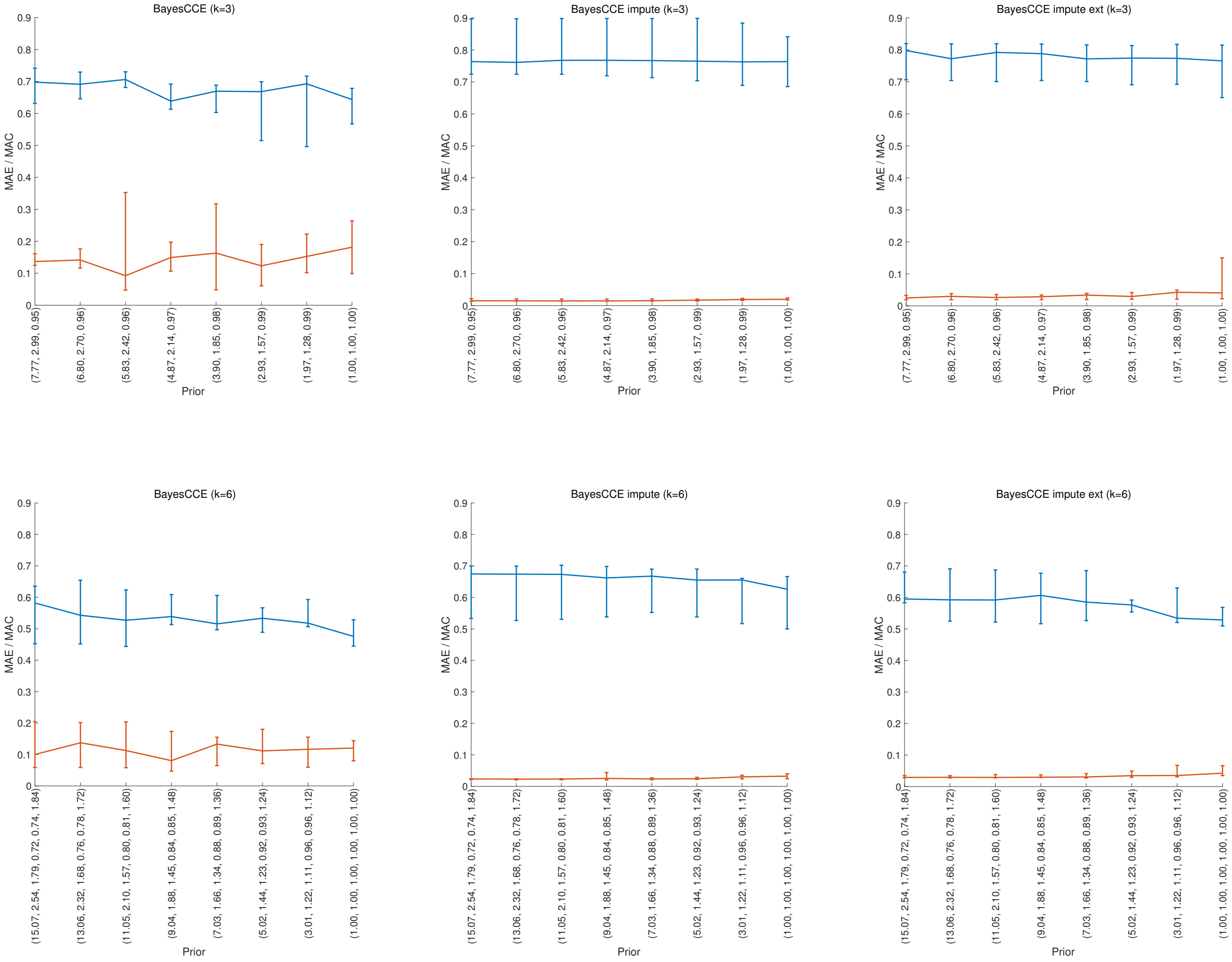

Figure 3.15: The performance of BayesCCE as a function of increasing noise introduced by the prior information, under the assumption of three constituting cell types in blood ($k = 3$): granulocytes, monocytes and lymphocytes (top panel), and under the assumption of six constituting cell types in blood ($k = 6$): granulocytes, monocytes and four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells; bottom panel). In this experiment, we evaluated BayesCCE, BayesCCE in a scenario wherein cell counts are known for 5% of the samples in the data (BayesCCE imp), and BayesCCE in a scenario wherein cell counts and methylation levels for samples from external data are included in the analysis (5% of the sample size; BayesCCE imp ext). For each method, presented are the values of mean absolute correlation (MAC) and mean absolute error (MAE) across all cell types as a function of the noise introduced into the prior information. Error bars indicate the performance across four data sets: Hannum et al. [94], Liu et al. [67], Hannon et al. I, and Hannon et al. II [95]. The range of the prior information was set between the prior estimated from real blood cell counts (see section 3.2) and a non-informative prior (a vector of ones).

3.4 Discussion

We introduce BayesCCE, a Bayesian method for estimating cell-type composition from heterogeneous methylation data without the need for methylation reference. We show mathematically and empirically the non-identifiability nature of the more straightforward reference-free NNMF approach for inferring cell counts, which tends to provide only linear combinations of the cell counts. In contrast, while we do not provide conditions for the uniqueness of a BayesCCE solution, our empirical evidence from multiple data sets clearly demonstrates the success of BayesCCE in providing desirable results of one component per cell type by leveraging readily obtainable prior information from previously collected data.

The parameters of the prior required by BayesCCE can be estimated by utilizing previous studies that collected cell counts from the tissue of interest. In our evaluation of the method, we used whole-blood methylation data, and we considered the classical definition of leukocyte cell types, which relies on cell surface markers. Considering other definitions of cell types is of potential interest; particularly, it would be interesting to examine to what extent BayesCCE and the reference-free methods can capture cell-type composition following a methylation-based definition of cell types (i.e. when defining cell types according to their methylation patterns).

Since BayesCCE captures cell composition variation under the classical definition of cell types by using the most dominant components of variation in the data, the main cell types of a natural methylation-based definition are expected to be a linear combination of the cell types under the classical definition. Much like in the experiments we presented here, wherein given a prior about the distribution of the cell types BayesCCE directed the solution towards an appropriate linear transformation, we would expect BayesCCE to perform similarly in the case of a methylation-based definition of cell types (given appropriate prior information about the distribution of cell types). Nevertheless, obtaining such a definition and evaluating BayesCCE under that definition would require obtaining appropriate single-cell methylation data, which is currently scarcely available. Moreover, deriving an actual meaningful definition of cell types given such data is a non-trivial problem. Therefore, until such definition and appropriate data are available, we are bounded to consider the classical definition of cell types.

Since BayesCCE requires a prior which can be estimated from previously collected cell counts without the need for any other genomic data, obtaining such as prior is relatively easy for many tissues, such as brain [107], heart [108] and adipose tissue [109]. Particularly, such data should be substantially easier to obtain compared to reference data from sorted cells for the corresponding tissues. Ideally, in order to learn the prior, one would want to use cell counts coming from the same population as the target population. Nevertheless, empirically, we observe that BayesCCE leverages the prior to direct the solution while still allowing enough flexibility, which makes it robust even to substantial deviations of the prior from the true underlying cell composition distribution. In fact, our results demonstrate that BayesCCE handles biases introduced by the prior remarkably well. Particularly, it allows to capture differences in cell compositions between different populations in the same study, thus providing an opportunity to study cell composition differences between different populations even in the absence of methylation reference.

Since no large data sets with measured cell counts are currently publicly available, we used a supervised method [85] for obtaining cell type proportion estimates, which were used as the ground truth in our experiments. Even though the method used for obtaining these estimates was shown to reasonably estimate leukocyte cell proportions from whole-blood methylation data in several independent studies [35, 79, 96], these estimates may have introduced biases into the analysis. Particularly, any inaccuracies introduced by the reference-based method could have directly affect the results of our evaluation. Our results indicate that such inaccuracies are more likely in some particular cell types over others. Failing to accurately estimate a particular cell type may be the outcome of various reasons. Notably, utilizing inappropriate reference data or failing to select a set of informative features that mark a particular cell type may dramatically affect its estimated values. Other reasons which are not methodological may also lead to inaccuracies of the estimates. For example, two cell types with very similar methylation patterns will be hardly distinguishable. In spite of the potential pitfalls of using estimates as a baseline for evaluation, we believe that our results on several independent data sets, including simulated data, and the use of a prior estimated from a large data set of high resolution cell counts, provide a compelling evidence for the utility of BayesCCE.

We further demonstrate that imputation of cell counts can be highly accurate when cell counts are available for some of the samples in the data. Particularly, based on our experiments, only as few as a couple of dozens of samples with known cell counts are needed in order to substantially improve performance. Moreover, in the general setup of BayesCCE, where no cell counts are known, each component corresponds to one cell type, however, not necessarily in the right scale and there is no automatic way to determine the identity of that cell type. In contrast, in the case of cell counts imputation, where cell counts are known for a subset of the samples, the assignment of components into cell types is straightforward. In addition, as we showed, BayesCCE is able to reconstruct cell counts up to a small absolute error (i.e. each component is scaled to form cell proportion estimates of one particular known cell type).

We note that in our evaluation of BayesCCE we considered only whole-blood data sets. Studying other tissues or biological conditions is clearly of interest. However, in the absence of other tissue-specific methylation references that were clearly shown to allow obtaining reasonable cell type proportion estimates, evaluation of performance based on tissues other than whole-blood will not be reliable. We therefore opt to focus on evaluating the performance of BayesCCE using multiple large whole-blood data sets. Importantly, beyond its potential utility for complex biological scenarios in which reference data is unavailable, BayesCCE may also provide an opportunity to improve cell count estimates in whole-blood studies in scenarios where the currently available reference data is not appropriate. Notably, in a recent work we have shown using multiple whole-blood data sets that ReFACTor outperforms the reference-based method in correcting for cell composition [92]. Differences in performance between ReFACTor (upon which BayesCCE relies for obtaining a starting point that captures the cell composition variation in the data) and the reference-based method are expected to be especially large in studies where the available reference data do not represent the individuals in the study well. We argue that this is likely to typically be the case, as the current go-to whole-blood reference consists of only six individuals [74], which represent a very specific and narrow population in terms of methylome altering factors, such as age [39], gender [40, 77], and genetics [70]. That said, large data sets with experimentally measured cell counts are required in order to fully investigate and demonstrate these claims.

We further note that in our benchmarking of BayesCCE with existing reference-free methods we considered only a subset of the available methods in the literature. Other reference-free methods that have been suggested in the context of accounting for cell composition in methylation data exist, however, these do not provide explicit components, but rather only implicitly account for cell composition variability in association studies. While in principle these methods can be modified to produce components, in this work we focused only on methods that can be readily used to provide explicit components for evaluation. We further note that several supervised and unsupervised decomposition methods have been suggested for estimating cell composition from gene expression [110, 111, 112, 113, 114]. However, these were refined for gene expression data and, to the best of our knowledge, none of these methods takes into account prior knowledge about the cell composition distribution as in BayesCCE. It remains of interest to investigate whether BayesCCE can be adapted for estimating cell composition from gene expression without the need for purified expression profiles.

Finally, our approach is based on finding a suitable linear transformation of the components found by ReFACTor [36]. It is therefore important to follow the guidelines for the application of ReFACTor, such as incorporation of methylation altering covariates; these guidelines were recently highlighted elsewhere [92, 93]. Since BayesCCE relies on the ReFACTor components, it is limited by their quality, and particularly, if the variability of some cell type is not captured by ReFACTor, BayesCCE will not be able to estimate that cell type well. Such a result is possible in scenarios where the variation of a particular cell type is substantially weaker than other sources of variation in the data (which are unrelated to cell-type composition); we note, however, that this potential limitation is not exclusive for ReFACTor or BayesCCE but rather a general limitation of all existing reference-free methods. BayesCCE will effectively provide the same result as ReFACTor if used for correcting for a potential cell-type composition confounder in methylation data. Since ReFACTor does not allow to infer direct cell count estimates but rather linear transformations of those, we suggest to use BayesCCE in cases in which a study of individual cell types is performed and therefore ReFACTor cannot be used. In case merely a correction for cell composition is desired, we suggest to use BayesCCE when cell counts are known for a subset of the samples, and otherwise to use ReFACTor.

CHAPTER 4

Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology

4.1 Background

Each cell type in the body of an organism performs a unique repertoire of required functions. Hence, disruption of cellular processes in particular cell types may lead to phenotypic alterations or development of disease. This presumption in conjunction with the complexity of tissue-level ("bulk") data has led to many cell-type-specific genomic studies, in which genomic features, such as gene expression levels, are assayed from isolated cell types in a group of individuals and studied in the context of a phenotype or condition of interest (e.g., [115, 116, 117, 118]).

In fact, in order to reveal cellular mechanisms affecting disease it is critical to study cell-type-specific effects. For example, it has been shown that cell-type-specific effects can contribute to our understanding of the principles of regulatory variation [119] and the underlying transcriptional landscape of heterogeneous tissues such as the human brain [120], it can provide a finer characterization of tumor heterogeneity [121, 122], and it may reveal disease-related pathways and mechanisms of genes that were detected in genetic association studies [123, 124]. Moreover, these findings are typically not revealed when a heterogeneous tissue is studied. For example, in [123] it has been shown that the FTO allele associated with obesity represses mitochondrial thermogenesis in adipocyte precursor cells. Particularly, in that study it is shown that the developmental regulators IRX3 and IRX5 had genotype-associated expression in primary preadipocytes, while genotype-associated expression was not observed in whole-adipose tissue, indicating that the effect was cell-type specific and restricted to preadipocytes.

In spite of the clear motivation to conduct studies with a cell-type-specific resolution, while developments in genomic profiling technologies have led to the availability of many large bulk data sets with hundreds or thousands of individuals (e.g., [43, 125, 126]), cell-type-specific data sets with a large number of individuals are still relatively scarce. Particularly, cell-type-specific studies are typically drastically restricted in their sample sizes owing to high costs and technical limitations imposed by both cell sorting and single-cell approaches. This restriction is especially profound for epigenetic studies with single-cell DNA methylation - while pioneering works on single-cell methylation have demonstrated significant advances (e.g. [127, 128, 129, 130]), profiling methylation with single-cell resolution is still limited in coverage and throughput and currently cannot be practically used to routinely obtain large-scale data for population studies (the most eminent recent studies included data from only a few individuals). This, in turn, substantially limits our ability to tackle questions such as identification of disease-related altered regulation of genes in specific cell types and mapping of diseases to specific manifesting cell types.

Technologies for profiling single-cell methylation are currently still under development, and some of these attempts will potentially allow sometime in the future for the analysis of cell-type-specific methylation across or within populations. However, even if such technologies emerge in the near future, the large number of existing bulk methylation samples that have been collected by now are still an extremely valuable resource for genomic research (e.g., more than 100,000 bulk profiles to date in the Gene Expression Omnibus (GEO) alone [11]). These data reflect years of substantial community-wide effort of data collection from multiple organisms, tissues, and under different conditions, and it is therefore of great importance to develop new statistical approaches that can provide cell-type-specific insights from bulk data.

Here, we introduce Tensor Composition Analysis (TCA), a novel computational approach for learning cell-type-specific DNA methylation signals (a tensor of samples by methylation sites by cell-types) from a typical two-dimensional bulk data (samples by methylation sites). Conceptually, TCA emulates the scenario in which each individual in the bulk data has been profiled with a single-cell resolution and then signals were aggregated in each cell population of the individual separately.

4.2 Methods

4.2.1 Enhancing epigenetic studies with cell-type-specific resolution

Different cell types are known to differ in their methylation patterns. Therefore, a bulk methylation sample collected from a heterogeneous tissue represents a combination of different signals coming from the different cell types in the tissue. Since cell-type composition varies across individuals, testing for correlation between bulk methylation levels and a phenotype of interest may lead to spurious associations in case the phenotype is correlated with the cell-type composition [72]. A widely acceptable solution to this problem is to incorporate the cell-type composition information into the analysis of the phenotype by introducing it as covariates in a regression analysis. Even though this procedure is useful for eliminating spurious findings, it does not take into account the fact that individuals are expected to vary in their methylation levels within each cell type (i.e. not just in their cell-type composition). Effectively, taking this approach results in an analysis that is conceptually similar to a study in which the cases and controls are matched on cell-type distribution, however, cell-type-specific signals are not explicitly modeled and leveraged.

In order to illustrate the above, consider the simple scenario, where the samples in the study are matched on cell-type distribution. Given no statistical relation between the phenotype and the cell-type composition, association studies typically assume a model with the following structure:

$$y_i = x_i\beta + \epsilon_i \tag{4.1}$$

Here, y_i represents the phenotypic level of individual i, x_i and β represent the bulk methylation level of individual i at a particular site under test and its corresponding effect size, and ϵ_i represents noise. This standard formulation assumes that a single parameter (β) describes the statistical relation between the phenotype and the bulk methylation level. We argue that this formulation is a major oversimplification of the underlying biology. In general, different cell types may have different statistical relations with the phenotype. Thus, a more realistic formulation would be:

$$y_i = \sum_{h=1}^{k} x_{ih}\beta_h + \epsilon_i \tag{4.2}$$

Here, $x_{i1}, ..., x_{ik}$ are the methylation levels of individual i in each of the k cell types composing the studied tissue and $\beta_1, ..., \beta_k$ are their corresponding cell-type-specific effects.

Applying a standard analysis as in Equation (4.1) to bulk data may fail to detect even strong cell-type-specific associations with a phenotype. For instance, consider the scenario of a case/control study, where the methylation of one particular cell type is associated with the disease. In this scenario, due to the signals arising from other cell types, the observed bulk levels may obscure the real association and not demonstrate a difference between the cases and controls; importantly, in general, merely taking into account the variation in cell-type composition between individuals does not allow the detection of the association (Figure 4.1). Thus, allowing analysis with a cell-type-specific resolution (i.e. obtaining $x_{i1}, ..., x_{ik}$ for each individual i) - beyond being required for revealing disease-manifesting cell types - is also important for the detection of true signals.

Notably, in the context of differential gene expression analysis, it has been previously suggested that cell-type-specific effects can be estimated by treating a phenotype of interest as a covariate (i.e. of the expression level) with potentially different effects on different cell types [131, 132]. Practically, this approach suggests to evaluate the effect of an interaction term (i.e. a multiplicative term) of the cell-type composition and the phenotype under a standard regression framework (i.e. by adding the interaction term to Equation (4.1)) [132]; equivalently, one may achieve the same goal by solving multiple decomposition problems

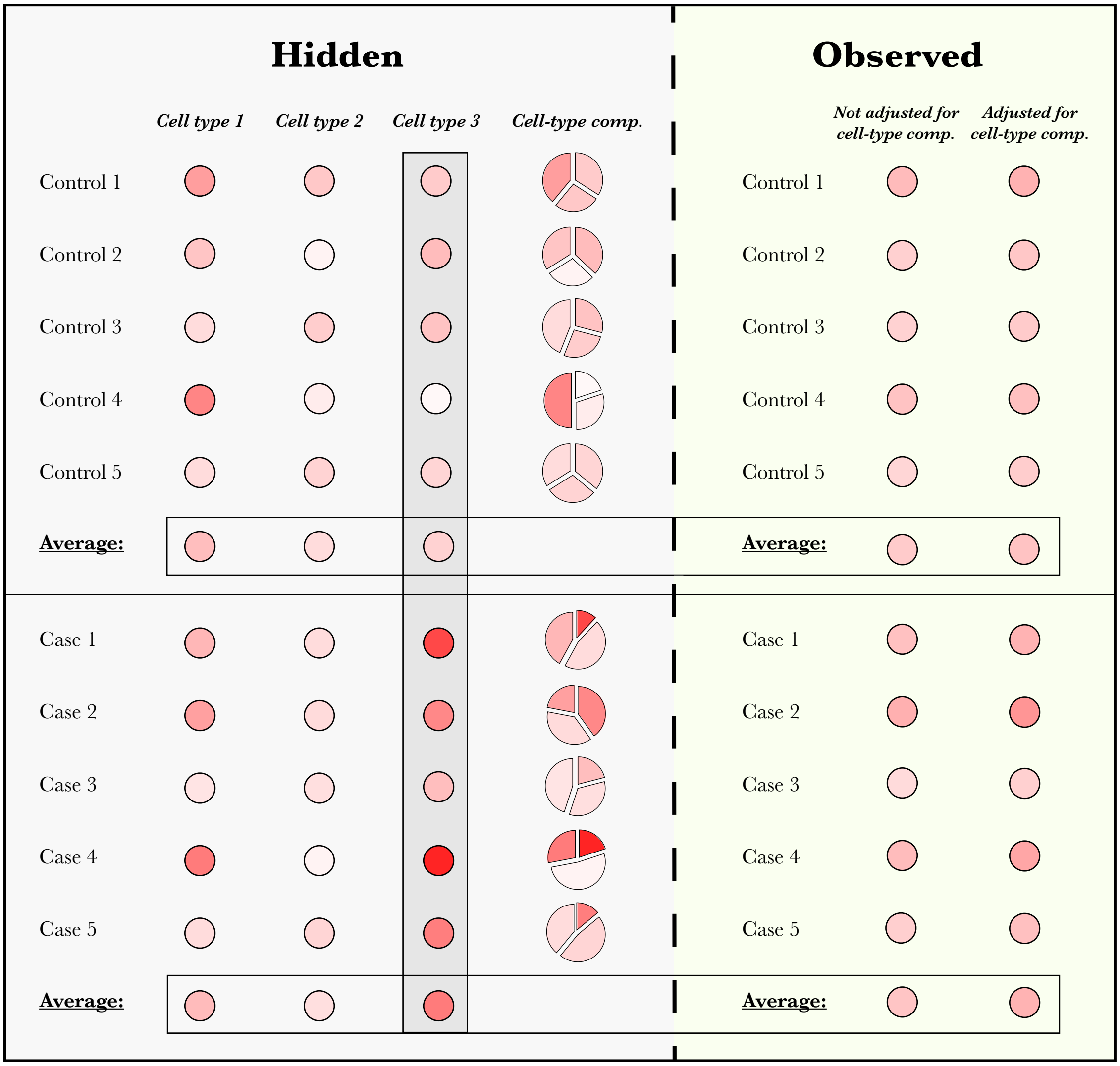

Figure 4.1: Observed bulk methylation levels may obscure cell-type-specific signals. Neither the observed methylation levels nor the observed levels after adjusting for the variability in cell-type composition can demonstrate a clear difference between cases and controls, in spite of a clear (unobserved) difference in cell type 3. Methylation levels are represented by a gradient of red color, and adjusted observed levels were calculated for each sample by removing the cell-type-specific mean levels, weighted by its cell-type composition.

(one for each possible value of the phenotype) [131]. In fact, this concept was recently applied and reported in the context of DNA methylation in attempt to detect cell-type-specific differences in methylation [133]. However, as we demonstrate below, a more detailed model of the variation in bulk methylation data allows a substantial improvement in power.

We propose a new model for DNA methylation, where we assume that the cell-type-specific

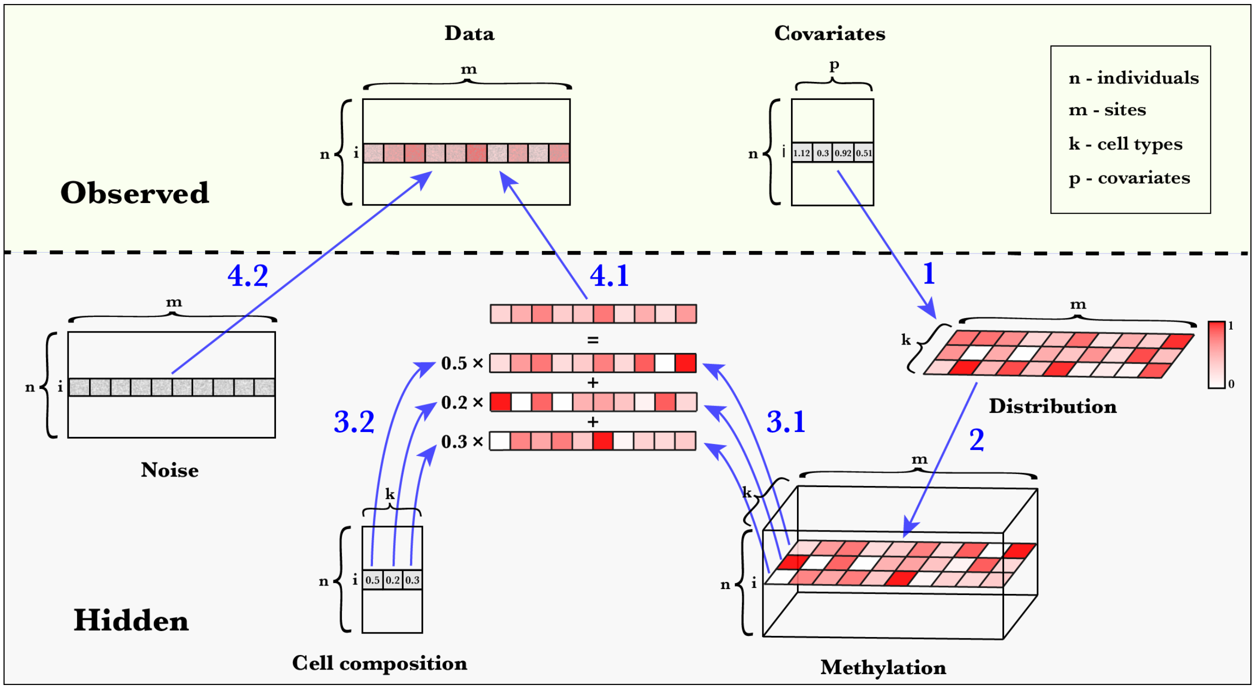

Figure 4.2: A summary of the TCA model for bulk DNA methylation data, presented as a four-steps generative model. Step 1: methylation altering covariates (e.g., age and sex) of a particular individual i can affect the methylation distribution of individual i. Step 2: the cell-type-specific methylomes of individual i are generated for each of the k cell types in the studied tissue. Step 3: the cell-type-specific methylomes of individual i (3.1) are combined according to the cell-type composition of the individual (3.2). Step 4: the true signal of the heterogeneous mixture (4.1) is distorted due to additional variation introduced by different sources of noise such as batch effects and other experiment-specific artificial variability (4.2); this results in the observed data. Methylation levels are represented by a gradient of red color

methylation levels of an individual are coming from a distribution that - up to methylation altering factors such as age [39] and sex [40] - is shared across individuals in the population. Based on this model, we developed Tensor Composition Analysis (TCA), a method for learning the unique cell-type-specific methylomes of each individual sample from its bulk data. We provide a detailed illustration of the model in Figure 4.2 and highlight the conceptual difference between TCA and a traditional decomposition approach in Figure 4.3. Here, we focus on the application of TCA for association studies, where we only implicitly consider the cell-type-specific methylomes of each individual by integrating over their distributions.

Importantly, TCA requires knowledge of the cell-type proportions of the individuals in the

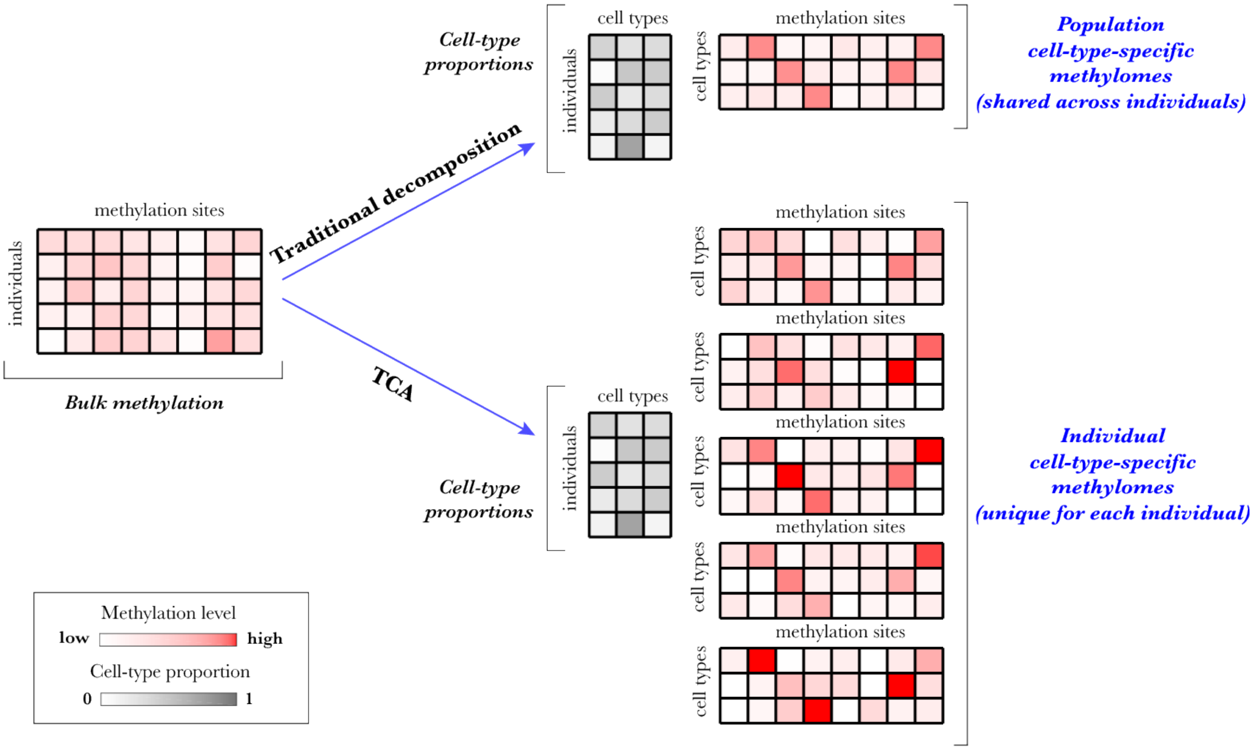

Figure 4.3: TCA versus a traditional decomposition approach. Given bulk DNA methylation data from a heterogeneous tissue, previous decomposition methods (e.g., PCA, ReFACTor [36], or a reference-based decomposition [85]) aim at estimating a matrix of the cell-type proportions of the individuals and a matrix of the cell-type-specific methylomes in the sample (shared across individuals). In contrast, TCA aims at estimating a matrix of the cell-type proportions of the individuals and - for each individual - a matrix of the unique cell-type-specific methylomes of the individual.

data. These can be computationally estimated using either a reference-based supervised approach [85] or a reference-free semi-supervised approach [134]; current reference-free unsupervised methods, however, are unable to provide reasonable estimates of cell-type proportions but rather only linear combinations of them [134]. Notably, in cases where only noisy estimates of the cell-type proportions are available (i.e. owing to inaccuracies of the computational method used for estimation), they can be used for initializing the optimization procedure of the TCA model, which can then provide improved estimates. As a result, as we show next, TCA performs well even in cases where only noisy estimates of the cell-type proportions are available.

In the following subsections, we summarize the model and mathematical methods. Further details, including the full mathematical derivations and the optimization procedures in TCA are given in Appendix A.

4.2.2 Modeling cell-type-specific variation in DNA methylation

Let Z^i_{hj} denote the value coming from cell type $h \in 1, ..., k$ at methylation site $j \in 1, ...m$ in sample $i \in 1, ...n$, we assume:

$$Z^i_{hj} | \mu_{hj}, \sigma_{hj} \sim N(\mu_{hj}, \sigma^2_{hj}) \tag{4.3}$$

In theory, the methylation status of a given site within a particular cell is a binary condition. However, unlike in the case of genotypes, methylation status may be different between different cells (even within the same individual, site and, cell type). We therefore consider a fraction of methylation rather than a fixed binary value. In array methylation data, possibly owing to the large number of cells used to construct each individual signal, we empirically observe that a normal assumption is reasonable.

Admittedly, normality may not hold for values near the boundaries (i.e. sites with mean methylation levels approaching 0 or 1); this can be addressed by applying variance stabilizing transformations such as a logit transformation (commonly referred to as M-values in the context of methylation) [135]. However, in practice, we ignore such consistently methylated or consistently unmethylated sites (e.g., in our experiments we discarded sites with mean value higher than 0.9 or lower than 0.1), which results in a set of sites that demonstrate an approximately linear relation with their respective M-values [135]. This makes the normality assumption reasonable and therefore widely accepted in the context of statistical analysis of DNA methylation.

Let $W \in \mathbb{R}^{k \times n}$ be a non-negative constant weights matrix of k cell types for each of the n samples (i.e. cell-type proportions; each column sums up to 1), we assume the following model for site j of sample i in the observed heterogeneous methylation data matrix X:

$$X_{ij} \quad = \quad \sum_{h=1}^{k} w_{hi} Z_{hj}^{i} + \epsilon_{ij}, \; \epsilon_{ij} \sim N(0, \tau^2) \tag{4.4}$$

where w_{hi} is the proportion of the h-th cell type of sample i in W, and ϵ_{ij} represents an additional component of measurement noise which is independent across all samples. We therefore get that X_{ij} follows a normal distribution with parameters that are unique for each individual i and site j. Put differently, we assume that the entries of X are independent but also different in their means and variances.

4.2.3 Tensor Composition Analysis (TCA)

Following the assumptions in (4.3) and in (4.4), the conditional probability of $Z_j^i = \left(Z_{1j}^i, ..., Z_{kj}^i\right)^T$ given X_{ij} can be shown (Appendix A) to satisfy

$$Pr(Z_j^i = z_j^i | X_{ij} = x_{ij}, w_i, \mu_j, \sigma_j, \tau) \propto exp\left(-\frac{1}{2}(a_{ij} - z_j^i)^T S_{ij}^{-1}(a_{ij} - z_j^i)\right) \tag{4.5}$$

where

$$\Sigma_j \quad = \quad diag(\sigma_{1j}^2, ..., \sigma_{kj}^2) \tag{4.6}$$

$$S_{ij} \quad = \quad \left(\frac{w_i w_i^T}{\tau^2} + \Sigma_j^{-1}\right)^{-1} \tag{4.7}$$

$$a_{ij} \quad = \quad S_{ij}\left(\frac{x_{ij}}{\tau^2} w_i + \Sigma_j^{-1} \mu_j\right) \tag{4.8}$$

Essentially, our suggested method, TCA, leverages the information given by the observed values $\{x_{ij}\}$ for learning a three-dimensional tensor consisted of estimates of the underlying values $\{z_{hj}^i\}$. This is done by setting the estimator $\hat{z}_j^i$ to be the mode of the conditional distribution in (4.5):

$$\hat{z}_j^i = a_{ij} = \left(\frac{w_i w_i^T}{\tau^2} + \Sigma_j^{-1}\right)^{-1} \left(\frac{x_{ij}}{\tau^2} w_i + \Sigma_j^{-1} \mu_j\right) \tag{4.9}$$

TCA requires the cell-type proportions W as an input. Given W, the parameters $\tau, \{\mu_j\}, \{\sigma_j\}$ can be estimated from the observed data under the assumption in (4.4). In practice, the cell-type proportions are typically unknown. In such cases, W can be estimated computationally using standard methods (e.g., [85, 134]) and then re-estimated under the TCA model in an alternating optimization procedure with the rest of the parameters in the model. The TCA model can further account for covariates, which may either directly affect Z_j^i (e.g., age and sex) or affect the mixture X_{ij} (e.g., batch effects). For more details and a full derivation of the conditional distribution of Z_j^i, while accounting for covariates, and for information about parameters inference see Appendix A.

In order to see why TCA can learn non-trivial information about the $\{z_{hj}^i\}$ values, consider a simplified case where $\tau = 0, \mu_{hj} = 0, \sigma_{hj} = 1$ for each h and a specific given j. In this case, it can be shown (Appendix A) that

$$Z_{hj}^i | X_{ij} = x_{ij} \sim N\left(\frac{w_{hi} x_{ij}}{\sum_{l=1}^k w_{li}^2}, 1 - \frac{w_{hi}^2}{\sum_{l=1}^k w_{li}^2}\right) \tag{4.10}$$

That is, given the observed value x_{ij}, the conditional distribution of Z_{hj}^i has a lower variance compared with that of the marginal distribution of Z_{hj}^i ($\sigma_{hj}^2 = 1$), thus reducing the uncertainty and allowing us to provide non-trivial estimates of the $\{z_{hj}^i\}$ values. This result further implies that in the context of DNA methylation, where the weights matrix W corresponds to a matrix of cell-type proportions, we should expect to gain better estimates for the $\{z_{hj}^i\}$ levels in more abundant cell types compared with cell types with typically lower abundance. For more details see Appendix A.

4.2.4 Applying TCA to epigenetic association studies

We next consider the problem of detecting statistical associations between DNA methylation levels and biological phenotypes. Let $X \in \mathbb{R}^{n \times m}$ be an individuals by sites matrix of methylation levels, and let Y denote an n-length vector of phenotypic levels measured from the same n individuals, typical association studies usually consider the following model for testing a particular site j for association with Y:

$$Y_i \;=\; X_{ij}\beta_j + e_i,\; e_i \sim N(0,\sigma^2) \tag{4.11}$$

where Y_i is the phenotypic level of individual i, β_j is the effect size of the j-th site, and e_i is a component of i.i.d. noise. For convenience of presentation, we omit potential covariates which can be incorporated into the model. In a typical EWAS, we fit the above model for each feature, and we look for all features j for which we have a sufficient statistical evidence of non-zero effect size (i.e. $\beta_j \neq 0$).

In principle, one can use TCA for estimating cell-type-specific levels, and then look for cell-type-specific associations by fitting the model in (4.11) with the estimated cell-type-specific levels (instead of directly using X). However, an alternative one-step approach can be also used. This approach leverages the information we gain about z^i_{hj} given that $X_{ij} = x_{ij}$ for directly modeling the phenotype as having cell-type-specific effects. Specifically, consider the following model:

$$Y_i = Z^i_{lj}\beta_{lj} + e_i, e_i \sim N(0,\phi^2) \tag{4.12}$$

where β_{lj} denotes the cell-type-specific effect size of some cell type of interest l. Provided with the observed information x_{ij}, while keeping the assumptions in (4.3) and in (4.4), it can be shown (Appendix A) that:

$$Y_i|X_{ij} = x_{ij} \sim N\left(\beta_{lj}\left(\mu_{lj} + \frac{w_{li}\sigma^2_{lj}\tilde{x}_{ij}}{\tau^2 + \sum_{h=1}^k w^2_{hi}\sigma^2_{hj}}\right), \phi^2 + \beta^2_{lj}\left(\sigma^2_{lj} - \frac{w^2_{li}\sigma^4_{lj}}{\tau^2 + \sum_{h=1}^k w^2_{hi}\sigma^2_{hj}}\right)\right) \tag{4.13}$$

$$\tilde{x}_{ij} = x_{ij} - \sum_{h=1}^{k} w_{hi}\mu_{hj} \tag{4.14}$$

This shows that directly modeling $Y_i|X_{ij}$ effectively integrates the information over all possible values of Z^i_{lj}. Given W, μ_j, σ_j, τ (typically estimated from X; Appendix A), we can estimate ϕ and the effect size β_{lj} using maximum likelihood. The estimate $\hat{\beta}_{lj}$ can be then tested for significance using a generalized likelihood ratio test. Similarly, we can consider a joint test for the combined effects of more than one cell type. A full derivation of the statistical test is described in Appendix A. Here, whenever association testing was conducted, we used this direct modeling of the phenotype given the observed methylation levels.

Finally, we note that in principle one can also use the model in Equation (4.4) for testing for cell-type-specific associations by treating the phenotype of interest as a covariate and estimating its cell-type-specific effect size. However, TCA provides a way to deconvolve the data into cell-type-specific levels, which is of independent interest beyond the specific application for association studies. Moreover, model directionality often matters, and the TCA framework allows us to directly model the phenotype rather than merely treat it as another covariate. Particularly, in the context of this work, it is known that methylation levels are actively involved in many cellular processes such as regulation of gene expression [136], thus, making DNA methylation a potential contributing determinant in disease (which further justifies the modeling of the phenotype as an outcome).

4.2.5 Implementation of TCA

A Matlab implementation of TCA was used for deriving all the results reported here, and an additional implementation in R was deposited as a CRAN package ('TCA'). The source code of both implementations is available from github at http://github.com/cozygene/TCA.

TCA requires for its execution a heterogeneous DNA methylation data matrix and corresponding cell-type proportions for the samples in the data. In case where cell counts are not available, TCA can take estimates of the cell-type proportions, which are then optimized with the rest of the parameters in the model. For the real data experiments, we used GLINT [93] for generating initial estimates of the cell-type proportions for the whole-blood data sets. GLINT provides estimates according to the Houseman et al. model [85], using a panel of 300 highly informative methylation sites in blood [96] and a reference data collected from sorted blood cells [74]. Given these estimates, we used the TCA model to re-estimate the cell-type proportions using the top 500 sites selected by the feature selection procedure of ReFACTor [36].

4.2.6 Data simulation

We first estimated cell-type-specific means and standard deviations in each site using reference data of methylation levels collected from sorted blood cells [74]. Since we expected cell-type-specific associations to be mostly present in CpG sites that are highly differentially methylated across different cell types, we considered cell-type-specific means and standard deviations from sites which demonstrated the highest variability in cell-type-specific mean levels across the different cell types. Using the estimated parameters of a given site, we generated cell-type-specific DNA methylation levels using normal distributions, conditional on the range $[0, 1]$. In cases where covariates were simulated to have an effect on the cell-type-specific methylation levels, the means of the normal distributions were tuned for each sample to account for its covariates and the corresponding effect sizes (shared across samples; Appendix A).

We generated cell-type proportions for each sample using a Dirichlet distribution with parameters set according to previous estimates from cell counts of 6 blood cell types [134]: 15.0727, 1.8439, 2.5392, 1.7934, 0.7240, and 0.7404, which correspond to Dirichlet parameters for granulocytes, monocytes and 4 sub-types of lymphocytes (CD4+, CD8+, B and NK cells). In the case of three constituting cell types (granulocytes, monocytes, and lymphocytes), we set the Dirichlet parameter of lymphocytes to be the sum of the parameters of all the lymphocyte sub-types. For the experiments with a nonparametric distribution of the cell-type proportions we sampled proportions of individuals from a pool of reference-based estimates that were estimated using a reference-based method [85] for samples in two data sets (described below) [67, 94].

Eventually, for each sample, we composed its methylation level at each site by taking a linear combination of the simulated cell-type-specific levels of that site, weighted by the cell composition of that sample, and added an additional i.i.d normal noise conditional on the range $[0, 1]$ to simulate technical noise ($\tau = 0.01$). In cases where covariates were simulated to have a global effect on the methylation levels (i.e. non-cell-type-specific effect, such as batch effects), we further added an additional component of variation for each sample according to its global covariates and their corresponding effect sizes.

4.2.7 Data sets

We used a total of five methylation data sets, all of which were collected using the Illumina 450K human DNA methylation array and are available from the Gene Omnibus Database (GEO). In more details, we used 3 methylation data sets that were previously collected in RA studies: a whole-blood data set by Liu et al. of 354 RA cases and 332 controls (GEO accession GSE42861) [67], a CD4+ methylation data set of 12 RA cases and 12 controls with matching age and sex (for each RA patient, a control sample with matching age and sex was collected) by Guo et al. (GEO accession GSE71841) [137], and cell-sorted methylation data collected from 63 female RA patients and 31 female control subjects in CD4+ memory cells, CD4+ naive cells, CD14+ monocytes, and CD19+ B cells (a total of 371 samples across four cell sub-types; GEO accession GSE131989); these cell-sorted data were originally described by Rhead et al. [138]. In addition, for replicating the association results with immune activity, we used another data set that was previously studied by Hannum et al. in the context of aging rates (n=656; GEO accession GSE40279) [94]. Finally, for the simulation experiments we used methylation reference of sorted leukocyte cell types collected in 6 individuals from the (GEO accession GSE35069) [74].

We processed the data similarly to a recently suggested normalization pipeline [97]. Specifically, we processed the raw IDAT files of the Liu et al. data set [67] and the Rhead et al. data set [138] (each cell sub-type separately) using the "minfi" R package [49] as follows. We removed 65 SNP markers and applied the Illumina background correction to all intensity values, while analyzing probes coming from autosomal and non-autosomal chromosomes separately. We considered a threshold of 10e-16 for the detection p-value of intensity values; probes with p-values higher than this threshold were treated as missing values, and samples with call rate <95% and probes with call rate <90% were excluded. Since IDAT files were not made available for the Hannum et al. data [94] and the Guo et al. data [137], we used the methylation intensity levels published by the authors. For each data set, we then performed a quantile normalization of the methylation intensity levels, subdivided by probe type, probe sub-type, and color channel, and imputed missing values using the ”impute” R package (using the function impute.knn). Eventually, we calculated Beta-normalized methylation levels based on the normalized intensity levels (according to the recommendation by Illumina).

We further excluded samples from the above data sets as follows. In the Liu et al. data set, we excluded two samples that demonstrated extreme values in their first two principal components (over four empirical standard deviations) and two more of the remaining samples that were regarded as outliers in the original study of Liu et al. In the Rhead et al. data set, we excluded a small batch that consisted of only 4 individuals, and in the Hannum et al. data set we removed six samples that demonstrated extreme values in their first two principal components (over four empirical standard deviations). The final numbers of samples remained for analysis in the Liu et al. data set, the Hannum data set and the Guo et al. data set were n=658, n=650, and n=24, respectively. The numbers of samples remained for analysis in the Rhead et al. data were n=89, n=88, n=90, and n=86 for the CD4+ memory cells, CD4+ naive cells, monocytes, and B cells, respectively.

Finally, for the association experiments, we discarded consistently methylated probes and consistently unmethylated probes from the data (mean value higher than 0.9 or lower than 0.1, respectively, according to the Liu et. al discovery data), and we further used GLINT [93] to exclude from the data CpGs coming from the non-autosomal chromosomes, as well as polymorphic and cross-reactive sites, as was previously suggested [37].

4.2.8 Power simulations

We simulated data and sampled for each site under test a normally distributed phenotype with additional effects of the cell-type-specific methylation levels of the site. We set the variance of each phenotype to the variance of the site under test, in order to eliminate the dependency of the power in the variance of the tested site (and therefore allow a clear quantification of the true positives rate under a given effect size). Particularly, when simulating an effect coming from a single cell type, we randomly generated a phenotype from a normal distribution with the variance set to the variance of the site under test in the specific cell type under test. Similarly, when simulating effects coming from all cell types, we randomly generated a phenotype from a normal distribution with the variance set to the total variance of the site under test (i.e. across all cell types).

We performed the power evaluation using simulated data with 3 constituting cell types (k=3) and using simulated data with 6 constituting cell types (k=6). We considered three scenarios across a range of effect sizes as follows: different effect sizes for different cell types (using s joint test), the same effect size for all cell types (using a joint test, under the assumption of the same effect for all cell types), and a scenario with only a single associated cell type (a marginal test). In the first scenario, effect sizes for the different cell types were drawn from a normal distribution with the particular effect size under test set to be the mean (with standard deviation $\sigma = 0.05$), and in the third scenario we evaluated the aggregated performance of all the marginal tests across all constituting cell types in the simulation. We further repeated the marginal test while stratifying the evaluation by cell type (i.e. the marginal test was performed under the third scenario for each cell type separately). In each of these experiment, we calculated the true positives rate of the associations that were reported as significant while adjusting for the number of sites in the simulated data.

For each scenario and for each number of constituting cell types, we simulated 10 data sets, each included 500 samples and 100 sites. Importantly, throughout the simulation study, we considered for each simulated data set the case where only noisy estimates of the cell-type proportions are available (and therefore need to be re-estimated together with the rest of the parameters in the TCA model). Specifically, for each sample in the data we replaced its cell-type proportions with randomly sampled proportions coming from a Dirichlet distribution with the original cell-type proportions of the individuals as the parameters. For each level of noise, these parameters were multiplied by a factor that controlled the level of similarity of the sampled proportions to the original proportions. Finally, for evaluating false positives rates, we followed the above procedure, however, without adding additional effects coming from methylation levels. We evaluated the false positives rate by considering the fraction of sites with p-value<0.05.

4.2.9 Analysis of immune activity

We used the Liu et al. data [67] as the discovery data (n=658) and the Hannum et al. data [94] as the replication data (n=650). Since we expected to observe associations with regulation of cell-type composition in CpGs that demonstrate differential methylation between different cell types, we considered for this analysis only CpGs that were reported as differentially methylated between different whole-blood cell types [72]. Specifically, we considered the sites in the intersection between the set of Bonferroni-significant CpGs that were reported as differentially methylated in whole-blood and the available CpGs in both the discovery and replication data sets; this resulted in a set of 50,123 CpGs that were available for this analysis.

We performed a standard linear regression analysis using GLINT [93] and a TCA analysis under the assumption of the same effect size in all cell types. In the analysis of the Liu et al. data we controlled for RA status, gender, age, smoking status, and known batch information, and in the analysis of the Hannum et al. data we controlled for gender, age, ethnicity and the first two EPISTRUCTURE principal components [70] in order to account for the population structure in this data set. In both data sets, in order to take into account potentially unknown technical confounding effects, we further included the first ten principal components calculated from the intensity levels of a set of 220 control probes in the Illumina methylation array, as suggested by Lenhe et al. [97] in an approach similar to the remove unwanted variation method (RUV) [139]. These probes are expected to demonstrate no true biological signal and therefore allow to capture global technical variation in the data.

In the replication analysis, we applied a Bonferroni threshold in reporting significance, controlling for the number of genome-wide significant associations that were reported in the discovery data. The results are summarized in Supplementary Data 1 in [140], where additional description for the associated genes is provided from GeneCards [141], the GWAS catalog [142], and GeneHancer [143].

4.2.10 Analysis of rheumatoid arthritis

We used the Liu et al. data [67] as the discovery data (n=658, 214,096 Cpgs). We applied a standard logistic regression analysis with the RA status as an outcome using GLINT [93] and TCA analysis: under the assumption of a single effect for all cell types (joint test), and for each of CD4+, CD14+, and CD19+, under the assumption of a single associated cell type (marginal test). In every analysis, we accounted for the same variables described in the immune activity analysis with this data set. In the TCA analysis, we additionally accounted for the first six ReFACTor components [36], calculated according to the most recent updated guidelines [92]. In order to test the associations reported by TCA for enrichment for the RA pathway, we used missMethyl [144], an R package that allows to run enrichment analysis for disease directly on CpGs (while accounting for gene length bias).

In the validation analysis with the Rhead et al. data, we applied a standard logistic regression analysis using GLINT [93] on each of the CD14+ (n=90) and CD19+ (n=86) data sets, while accounting for age, smoking status, and batch information. Since the Rhead et al. data included sorted-cell methylation from two sub-types of CD4+, for the replication analysis of CD4+ (n=81) we performed for each site a logistic regression analysis using both its CD4+ naive cells methylation levels and CD4+ memory cells methylation.

Taking a standard regression approach in the analysis of the Guo et al. CD4+ sorted methylation data resulted in a severe inflation in test statistic. Since the cases and controls in the sample were matched for age and sex, we suspected that technical variation might have led to this inflation. In order to test that, we calculated the first principal component of control probes, similarly to the approach taken in the analysis of the Liu et al. data. However, since IDAT files were not available for the Guo et al .data, and therefore the same set of 220 control probes that were used in the Liu et al. data were not available, we used the methylation intensity levels of the 220 sites with the least variation in the data as control probes. Indeed, we found that the first PC of the control probes corresponds to the case/control status in the data almost perfectly (r=0.91, p-value=6.29e-10). As a result, p-values obtained using a standard analysis of the Guo et al. data set are not reliable. We therefore considered the following nonparametric procedure. We ranked the sites according to their absolute difference in mean methylation levels between cases and controls, and considered a simple enrichment test, wherein the p-value of a site was determined as its rank divided by the total number of sites in the ranking.

The results are summarized in Supplementary Data 2 in [140], where additional description for the associated genes is provided from GeneCards [141], the GWAS catalog [142], and GeneHancer [143].

4.2.11 Application of CellDMC and HIRE

We applied CellDMC using the corresponding R package by Zheng et al. [133], and provided it with the true cell-type proportions as an input throughout our simulation study, and with the same covariates we used for TCA in the real data analysis. We further applied HIRE using the corresponding R package by Luo et al. [145]. Unlike CellDMC, HIRE treats the cell-type proportions as parameters that are being estimated as part of the optimization process. Therefore, in order to provide it with a similar advantage to CellDMC, which was given access to the true cell-type proportions in the simulation study, we assigned the initial cell-type proportion estimates in the HIRE code to be the true cell-type proportions.

Since both CellDMC and HIRE provide only test statistics and p-values for the effects of individual cell types (i.e. only for marginal tests and not for a joint, CpG-level test), in the power simulations with effects in multiple cell types we considered a CpG to be associated with the phenotype if it had a significant association with at least one of the cell types. To make our benchmarking of TCA with these methods conservative, we allowed a favorable procedure for CellDMC and HIRE in these cases by not accounting for the number of cell types (i.e. just for the number of CpGs) when calculating true positive rates.

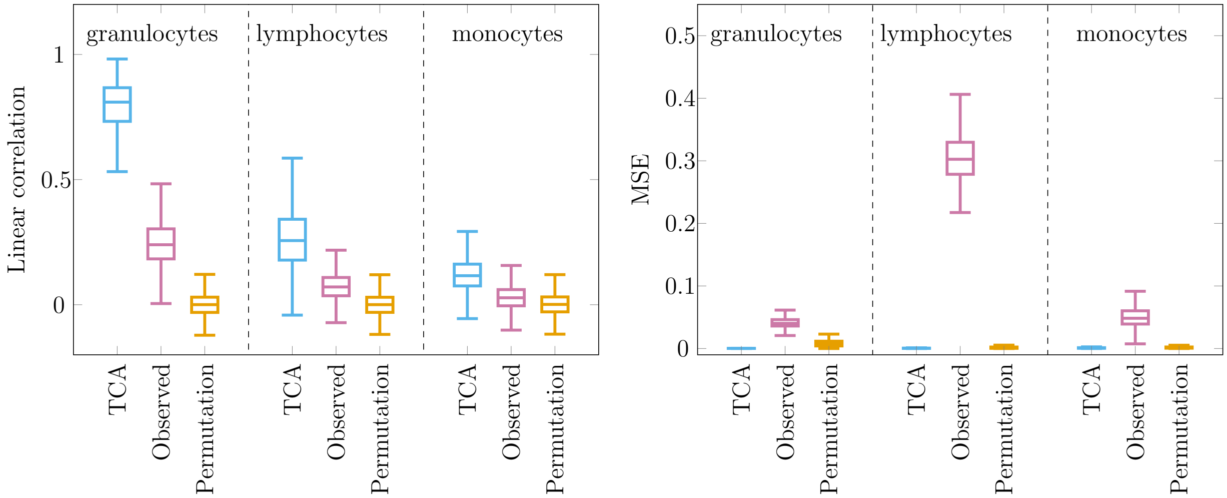

Figure 4.4: Reconstructing cell-type-specific methylation levels from simulated bulk whole-blood data with three constituting cell types ($k = 3$; 250 samples, 250 sites). Three approaches were evaluated in capturing the cell-type-specific levels of each site j and cell type h across all individuals $z_{hj} = (z_{hj}^1, ..., z_{hj}^n)$: TCA, TCA after permuting the observed data matrix ("Permutation") and directly using the observed bulk data ("Observed"; i.e. using the bulk as the estimate for the cell-type-specific levels of each cell type). For each of the evaluated approaches and for each of the simulated cell types (ordered by their mean abundance), presented are the distributions of the linear correlation between z_{hj} and its estimate $\hat{z}_{hj}$ across all sites j and across ten simulated data sets (left), and the distribution of the MSE between z_{hj} and its estimate $\hat{z}_{hj}$ across all sites j and across ten simulated data set (right). The central mark on each box indicates the median, and the bottom and top edges indicate the 25th and 75th percentiles, respectively.

4.3 Results

4.3.1 Detecting cell-type-specific associations using TCA

In order to empirically verify that TCA can learn cell-type-specific methylation levels, we first leveraged whole-blood methylation data collected from sorted leukocytes [74] to simulate heterogeneous bulk methylation data. While the bulk data captured the cell-type-specific signals to some extent, as expected, TCA performed substantially better (Figures 4.4 and 4.5).

We next evaluated the performance of TCA in detecting cell-type-specific associations by simulating whole-blood methylation and corresponding phenotypes with cell-type-specific effects. We compared the performance of TCA with a standard regression analysis of the bulk levels and with the method CellDMC, an interaction-based test that was recently evaluated

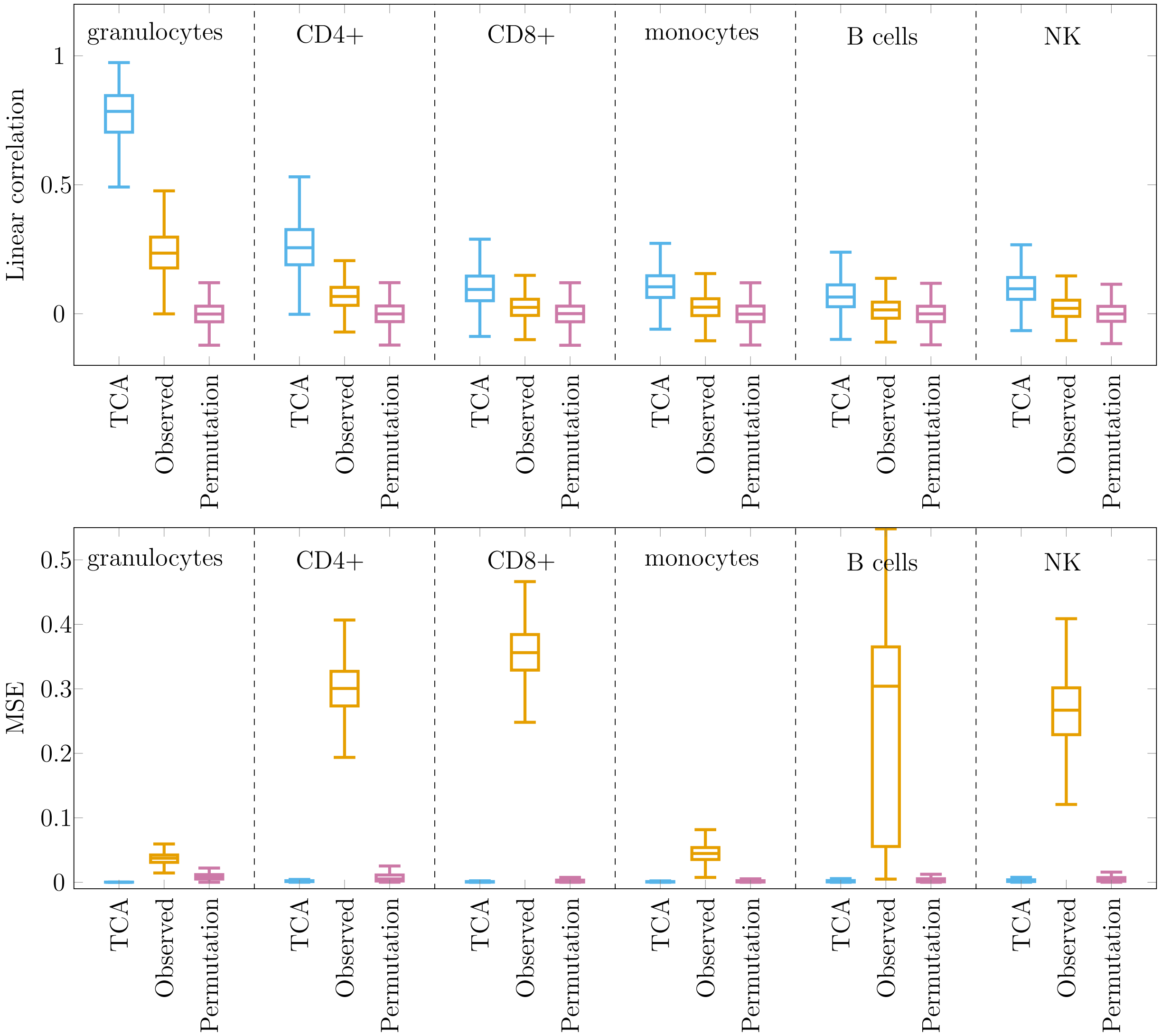

Figure 4.5: Reconstructing cell-type-specific methylation levels from simulated bulk whole-blood data with six constituting cell types ($k = 6$; 250 samples, 250 sites). Three approaches were evaluated in capturing the cell-type-specific levels of each site j and cell type h across all individuals $z_{hj} = (z_{hj}^1, ..., z_{hj}^n)$: TCA, TCA after permuting the observed data matrix ("Permutation") and directly using the observed bulk data ("Observed"; i.e. using the bulk as the estimate for the cell-type-specific levels of each cell type). For each of the evaluated approaches and for each of the simulated cell types (ordered by their mean abundance), presented are the distributions of the linear correlation between z_{hj} and its estimate $\hat{z}_{hj}$ across all sites j and across ten simulated data sets (top), and the distribution of the MSE between z_{hj} and its estimate $\hat{z}_{hj}$ across all sites j and across ten simulated data set (bottom). The central mark on each box indicates the median, and the bottom and top edges indicate the 25th and 75th percentiles, respectively.

in the context of detecting cell-type-specific associations with methylation [133]. Notably, we provided CellDMC with the true underlying cell-type proportions as an input. Beyond introducing interaction terms into a standard regression framework, CellDMC also considers additive effects of the cell-type composition. Given the true cell-type proportions, it therefore achieves a perfect linear correction for cell-type composition. Hence, CellDMC practically reflects in our experiments an upper bound for the performance of any standard method that merely accounts for linear differences in cell-type composition across individuals.

Our experiments verify that TCA yields a substantial increase in power over the alternatives under different scenarios Particularly, in its worst performing scenario, TCA achieved a median of 2.25 fold increase in power (across all tested effect sizes) over the standard regression approach and a median of 11.15 fold increase in power in the best performing scenario (Figure 4.6); compared with CellDMC, TCA achieved a median of between 1.46 and 12.25 fold increase in power across all scenarios. Repeating these experiments while including cell-type-specific affecting covariates and under a nonparametric distribution of the cell-type proportions (i.e. rather than a parametric one) demonstrated similar results (Figure 4.7).

Remarkably, TCA demonstrated the highest improvement in a scenario where all cell types had the exact same effect size, although this is intuitively a favorable scenario for a standard regression analysis, which does not model cell-type-specific signals (Figure 4.6). Interestingly, in spite of the high power achieved by TCA, we found it to be conservative (i.e. less false positives than expected; Figure 4.8); this can be explained by the optimization procedure of the model (Appendix A).

Finally, we performed an additional power analysis stratified by cell types, which, once again, showed that TCA robustly outperforms the alternative approaches (Figures 4.9 and 4.10). This analysis further revealed that under the scenario of a single causal cell type, TCA achieved better power when the causal cell type was highly abundant (as opposed to lowly abundant); these results are expected, given that bulk signals are mostly dominated by abundant cell types. For instance, considering a moderate effect size corresponding to a signal-to-noise ratio of 1, we found that TCA achieved a median power of 1 and 0.52 in granulocytes and CD4+ cells (the two most abundant cell types; mean abundance of 0.67 and 0.11, respectively), yet only a limited power in the less abundant cell types; for example,

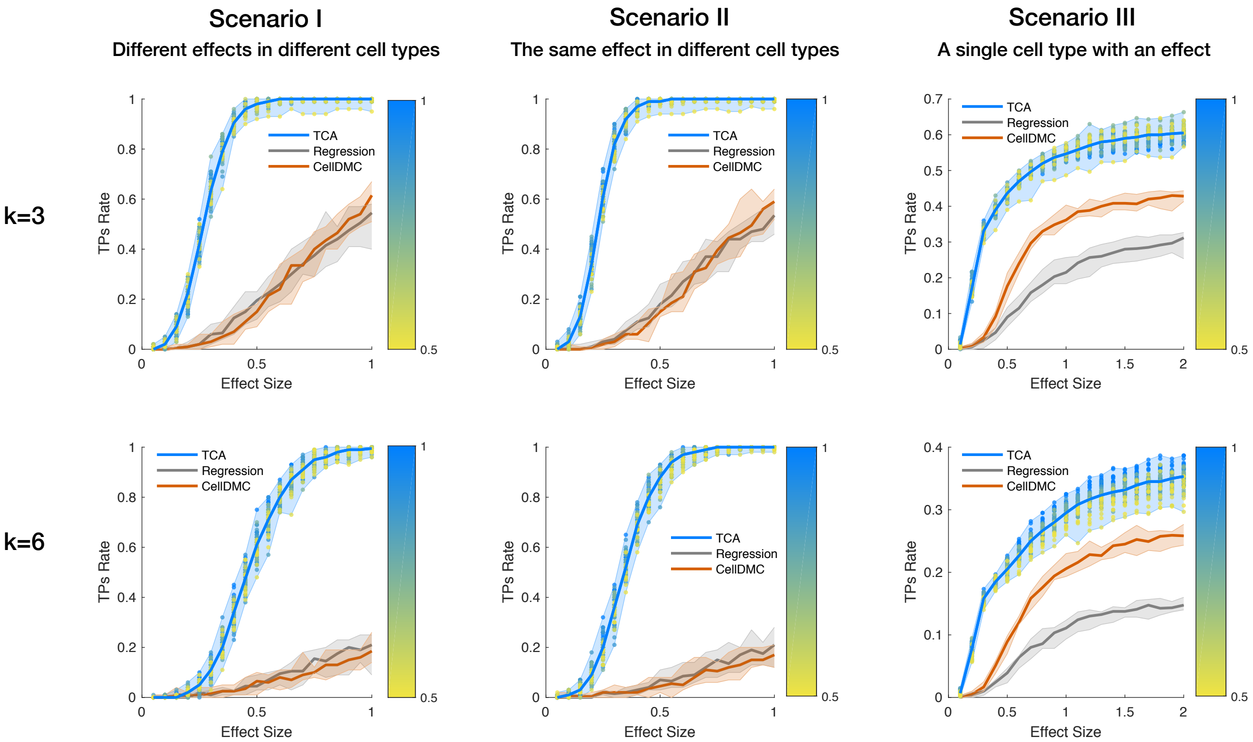

Figure 4.6: An evaluation of power for detecting cell-type-specific associations with DNA methylation. Performance was evaluated using three approaches: TCA, a standard linear regression with the observed bulk data, and CellDMC with the true cell-type proportions as an input. The numbers of true positives (TPs) were measured under three scenarios using a range of effect sizes: different effect sizes for different cell types (Scenario I), the same effect size for all cell types (Scenario II), and a single effect size for a single cell type (Scenario III); each of the scenarios was evaluated under the assumption of three constituting cell types (k=3; top row) and six constituting cell types (k=6; bottom row). Lines represent the median performance across 10 simulations and the colored areas reflect the results range across the multiple executions. The colored dots reflect the results of TCA under different initializations of the cell-type proportion estimates (i.e. different levels of noise injected into TCA), where the color gradients represent the mean absolute correlation of the initial estimates with the true values (across all cell types).

in the two least abundant cell types considered, B cells and NK cells (mean abundance 0.03 for both), TCA could only achieve a median power of 0.08 and 0.03 under the same effect size (Figure 4.9).

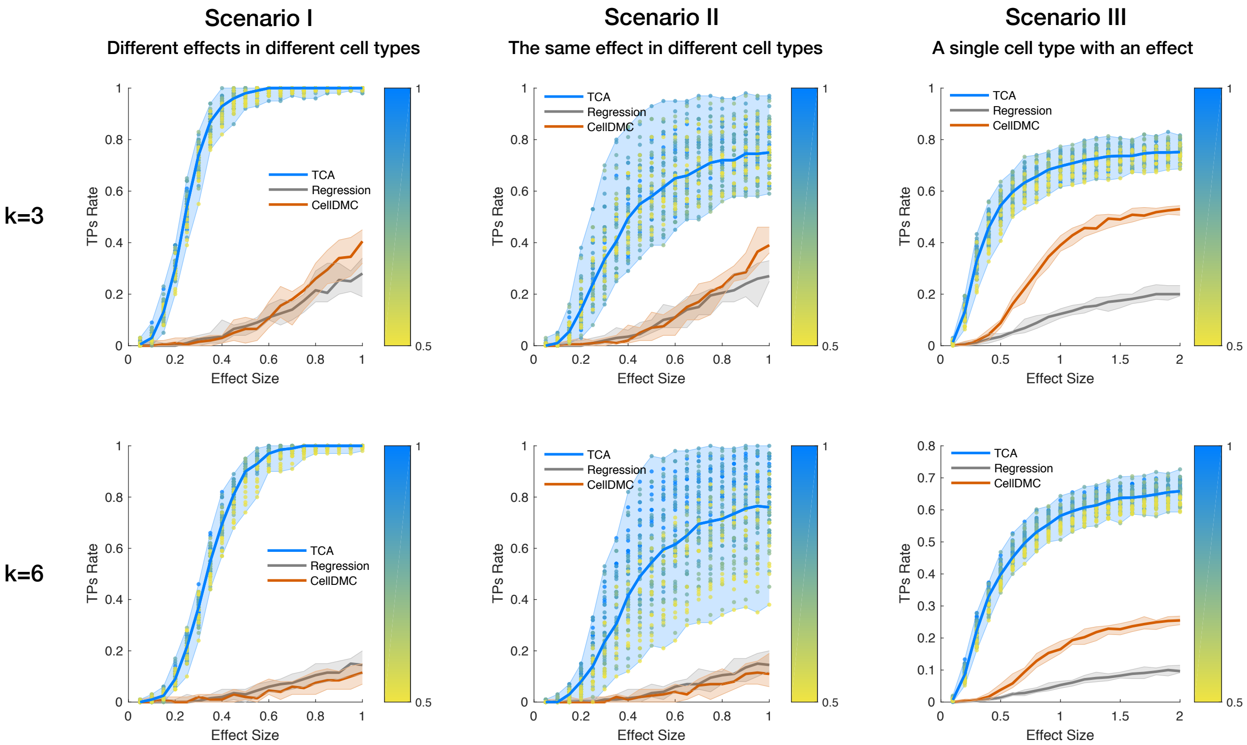

Figure 4.7: An evaluation of power for detecting cell-type-specific associations with DNA methylation while including cell-type-specific affecting covariates and using a nonparametric distribution of the cell-type proportions. Performance was evaluated using three approaches: TCA, a standard linear regression with the observed bulk data, and CellDMC with the true cell-type proportions as an input. The numbers of true positives (TPs) were measured under three scenarios using a range of effect sizes: different effect sizes for different cell types (Scenario I), the same effect size for all cell types (Scenario II), and a single effect size for a single cell type (Scenario III); each of the scenarios was evaluated under the assumption of three constituting cell types (k=3; top row) and six constituting cell types (k=6; bottom row). Lines represent the median performance across 10 simulations and the colored areas reflect the results range across the multiple executions. The colored dots reflect the results of TCA under different initializations of the cell-type proportion estimates (i.e. different levels of noise injected into TCA), where the color gradients represent the mean absolute correlation of the initial estimates with the true values (across all cell types).

4.3.2 Cell-type-specific differential methylation in immune activity

In general, the methylation levels in a particular cell type are not expected to be related to the tissue cell-type composition. Therefore, in the analysis of sorted-cell or single-cell methylation, there is no need to account for cell-type composition. In contrast, it is now widely acknowledged that in analysis of bulk methylation one has to account for cell-type composi-

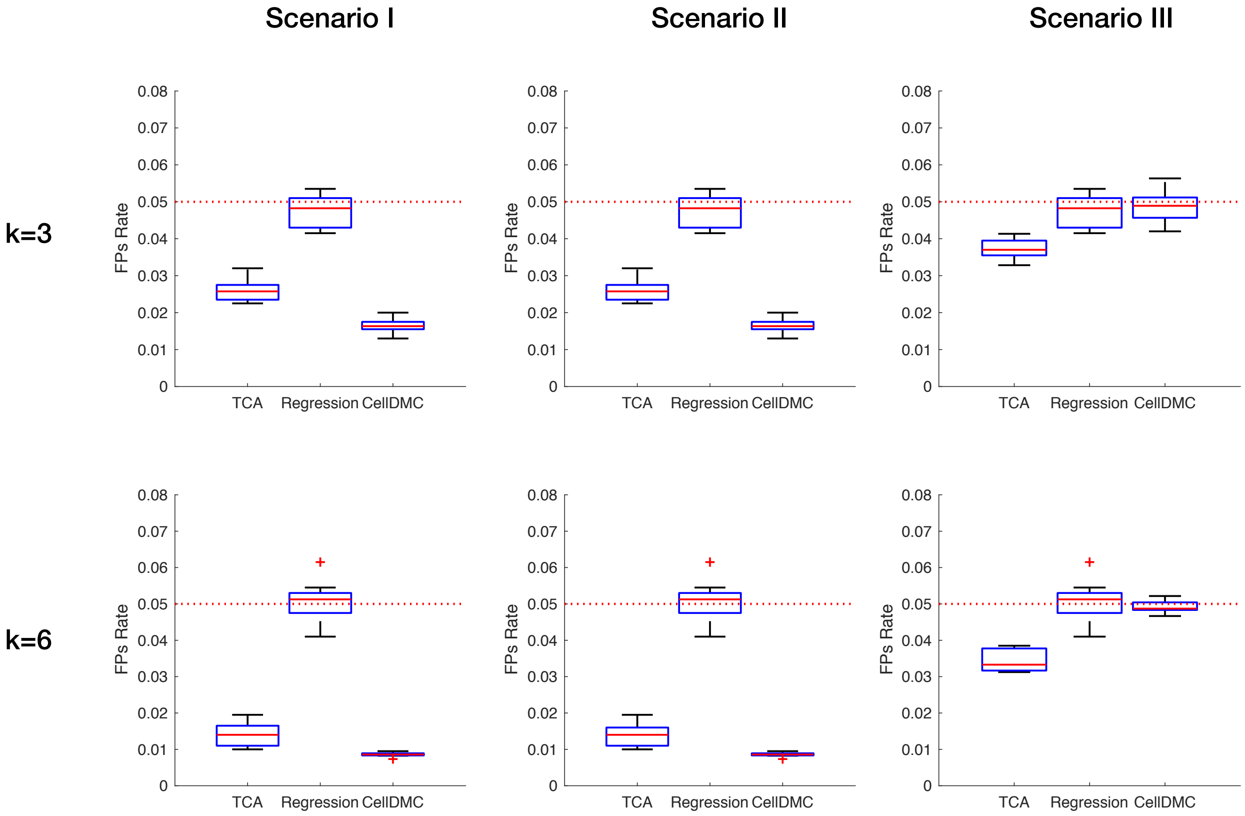

Figure 4.8: An evaluation of false positives rates in association testing with DNA methylation. Performance was evaluated using three approaches: TCA, a standard linear regression with the observed bulk data, and CellDMC with the true cell-type proportions as an input. The proportions of false positives (FPs) were measured under three scenarios using a range of effect sizes: different effect sizes for different cell types (Scenario I), the same effect size for all cell types (Scenario II), and only a single effect size for a single cell type (Scenario III); each of the scenarios was evaluated under the assumption of three constituting cell types (k=3) and six constituting cell types (k=6). Boxplots reflect results across 10 simulations. The central mark on each box indicates the median, and the bottom and top edges indicate the 25th and 75th percentiles, respectively.

tion in cases where it is correlated with the phenotype of interest [72]. For a phenotype that is highly correlated with the cell-type composition, such a correction of bulk methylation data is expected to reduce true underlying signals, potentially resulting in no findings (i.e. false negatives). As opposed to analysis of bulk data, cell-type specific analysis would not reduce the signal in this case. To demonstrate this, we consider an extreme case where the phenotype is the cell-type composition. Specifically, we defined the level of immune activity of an individual as its total lymphocyte proportion in whole-blood, and aimed at finding

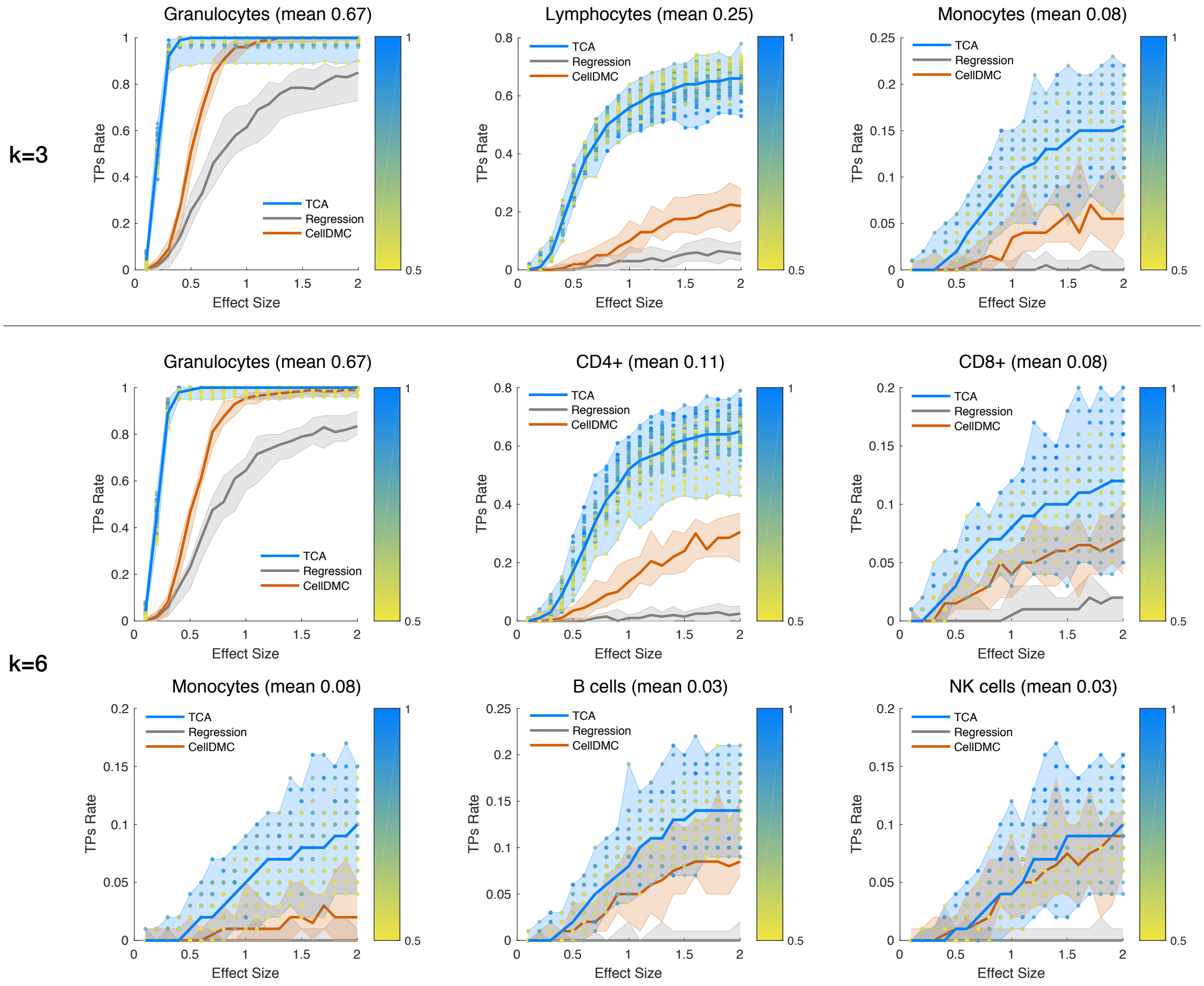

Figure 4.9: An evaluation of power for detecting cell-type-specific associations with DNA methylation, stratified by cell types (with the mean abundance of each cell type noted). Performance was evaluated using three approaches: TCA, a standard linear regression with the observed bulk data, and CellDMC with the true cell-type proportions as an input. The numbers of true positives were measured under a scenario where only a single effect size for a single cell type exists, both in the case of three constituting cell types (k=3) and six constituting cell types (k=6). The colored areas reflect the results range across 10 simulations, and the colored dots reflect the results of TCA under different initializations of the cell-type composition estimates (i.e. different levels of noise injected into TCA), where the color gradients represent the mean absolute correlation of the initial estimates with the true values (across all cell types).

methylation sites that are associated with regulation of immune activity.

Since bulk methylation data is a composition of signals that depend on to the cell-type proportions, a standard regression approach with whole-blood methylation is expected to

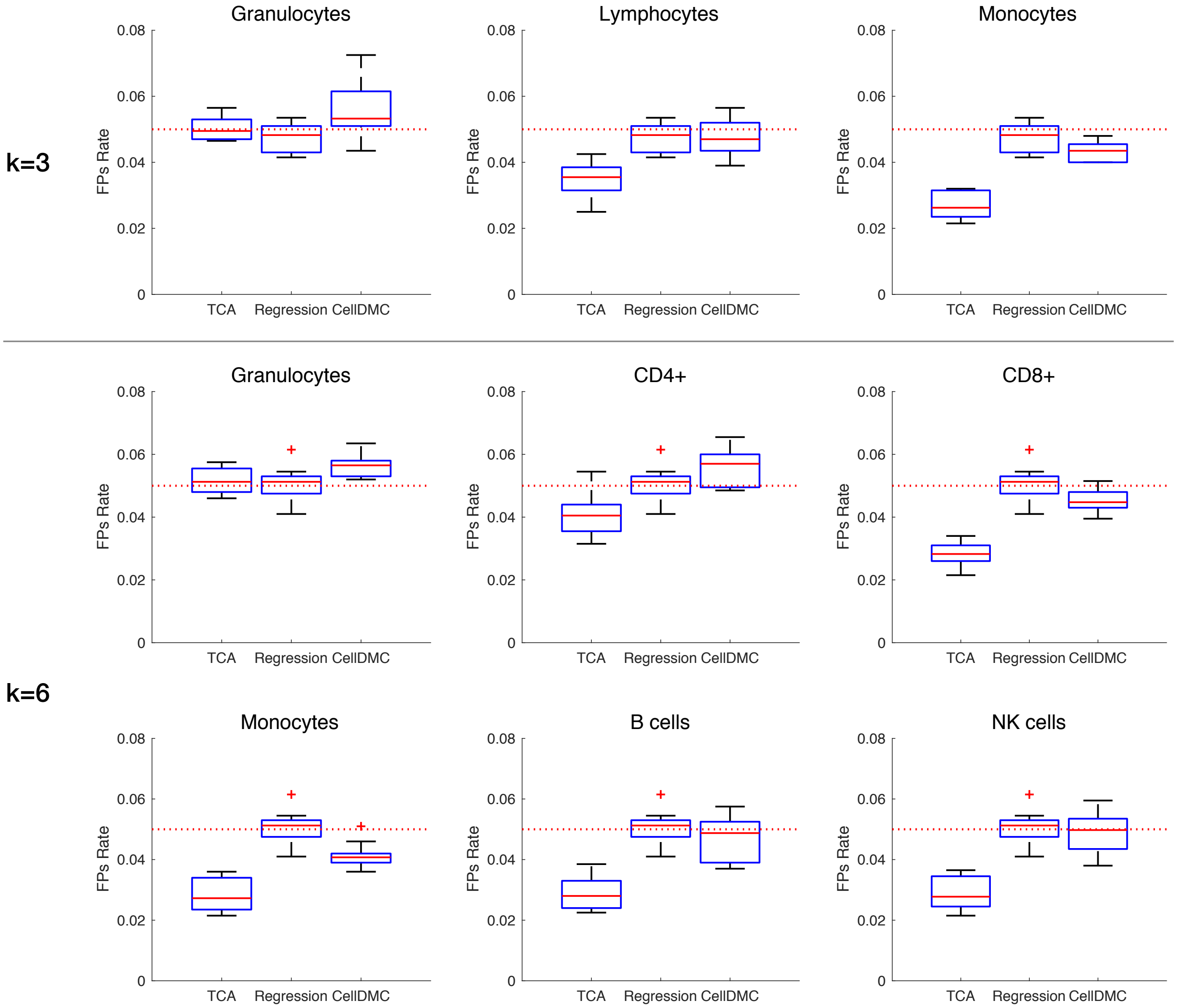

Figure 4.10: An evaluation of false positives rates in association testing with DNA methylation, stratified by cell types. Performance was evaluated using three approaches: TCA, a standard linear regression with the observed bulk data, and CellDMC with the true cell-type proportions as an input. The proportions of false positives (FPs) were measured under a scenario where only a single effect size for a single cell type exists, both in the case of three constituting cell types (k=3) and six constituting cell types (k=6). Boxplots reflect results across 10 simulations. The central mark on each box indicates the median, and the bottom and top edges indicate the 25th and 75th percentiles, respectively.

fail to distinguish between false and true associations with immune activity. We verified this using whole-blood methylation data from a previous study by Liu et al. ($n = 658$) [67]. Importantly, accounting for the cell-type composition in this case would eliminate any true signal in the data, as the immune response phenotype is perfectly defined by the cell-type composition.

We next performed cell-type-specific analysis. Applying CellDMC resulted in a massive inflation in test statistic, which failed to distinguish between false and true associations (Figure 4.11a). Using TCA, in contrast, resulted in 8 experiment-wide significant associations (p-value<9.87e-07; Figure 4.11b and Supplementary Data 1 in [140]). Importantly, 6 of the associated CpGs reside in 5 genes that were either linked in GWAS to leukocyte composition in blood or that are known to play a direct role in regulation of leukocytes: *CD247*, *CLEC2D*, *PDCD1*, *PTPRCAP*, and *DOK2* (Supplementary Data 1 in [140]). The remaining associated CpGs reside in the genes *SDF4* and *SEMA6B*, which were not previously reported as related to leukocyte composition. Using a second large whole-blood methylation data set (n=650) [94], we could replicate the associations with 4 out of the 7 genes (*PTPRCAP*, *DOK2*, *SDF4* and *SEMA6B*; p-value<0.0063; Supplementary Data 1 in [140]). Our results are therefore consistent with the possibility that methylation modifications in these genes are involved in regulation of immune activity.

4.3.3 Cell-type-specific differential methylation in rheumatoid arthritis

RA is an autoimmune chronic inflammatory disease which has been previously related to changes in DNA methylation [146, 147]. In order to further demonstrate the utility of TCA, we revisited the largest previous whole-blood methylation study with RA by Liu et al. ($n = 658$) [67].

As a first attempt to detect associations between methylation and RA status, we applied a standard regression analysis, which yielded 6 experiment-wide significant associations (p-value<2.33e-7 ;Figure 4.11c and Supplementary Data 2 in [140]), overall in line with previous studies that analyzed this data set [36, 87]. In order to allow an intuitive comparison with a standard regression, we performed a second analysis under the TCA model while assuming a single effect size in all cell types, which is expected to be a favorable scenario for a standard regression analysis. Remarkably, TCA found 15 experiment-wide significant CpGs, 11 of which were not reported by the standard regression analysis. Altogether, these

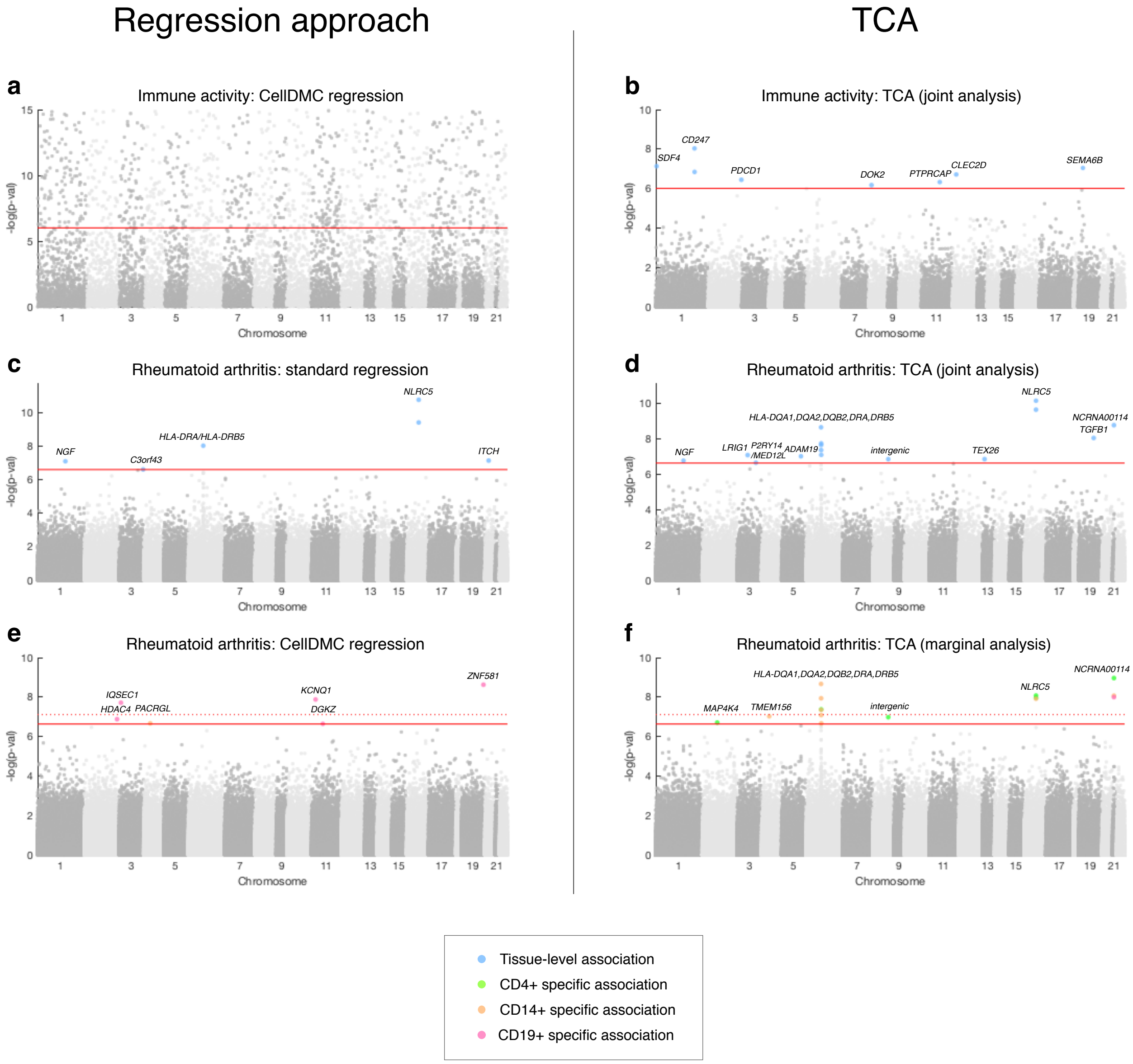

Figure 4.11: Results of the association analysis with level of immune activity and with rheumatoid arthritis in the Liu et al. whole-blood methylation data, presented by Manhattan plots of the -log10 P-values for the association tests. (a-b) Shown are results with immune activity using CellDMC (results subsampled and truncated for visualization) and using TCA. (c-d) Shown are results of the RA analysis using standard regression and using TCA under the assumption of a single effect size for all cell types. (e-f) Shown are results of a cell-type-specific analysis of RA using CellDMC and using TCA. Solid horizontal red lines represent the experiment-wide significance threshold, and dotted horizontal red lines represent the significance threshold adjusted for three experiments corresponding to the three cell types.

15 associations highlighted RA as an enriched pathway (p-value=1.45e-07; Figure 4.11d and Supplementary Data 2 in [140]).

The presumption that only some particular immune cell-types are related to the pathogenesis of RA, have led to studies with methylation collected from sorted populations of leukocytes (e.g., [137, 138, 148]). In a recent study by Rhead et al., some of us investigated differences in methylation patterns between RA cases and controls using data collected from sorted cells [138]. Particularly, methylation levels were collected from two sub-populations of CD4+ T cells (memory cells and naive cells; n=90, n=88), CD14+ monocytes (n=90), and CD19+ B cells (n=87). Although this study involved a considerable data collection effort in an attempt to provide insights into the methylome of RA patients at a cell-type-specific resolution, it does not allow the detection of experiment-wide significant associations, possibly owing to the limited sample size.

In order to reconcile with the sample size limitation in the sorted data by Rhead et al., we considered it for validation of the results reported by TCA in the large whole-blood data rather than for detecting novel associations. We found that 11 of the 15 CpGs reported by TCA (and 4 of the 6 CpGs reported by a standard regression) had a significant p-value at level 0.05 in at least one of the cell types, reflecting a high consistency with the results reported by TCA.

We next used TCA to test for associations in each of CD4+, CD14+, and CD19+ cells separately (i.e. a marginal test for each cell type, without the restriction of a single effect size). This analysis reported 15 cell-type-specific associations with 11 CpGs: 6 associations in CD4+, 8 in CD14+, and one association in CD19+ cells (p-value<2.33e-07; Figure 4.11f and Supplementary Data 2 in [140]). Considering a more stringent significance threshold in order to account for the three separate experiments we conducted for the three cell types resulted in 10 cell-type-specific associations with 7 CpGs (p-value<7.78e-08; Figure 4.11f). We further found these CpGs to be enriched for involvement in the RA pathway (p-value=9.47e-07); particularly, 4 of these CpGs reside in HLA genes (or in an intergenic HLA region) that were previously reported in GWAS as RA genetic risk loci: *HLA-DRA*, *DRB5*, *DQA1*, and *DQA2*.

We further sought to evaluate the 15 associations found by the TCA marginal test using sorted data. We found that in the Rhead et al. data 4 of the 6 associations in CD4+ and 4 of the 8 associations in CD14+ had a significant p-value at level 0.05, with all associations having overall low p-values (p-value $\leq$ 0.35 for all 15 associations; Supplementary Data 2 in [140]). Following the enrichment in small p-values, considering a false discovery rate (FDR) criterion for the entire set of 15 associations revealed significant q-values at level 0.05 for all 15 associations. We further considered an additional data set with sorted CD4+ methylation from an RA study by Guo et al. (n=24) and found it to be consistent (p-value<0.05) with 3 of the 4 CD4+ associations that were verified in the Rhead et al. data.

Notably, applying CellDMC as an alternative approach for detecting cell-type-specific associations in CD4+, CD14+, and CD19+ resulted in 6 genome-wide significant hits: one in CD14+ and five in CD19+ (and only three hits in CD19+ if accounting for the three separate experiments; Figure 4.11e). However, none of these 6 hits were found to be significant in the sorted cells data by Rhead et al. (p-value>0.05), thus, echoing our conclusions from the power simulation showing a substantial gap in power between TCA and CellDMC.

Finally, we note that the lack of evidence (from the sorted cells data) for some of the associated CpGs may be explained in part by the fact that each data set was collected from a different population; specifically, Liu et al. studied a Swedish population, Rhead et al. studied a heterogeneous European population, and Guo et al. studied a Han Chinese population. In the case of TCA, another possibility is that it did not attribute the correct cell types to some of the associations. A support for this possibility is given by the fact that two associations (cg16411857 and cg22812614) were attributed to CD4+, however were supported by the sorted data to be CD14+ specific, and another association (cg11767757) was attributed to all cell types, however, was only supported by the sorted data to be CD14+ specific.

4.4 Discussion

We proposed a methodology that can reveal novel cell-type-specific associations from bulk methylation data, i.e., without the need to collect cost prohibitive cell-type-specific data. This methodology is particularly useful in light of the large number of bulk samples that have been collected by now, and due to the fact that currently single-cell methylation technologies are not practically scalable to large population studies. Importantly, we found that TCA is substantially superior to a standard regression analysis with interaction terms between the cell-type proportions and the phenotype, while adequately controlling for false positive rate, even in the case where all cell types share the same effect size. We therefore suggest that TCA should always be preferred in analysis of bulk methylation data.

Notably, a recent attempt to provide cell-type-specific context in genetic studies aims at identifying trait-relevant tissues or cell types by leveraging genetic data and known tissue or cell-type-specific functional annotations [149, 150]. This approach yielded some promising results in relating trait-associated genetic loci to relevant tissues and cell types. However, it is limited to only one particular task and it is bounded by design to consider only genetic signals, whereas non-genetic signals are often also of interest in genomic studies. Moreover, this approach can only suggest an implicit cell-type-specific context by binding known annotations with heritability. In contrast, the approach taken in TCA allows the extraction of explicit cell-type-specific signals, which can potentially allow many opportunities and applications in biological research. We further note that around the time of submitting this work, another model similar to TCA appeared as a preprint by Luo et al. [145]. For completeness, we verified that TCA performs substantially better than the method by Luo et al.; given that the latter was not published at the time of developing TCA, we did not include this evaluation in this work.

A potential limitation of TCA is the need for rarely available cell-type proportions as an input. We alleviate this issue by allowing TCA to get estimates of the cell-type proportions using standard methods [85, 134] and then re-estimating them following the TCA model. As we showed, this allows TCA to provide good results even when just noisy estimates of the cell-type proportions are available. In practice, obtaining such estimates can be done using either a reference-based approach [85] or a semi-supervised approach [134], in case a methylation reference is not available for the studied tissue.

Our experiments and mathematical results show that TCA can extract cell-type-specific signals from abundant cell types better compared with lowly abundant cell types. Another potential limitation is expected to be in the case where the proportion of one cell type strongly covary with the proportion of a second cell type. In case of a true association in just one of the two cell types, performing a marginal association test on each cell type separately might fail to effectively distinguish between the signals of the two cell types and report an association in both cell types. In light of these limitations, we suggest that future studies include small replication data sets from sorted or single cells. Future work might be able to alleviate this issue by modeling the covariance of the cell-type proportions.

We further note that around the time of developing TCA, another model similar to TCA appeared as a preprint by Luo et al. [145] (HIRE). For completeness, we verified that TCA performs substantially better than the method by Luo et al. (data not shown); given that the latter was not published by the time of finalizing the work on TCA, we separated this evaluation from the our benchmarking.

Finally, in this work we focus on the application of TCA to epigenetic association studies. However, TCA can be formulated as a general statistical framework for obtaining underlying three-dimensional information from two-dimensional convolved signals, a capability which can benefit various domains in biology and beyond.

CHAPTER 5

Tensor Composition Analysis: relation to other models and further analysis

5.1 Background

Genomic markers are known to demonstrate differences between different cell types. Yet, differential analysis in genomics is typically performed using tissue-level bulk data, owing to high costs and practical limitations with generating large-scale data at cell-type-specific resolution. This has led to the development of computational methods that aim at performing differential analysis experiments at the cell-type level using tissue-level data.

A first attempt in that direction suggested to estimate differential expression at a cell-type level by solving a separate decomposition problem for each of two populations of interest (e.g., cases and controls for a particular condition) [131]. A later work, which, as we show here, essentially further generalized the two-way decomposition approach, allowed to consider non-categorical phenotypes by using a standard linear regression framework with interaction terms between the cell-type proportions of the samples in the data and a phenotype and interest [132]. More recently, the same interaction model was suggested and applied independently by several groups for calling differential DNA methylation at the cell-type level [133, 151, 152].

A different approach that was suggested here in Chapter 4 - Tensor Composition Analysis (TCA) - models cell-type-specific variation as individual-specific, thus assuming that the two-dimensional input data (individuals by CpGs in the case of methylation) is coming from a three-dimensional structure (individuals by CpGs by cell types). Notably, this perspective was implicitly suggested independently by another group as well at the same time [153], although, as we later discuss, they did not consider the entire span of aspects that were presented as part of the TCA framework.

Below, we go over the technical details of the decomposition approach for detecting differential methylation at the cell-type level. Then, we describe the TCA approach and relate it mathematically to the more standard decomposition and interaction models, as well as provide a theoretical justification for TCA through asymptotic analysis. Finally, we further evaluate and discuss different alternatives for statistical testing and setting model directionality that are possible within the TCA framework, while providing a thorough comparison with the interaction model.

5.2 Methods

5.2.1 A two-way decomposition for differential analysis with binary phenotypes

Let $X \in \mathbb{R}^{n \times m}$ be tissue-level bulk data collected from m features in n individuals. A standard decomposition problem assumes the following model:

$$X = W^T Z + E \tag{5.1}$$

Here, $W \in \mathbb{R}^{k \times n}$ is a matrix containing for each individual their fractions of each of k different cell types assumed to compose the tissue from which X was collected, $Z \in \mathbb{R}^{k \times m}$ is a matrix with cell-type-specific signatures for each of the k cell types in each of the m features, and $E \in \mathbb{R}^{n \times m}$ represents noise.

Many different versions that differ in their assumptions on the components of the decomposition (i.e. W, Z, E) have been proposed, and numerous applications in genomics and beyond have successfully employed a decomposition approach. Critically, estimating W, Z jointly under the model in (5.1) (unsupervised decomposition) may not be identifiable, even when introducing certain constraints on the solution of the decomposition (such as non-negativity constraints and requiring the cell-type fractions of each individual to sum up to 1) [134]. To circumvent this, one can learn and fix either Z or W by using external reference data collected from purified cells for the former or, alternatively, measuring cell counts for the individuals in the data for the latter; this supervised decomposition approach allows to avoid the non-identifiability issue [85, 134].

Solving a decomposition problem provides us with cell-type-specific signatures that are shared across all individuals in the data; thus, essentially assuming that all individuals have the exact same values at the cell-type level. This assumption, which is known to be unjustified biologically, is inherent in the classical decomposition formulation. Consequently, this model does not allow us to interrogate variance at the cell-type level, let alone to perform differential expression at the cell-type level. The first relaxation of this assumption in genomics was proposed by Shen-Orr et al. in the context of differential expression analysis [131], where the authors considered the following model for tissue-level expression data:

$$x_{ij} = \sum_{h=1}^{k} w_{hi} z_{hj}^{g} + \epsilon_{ij} \tag{5.2}$$
$$\epsilon_{ij} \sim N(0, \sigma_g^2)$$

Here, x_{ij} corresponds to the expression level of gene j in individual i (row i, column j in X), w_{hi} corresponds to the fraction of cell type h in individual i (row h, column i in W), and z_{hj} corresponds to the expression of gene j in cell-type h (row h, column j in Z). The parameter $g \in \{0, 1\}$ represents two possible values for Z, one for each of two groups of individuals (e.g., cases and controls) and ϵ_{ij} is assumed to be normally distributed with a possibly different variance for each group g.

Shen-Orr et al. experimentally measured W and solved (5.2), which essentially corresponds to solving a supervised version of the decomposition model in (5.1) twice, once for each group g, while fixing W with the measured cell counts. Provided with estimates for $\{z_{hj}^{g}\}_{h,j,g}$, the authors statistically tested for differences between z_{hj}^{0} and z_{hj}^{1} for each h, j, which allowed to interrogate differential expression of gene j in cell type h across the two groups of individuals.

5.2.2 Regression with interaction terms as a generalized decomposition for quantitative phenotypes

The two-way decomposition approach can in principle be extended to a multi-way decomposition, where a separate decomposition model is fitted for each possible value of a categorical phenotype of interest. However, is it not immediately clear how this approach can be extended to quantitative phenotypes. As we next show, this can be addressed within a regression framework.

Westra et al. [132] employed a regression model with interaction terms in the context of differential expression as follows. Let y be an n-length vector with phenotypic values for the same n individuals as in the data matrix X, keeping the previous notations, the authors considered the following model for tissue-level expression:

$$x_{ij} = \sum_{h=1}^{k} w_{hi}\mu_{hj} + \sum_{h=1}^{k} w_{hi} y_i \gamma_{hj} + \epsilon_{ij} \tag{5.3}$$
$$\epsilon_{ij} \sim N(0, \sigma^2)$$

where μ_{hj} is the cell-type-specific signature of gene j in cell type h, y_i is the phenotypic value of individual i and γ_{hj} is the effect size - specific to gene j - of the interaction (i.e. multiplicative) term between y and the proportion of cell type h.

For a binary phenotype y, it is easy to see that the model in (5.3) can be rephrased as a two-way decomposition model as in (5.2):

$$\begin{aligned}\sum_{h=1}^{k} w_{hi}\mu_{hj} + \sum_{h=1}^{k} w_{hi} y_i \gamma_{hj} + \epsilon_{ij} &= \sum_{h=1}^{k} w_{hi}\tilde{z}_{hj}\mathbb{I}\{y_i = 0\} + \sum_{h=1}^{k} w_{hi} z'_{hj}\mathbb{I}\{y_i = 1\} + \epsilon_{ij} && (5.4)\\ &\equiv \sum_{h=1}^{k} w_{hi} z^{g}_{hj} + \epsilon_{ij} && (5.5)\end{aligned}$$

where g in (5.5) represents the possible values for z_{hj}, corresponding to the two values of the phenotype y. Of note, the component of variation ϵ_{ij} can be treated separately for each group in y.

The above illustrates the motivation for performing cell-type-specific differential analysis by testing the coefficients $\{\gamma_{hj}\}$ in (5.3) as an evidence for different effects for different values of y. Clearly, rephrasing the decomposition formulation as a regression problem with interactions not only allows us to consider non-categorical phenotypes, but it also allows a straightforward model fitting and statistical analysis by leveraging the well-established tools from standard regression analysis. Particularly, covariates can be readily included in the model and tested as in standard regression:

$$x_{ij} = \sum_{h=1}^{k} w_{hi}\mu_{hj} + \sum_{h=1}^{k} w_{hi}y_i\gamma_{hj} + \sum_{d=1}^{p} c_{id}\delta_{jd} + \epsilon_{ij} \tag{5.6}$$

Here, c_{id}, δ_{jd} are the value of the d-th covariate (out of a total of p covariates) of individual i and its corresponding effect size, which is specific to feature j.

The interaction model in (5.6) was recently presented and applied by several groups for calling differential DNA methylation at cell-type resolution [133, 151, 152]. Most notably, it was used in CellDMC by Zheng et al. [133].

5.2.3 TCA: a deconvolution-based solution

While the interaction model in (5.6) improves upon the classical decomposition model in (5.1) by allowing differences in cell-type-specific profiles between individuals that are coming from different groups, it is still limited by the assumption that cell-type-specific levels are constant across individuals when conditioning on the phenotype of interest (whether it is categorical or continuous). Put differently, if we consider a case/control scenario as an illustrative example, the interaction model assumes that all cases (and similarly for controls) share the exact same cell-type-specific levels. The inter-group variation under this model stems only from fixed effects, while ignoring possible effects of additional factors at the cell-type level as well as intrinsic variability per individual (i.e. individual-specific unexplained biological noise).

TCA models the assumption that different individuals may demonstrate differences in cell-type-specific profiles owing to multiple factors affecting at the cell-type level and owing to individual-specific intrinsic variability [140]. In more details, the TCA model for methylation at the cell-type level makes the following assumption:

$$Z^i_{hj} = \mu_{hj} + \sum_{d=1}^{p_1} c^{(1)}_{id} \gamma^j_{hd} + \epsilon^i_{hj} \tag{5.7}$$
$$\epsilon^i_{hj} \sim N(0, \sigma^2_{hj})$$

where Z^i_{hj} is a random variable that represents the level in methylation site j and cell type h for the i-th individual, μ_{hj}, σ_{hj} are the population-level mean value and standard deviation for site j and cell type h, and $c^{(1)}_{id}, \gamma^j_{hd}$ are the value of the d-th covariate (out of a total of p_1 covariates) of individual i and its corresponding effect size that is specific to methylation site j and cell type h. Note that (5.7) considers p_1 covariates that are assumed to affect methylation at the cell-type level.

The TCA model further assumes the following for the observed tissue-level data:

$$X_{ij} = \sum_{h=1}^{k} w_{hi} Z^i_{hj} + \sum_{d=1}^{p_2} c^{(2)}_{id} \delta_{jd} + \epsilon_{ij} \tag{5.8}$$
$$\epsilon_{ij} \sim N(0, \tau^2)$$

where ϵ_{ij} is an i.i.d. component of variation and $c^{(2)}_{id}, \delta_{jd}$ are the value of the d-th covariate (out of a total of p_2 covariates) of individual i and its corresponding effect size that is specific to site j. Note that (5.8) considers p_2 covariates that are assumed to affect the tissue-level methylation mixtures (i.e. rather than the methylation at the cell-type level; such covariates can be, for example, batch information, that may affect the mixture regardless of the cell composition and the underlying cell-type-specific signals).

The above model reflects the assumption that the two-dimensional input data (X; individuals by sites) is coming from a three-dimensional underlying structure ($\{z^i_{hj}\}_{h,j,i}$; individuals by sites by cell types). The TCA framework allows for learning the cell-type-specific levels of an individual from their tissue-level data, thus performing a deconvolution of the observed signals in X into their underlying signals.

We make a distinction between a deconvolution that aims at obtaining the complete three-dimensional underlying signals and a decomposition that aims at obtaining population-level properties from the data, as in (5.1) and (5.6). Performing a deconvolution in this case becomes possible by the fact that both the data points and the entries of the underlying three-dimensional tensor are treated as random variables. Consequently, we can look at the conditional distribution of the tensor given the observed data and use it for inferring the tensor values of the observed data. Particularly, the TCA estimator is defined as:

$$\hat{z}_{hj}^{i} = \mathrm{E}[Z_{hj}^{i}|X_{ij} = x_{ij}; \Theta] \tag{5.9}$$

where Θ represents the parameters in (5.7) and in (5.8) - these are unknown, however, they can be estimated from the data [140]. Particularly, W is assumed to be known in a typical application of TCA, however, we developed an alternative optimization procedure for re-estimating W from the data in the case where only low-quality estimates of the cell-type proportions are available [140].

Notably, even though the model in (5.6), as presented in CellDMC, does not make a distinction between cell-type-specific covariates and mixture-level covariates, it can allow so by considering interaction terms between the cell-type proportions and the covariates that are assumed to have cell-type-specific effects. Yet, TCA further handles individual-specific intrinsic variability (i.e. the component ϵ_{hj}^{i} in (5.7)), which is not modeled by CellDMC.

In parallel to the introduction of TCA, another group presented essentially the same model as in (5.7)-(5.8) [153] and applied it for calling cell-type-specific differential methylation by treating the phenotype of interest as a cell-type-specific covariate. The TCA framework, however, is more general: it allows both to treat the phenotype as a covariate (i.e. as an explaining variable of methylation; see next subsection) and, as we later discuss, to directly model the phenotype (i.e. as the explained variable, while considering methylation levels as explaining variables). In addition, as discussed above, TCA allows to explicitly estimate cell-type-specific methylation profiles for each individual.

5.2.4 Relating the TCA model to the interaction model

In the chapter introducing TCA (Chapter 4), we primarily focused on directly modeling a phenotype of interest as affected by cell-type-specific methylation, while making the assumptions in (5.7)-(5.8) for the methylation levels (see further details in the next subsection). However, since the model in (5.7)-(5.8) can take into account covariates that affect methylation levels (or mediating components thereof) at the cell-type level, the phenotype of interest can also be treated as a cell-type-specific covariate.

Let y be a phenotype of interest that may affect methylation at the cell-type level, assuming no other factors affect methylation for simplicity, the TCA model can be formulated as follows:

$$Z_{hj}^i = \mu_{hj} + y_i \gamma_h^j + \epsilon_{hj}^i, \epsilon_{hj}^i \sim N(0, \sigma_{hj}^2) \tag{5.10}$$

$$X_{ij} = \sum_{h=1}^{k} w_{hi} Z_{hj}^i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \tau^2) \tag{5.11}$$

Disregarding the component of intrinsic variation at the cell-type level (i.e. ϵ_{hj}^i) by assuming $\sigma_{hj} = 0$, we get:

$$z_{hj}^i = \mu_{hj} + y_i \gamma_h^j \tag{5.12}$$

where z_{hj}^i is now a constant value conditional on y_i, γ_h^j (and therefore we make a distinction from the notation Z_{hj}^i, which represents a random variable with non-zero variance). Under this assumption, the model of X can be summarized as follows:

$$\begin{aligned} x_{ij} &= \sum_{h=1}^{k} w_{hi} z_{hi} + \epsilon_{ij} && (5.13) \\ &= \sum_{h=1}^{k} w_{hi} \mu_{hj} + \sum_{h=1}^{k} w_{hi} y_i \gamma_h^j + \epsilon_{ij} && (5.14) \end{aligned}$$

where $\epsilon_{ij} \sim N(0, \tau^2)$, and where, as before, we make a distinction between x_{ij} here and X_{ij}

in (5.11) (where cell-type-specific components are non-trivial random variables). This model is exactly the model in (5.3), which is the one used in CellDMC.

We conclude that assuming no intrinsic variability at the cell-type level (i.e. setting $\sigma_{hj} = 0$) yields TCA as equivalent to the CellDMC model, which reveals CellDMC as a degenerate case of TCA. Unlike CellDMC, the generality of TCA allows it to absorb and account for intrinsic variability at the cell-type level, and for that reason, in principle, TCA is expected to perform better than CellDMC in cases where intrinsic cell-type variation exists and equally well as CellDMC in cases where intrinsic cell-type variation does not exist or is minimal.

Importantly, the above result is specific to the case where methylation is assumed to be affected by the phenotype. As we discuss next, the TCA framework also allows to accommodate the assumption that the phenotype of interest if statistically affected by methylation (either directly or by a mediating component). In that case, CellDMC cannot be clearly related to TCA, as it does not allow an analogous modeling of the phenotype.

5.2.5 Changing the model directionality: modeling the phenotype within the TCA framework

All the models we discussed so far aim at explaining methylation, and differential methylation analysis is made possible within these models by testing whether a phenotype of interest statistically affect the methylation levels under test. Essentially, this is done by including the phenotype as a covariate in the model of the methylation levels. In practice, however, the directionality of the true underlying model may be different. That is, the phenotype of interest may be affected by methylation levels (or by a component that is statistically captured by methylation). In those cases, the models discussed above are statistically unjustified and, as we later show, empirically lead to worse performance as a result of making an incorrect assumption. In what follows, we denote the assumption that methylation *affects* the phenotype (or a mediating component thereof) by $Y|X$, and we denote the assumption that methylation is *affected* by the phenotype (or by a mediating component thereof) inversely as $X|Y$.

The TCA framework allows to accommodate the assumption that methylation affects the phenotype by directly modeling the phenotype. Particularly, it considers the following model for testing a given methylation site j:

$$Y_i = \sum_{h=1}^{k} Z_{hj}^{i} \beta_{hj} + e_i \tag{5.15}$$
$$e_i \sim N(0, \phi^2)$$

Here, Y_i is the phenotypic level of individual i (we make a distinction from our previously defined y_i to reflect the fact that the phenotype is now a function of the random variables $\{Z_{hj}^{i}\}_{h,j}$) and β_{hj} is the effect size of the methylation level in cell type h at site j on the phenotype.

Notably, depending on the context and phenotype of interest, assuming $Y|X$ as in (5.15) may be more appropriate (and interesting) than the alternative assumption $X|Y$. As an example, consider our analysis in Chapter 4, where we applied TCA to previously studied data with rheumatoid arthritis [67]. In this analysis, three of the associated CpGs we found are highly heritable (cg13081526, cg18816397, cg13778567; more than 50% of their variances can be explained by their cis-SNPs [70]). These findings are consistent with the possibility that methylation mediates causal genetic effects on rheumatoid arthritis, which rationalizes the assumption $Y|X$.

Fitting the model in (5.15) within the TCA framework can be done in one of two ways. First, cell-type-specific methylation can be inferred following (5.9), which then allows to employ a standard regression analysis (i.e. using the tensor estimates as the explaining variables). Second, we can consider the conditional distribution of the phenotype given $\{Z_{hj}^{i}\}_{h,j,i}$, however, since these are random variables with unobserved values they need to be integrated over. We can do so by using the conditional distribution $Z_{hj}^{i}|X_{ij}$ as follows:

$$\begin{aligned} & \int_{z_1} ... \int_{z_k} Pr\left(Y_i = y_i | \{Z_{hj}^{i} = z_h\}_{h=1}^{k}\right) Pr\left(\{Z_{hj}^{i} = z_h\}_{h=1}^{k} | X_{ij} = x_{ij}\right) dz_1 ... dz_k & (5.16) \\ = \; & Pr\left(Y_i = y_i | X_{ij} = x_{ij}\right) & (5.17) \end{aligned}$$

Therefore, in this second approach, TCA fits the conditional model $Y_i|X_{ij} = x_{ij}$, which can be expressed in terms of the effect sizes $\{\beta_{hj}\}_{h,j}$, thus allowing to estimate and statistically test them [140].

Of note, taking either approach to fit the model in (5.15) still requires to obtain estimates for the parameters of the methylation model in (5.7)-(5.8) [140]. These can then be used in the first approach for estimating $\{Z^i_{hj}\}_{h,j,i}$ following (5.9) or in the second approach for fitting the distribution $Y_i|X_{ij} = x_{ij}$. Regardless of which approach is taken, under the $X|Y$ assumption, the phenotype should not be considered as a cell-type-specific covariate as before in the $X|Y$ assumption.

5.2.6 A theoretical justification for TCA over standard regression

Let $Y \in \mathbb{R}^n$ be an outcome across n observations and $X \in \mathbb{R}^n$ be an explaining variable. A standard linear regression model assumes:

$$Y_i = X_i\beta + \epsilon_i \tag{5.18}$$

$$\epsilon_i \sim (0, \sigma^2) \tag{5.19}$$

In this classical formulation X is assumed to be fixed. However, in many types of problems, it may be more appropriate to treat X as random. TCA considers the case where X is random and is coming from a mixture of distributions (corresponding to the different sources that compose the mixture; in our case with methylation, sources correspond to cell types). Under this assumption, we would want to take into account the variation coming from the different sources.

Consider the following simplified setup of the assumptions in TCA:

$$X_i = \sum_{h=1}^{k} W_{hi} Z_{hi} \tag{5.20}$$

$$Z_{hi} \sim N(0,1) \tag{5.21}$$

$$\vec{W}_i = (W_{1i}, ..., W_{ki})^T \sim F_\Theta \tag{5.22}$$

where each of the sets $\{Z_{hi}\}_{h,i}, \{\vec{W}_i\}_i$ is i.i.d. across observations. Under this setup, we can consider the following model, which allows different effects for the different sources composing X:

$$Y_i = \sum_{h=1}^{k} Z_{hi}\beta_h + \epsilon_i \tag{5.23}$$

This model is clearly richer than the regression model in (5.18), and we are interested in understanding the expected differences between applying TCA and taking a standard linear regression approach under the assumptions in (5.20)-(5.23) (i.e. using the TCA estimator versus a linear regression estimator). To further simplify our analysis, consider the special case where there is an effect only in a single source l. Put differently, we change the model of the outcome Y in (5.23) to the following:

$$Y_i = Z_{li}\beta + \epsilon_i \tag{5.24}$$

$$\epsilon_i \sim N(0, \sigma^2) \tag{5.25}$$

Below, we show an asymptotic analysis comparing the TCA estimator to a standard linear regression estimator under these settings. In our analysis, we consider the two-step approach of TCA for model fitting under the assumption $Y|X$; that is, the values $\{z_{hi}\}_{h,i}$ are first explicitly estimated and then used as the explaining variables in a standard linear regression analysis.

Denote the estimator of a standard (univariate) linear regression by $\hat{\beta}_n^{\text{REG}}$, where n denotes the number of observations. Further define

$$a_n^{\text{REG}} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \tag{5.26}$$

$$b_n^{\text{REG}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \tag{5.27}$$

We get that

$$\hat{\beta}_n^{\text{REG}} = \frac{a_n^{\text{REG}}}{b_n^{\text{REG}}} \tag{5.28}$$

We will be using the following (Slutsky's Theorem):

$$\frac{\sqrt{n}\left(a_n^{\text{REG}} - \mathrm{E}\left[a_1^{\text{REG}}\right]\right)}{b_n^{\text{REG}}} \xrightarrow{d} N\left(0, \frac{\mathrm{V}\left[a_1^{\text{REG}}\right]}{\mathrm{E}\left[b_1^{\text{REG}}\right]^2}\right) \tag{5.29}$$

This relation holds since:

$$\sqrt{n}\left(a_n^{\text{REG}} - \mathrm{E}\left[a_1^{\text{REG}}\right]\right) \xrightarrow{d} N\left(0, \mathrm{V}\left[a_1^{\text{REG}}\right]\right) \tag{5.30}$$

$$b_n^{\text{REG}} \xrightarrow{p} \mathrm{E}\left[b_1^{\text{REG}}\right] \tag{5.31}$$

where (5.30) is given by the central limit theorem and (5.31) is given by the law of large numbers.

In this case

$$\mathrm{E}\left[a_1^{\text{REG}}\right] = \beta \mathrm{E}[W_{l1}] \tag{5.32}$$

$$\mathrm{E}\left[b_1^{\text{REG}}\right] = \sum_{h=1}^{k} \mathrm{E}[W_{h1}^2] \tag{5.33}$$

$$\mathrm{V}\left[a_1^{\text{REG}}\right] = (\beta^2 + \sigma^2)\sum_{h=1}^{k} \mathrm{E}[W_{h1}^2] + \beta^2\left(\mathrm{E}[W_{l1}^2] + \mathrm{V}[W_{l1}]\right) \tag{5.34}$$

Using (5.29), we get the following asymptotic distribution for the estimator of linear regression:

$$\hat{\beta}_n^{\mathrm{REG}} \xrightarrow{d} N\left(\frac{\beta \mathrm{E}[W_{l1}]}{\sum_{h=1}^k \mathrm{E}[W_{h1}^2]}, \frac{(\beta^2+\sigma^2)\sum_{h=1}^k \mathrm{E}[W_{h1}^2] + \beta^2\left(\mathrm{E}[W_{l1}^2] + \mathrm{V}[W_{l1}]\right)}{n\sum_{h=1}^k \mathrm{E}[W_{h1}^2]}\right) \tag{5.35}$$

Next, we derive the asymptotic distribution of the TCA estimator. First, recall that the TCA estimator of Z_{li} in this case is

$$\hat{Z}_{li} = \mathrm{E}\left[Z_{li} | X_i = x_i, \vec{W}_i = (w_{1i}, ..., w_{ki})^T\right] = \frac{w_{li}x_i}{\sum_{h=1}^k w_{hi}^2} \tag{5.36}$$

Similarly to the case of the regression estimator, we define

$$a_n^{\mathrm{TCA}} = \frac{1}{n}\sum_{i=1}^n \hat{Z}_{li} Y_i \tag{5.37}$$

$$b_n^{\mathrm{TCA}} = \frac{1}{n}\sum_{i=1}^n \hat{Z}_{li}^2 \tag{5.38}$$

$$\hat{\beta}_n^{\mathrm{TCA}} = \frac{a_n^{\mathrm{TCA}}}{b_n^{\mathrm{TCA}}} \tag{5.39}$$

which gives us

$$\mathrm{E}\left[a_1^{\mathrm{TCA}}\right] = \beta \mathrm{E}\left[\frac{W_{li}^2}{\sum_{h=1}^k W_{hi}^2}\right] \tag{5.40}$$

$$\mathrm{E}\left[b_1^{\mathrm{TCA}}\right] = \mathrm{E}\left[\frac{W_{li}^2}{\sum_{h=1}^k W_{hi}^2}\right] \tag{5.41}$$

$$\mathrm{V}\left[a_1^{\mathrm{TCA}}\right] = (\beta^2+\sigma^2)\mathrm{E}\left[\frac{W_{li}^2}{\sum_{h=1}^k W_{hi}^2}\right] + \beta^2\left(\mathrm{E}\left[\frac{W_{li}^2}{\sum_{h=1}^k W_{hi}^2}\right] + 2\mathrm{V}\left[\frac{W_{li}^2}{\sum_{h=1}^k W_{hi}^2}\right]\right) \tag{5.42}$$

Using (5.29) similarly as in the regression case, we get the following asymptotic distribution for the estimator of TCA:

$$\hat{\beta}_n^{\mathrm{TCA}} \xrightarrow{d} N\left(\beta, \frac{1}{n}\left(2\beta^2+\sigma^2\right) + \frac{2\beta^2}{n}\frac{\mathrm{V}\left[\frac{W_{li}^2}{\sum_{h=1}^k W_{hi}^2}\right]}{\mathrm{E}\left[\frac{W_{li}^2}{\sum_{h=1}^k W_{hi}^2}\right]}\right) \tag{5.43}$$

Asymptotically, TCA provides an unbiased estimator for β. On the other hand, a standard linear regression estimator yields a biased estimator in case $\beta \neq 0$; particularly, it will typically underestimate β, except for cases where $\mathrm{E}[W_{li}] > \sum_{h=1}^{k} \mathrm{E}[W_{hi}^2]$, in which case it will overestimate β. These results provide an insight into the empirical differences we observe between TCA and and a standard regression approach (see Subsection 4.3.1).

5.2.7 Software and computational tools

We applied TCA using the `TCA` R package version 1.2.0 (available on CRAN); source code is available from github at `github.com/cozygene/TCA`. For the application of CellDMC, we used the `CellDMC` function in the EpiDISH R package version 2.2.0 (available on Bioconductor) which provides an implementation of CellDMC.

In the application of TCA under the assumption $X|Y$, we used the `tca` function, which provides p-values for the estimated parameters in the model under a marginal conditional test and under a joint test; these are given in the output fields `gammas_hat_pvals` and `gammas_hat_pvals.joint` of the `tca` function. In the application of TCA under the assumption $Y|X$, we used the `tcareg` function, while setting the argument `test` to the requested type of test (e.g., marginal or marginal conditional). Throughout our experiments, unless stated otherwise, we used the fast mode of the `tca` function by setting `vars.mle = FALSE` and the fast mode of the `tcareg` function by setting `fast_mode = TRUE` (see next subsection for details).

5.2.8 Fast optimization of the TCA model

We previously applied an alternating maximum-likelihood-based optimization procedure for fitting the TCA model [140]. Learning the mean parameters in the model in equations (5.7)-(5.8) given the variance parameters is a convex problem that can be solved efficiently by formulating it as a constrained regression problem, yet, estimating the variance parameters (i.e. given estimates of the means) is a non-convex problem. For that reason, we employed a gradient-based optimization for the variances. This resulted in relatively long runtimes of the function tca in the TCA R package, especially for large data such as methylation arrays that include hundreds of thousands of features.

Maximum-likelihood estimation is perhaps the most common approach for fitting statistical models, however, alternatives do exist. Particularly, the generalized method of moments (GMM) allows to estimate model parameters by defining moment conditions, which are essentially sets of equations that are constructed from the model parameters and the data [154]. Given that several assumptions on the moment conditions are met, parameter estimation with proven statistical properties can then be performed by solving efficient quadratic programming problems [154, 155].

We applied the GMM technique for a fast estimation of the variances in the TCA model. Particularly, for each methylation site, we estimated the cell-type-specific variance parameters of all cell-types jointly. In order to do so, for each site, we define a set of moment conditions, one per each individual sample in the data. Each such individual-based moment condition formulates an estimator for the variance of the methylation level of the individual in the particular site under consideration. Since the variance of each given individual is a function of both the individual-specific cell-type proportions and the variance parameters of all cell types, these moment conditions can be used to estimate the variance parameters under the GMM framework [154].

We updated the tca function in the TCA package to include an argument vars.mle, which can set tca to use either the original maximum-likelihood estimation procedure for the variances (if set to TRUE) or the alternative, GMM-based procedure (if set to FALSE).

The TCA framework further allows for statistical testing under the assumption that cell-type-specific methylation affect the phenotype of interest or a mediating component thereof (i.e. $Y|X$; see subsection 5.2.5); this assumption is implemented in the TCA package within the tcareg function. As discussed in the original TCA paper, there are two approaches to perform statistical testing under the assumption $Y|X$. First, we can use a two-step approach of obtaining estimates for the cell-type-specific levels (using the tensor function in the TCA package), which can then be associated with the phenotype under a standard regression framework. Second, we can consider a single-step approach of directly using the conditional distribution of the phenotype given the data for statistical inference and testing (see subsection 5.2.5).

Previously, we implemented only the one-step approach in `tcareg`. In order to streamline the faster statistical testing that is allowed by the two-step approach, we updated the `tcareg` function accordingly to include an argument `fast_mode`, which can set `tcareg` to use either the previously implemented one-step approach (if set to `FALSE`) or the faster two-step approach (if set to `TRUE`).

5.2.9 Simulation study

We designed our simulation study similarly to a recently suggested pipeline [156] as follows. For each data set we simulated, we generated tissue-level bulk data for a subset of the methylation sites that are available in the Reinius et al. data [74]: 1,000 sites picked at random and an additional set of 333 reference CpGs that are used in the software EpiDISH for reference-based estimation of cell-type proportions [157]. For simulating methylation levels of an individual, we first sampled cell-type-specific methylation levels for the 1,333 sites in each of six major immune cell types (CD4+, CD8+, granulocytes, monocytes, B cells, and natural killer cells) using Beta distributions that we learned from the purified methylation profiles of these cell types in the Reinius et al. data (n=6 for each cell type) [74]. Eventually, we constructed tissue-level methylation values by linearly mixing them according to cell-type proportions that we sampled from a pool of estimates we obtained by applying the reference-based method EpiDISH to the Hannum et al. data [94].

In the experiments under the assumption $X|Y$ (i.e. methylation is affected by the phenotype), unless otherwise stated, each data set we generated was consisted of 500 individuals (to reflect a typical sample size in association studies), out of which 250 were cases and 250 controls. For simulating differentially methylated cell-types, we first selected 100 sites and cell types at random (the number of cell types was determined by the specific scenario under consideration as explained later), while requiring the selected sites to exhibit either low

(<0.2) or high (>0.8) average methylation levels in the specific cell-types to be altered (the former were used for simulating hypermethylation in cases and the latter for hypomethylation in cases). Then, we altered the cell-type-specific methylation of cases in the selected sites and cell types based on the following equations:

$$\gamma = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{2}}} \tag{5.44}$$

$$\sigma_1 = \sqrt{\frac{\mu_1(1-\mu_1)}{\mu_2(1-\mu_2)}}\sigma_2 \tag{5.45}$$

Here, considering one particular differentially methylated cell type in a given methylation site, γ denotes the effect size and $\{\mu_1, \sigma_1^2\}, \{\mu_2, \sigma_2^2\}$ denote sets of the mean and variance of the methylation levels of the particular site and cell type in the cases and controls groups, respectively. Setting μ_1, σ_1 was done using the above equations given a particular effect size under consideration and given the parameters $\{\mu_2, \sigma_2\}$, which were set to the mean and variance of the beta distributions that were estimated based on the Reinius et al. data as explained above. Given $\{\mu_1, \sigma_1^2\}$, these were considered as the mean and variance of a beta distribution from which methylation levels were sampled for cases.

In the experiments under the assumption $Y|X$ (i.e. methylation affects the phenotype), unless otherwise stated, each data set we generated consisted of 500 individuals; these were similarly generated as in the $X|Y$ simulations, with the exception that methylation levels of a specific cell type in a specific site were sampled from a single beta distribution (i.e. the same distribution for all individuals, as opposed to the separate distribution for cases and controls in the $X|Y$ simulations). We randomly selected 100 differentially methylated sites and cell types as performed previously in the $X|Y$ simulations, and then simulated a phenotype for each differentially methylated site j as follows:

$$y_i^j = w_i^T \alpha_j + \sum_{h \in S_j} Z_{hj}^i \beta_{hj} + \epsilon_{ij} \tag{5.46}$$

$$\alpha_j \sim N(0,1) \tag{5.47}$$

$$\epsilon_{ij} \sim N(0, \sigma_j^2) \tag{5.48}$$

$$\sigma_j^2 = \sum_{h \in S_j} \sigma_{hj}^2 \tag{5.49}$$

where y_i^j is the phenotypic value of individual i, w_i,α_j are the individual's cell-type proportions and their corresponding effect sizes in site j, respectively, S_j is the set of all differentially methylated cell types in site j, Z_{hj}^i, β_{hj} are the cell-type-specific methylation of individual i in site j and cell type h and the corresponding effect size, respectively, and σ_{hj} is the standard deviation of cell type h in site j.

The rest of the simulated sites that are not differentially methylated were tested against one of the 100 simulated phenotypes (selected at random). Notably, Equation (5.49) allows a meaningful definition of effect sizes by accounting for changes in variability between different methylation sites and cell types. Further, β_{hj} can be either positive or negative under the above formulation, however, in our final evaluations (i.e. in the figures) we consider the absolute value of β_{hj} as the effect size.

In both the $X|Y$ and $Y|X$ experiments, we considered four scenarios: unidirectional change in one cell type, unidirectional changes in two cell types, bidirectional changes in two cell types, and bidirectional changes in three cell types. For evaluation metrics, we considered sensitivity, specificity, and precision (positive predictive value ; PPV). In our calculation of sensitivity, we did not require that the direction of a predicted association must match the direction of the true effect in order to be counted as a true positive. We disregarded this seemingly reasonable criterion for reasons that are related to the inherent multicollinearity in TCA and CellDMC (see Appendix B). Yet, we did rerun all experiments while taking this additional criterion into account in the definition of sensitivity, and we found that results are essentially the same in all experiments (and conclusions therefore remain the same throughout our analysis; data not shown), with the exception of marginal tests in the bi-directional scenarios, wherein the performance of sensitivity is greatly affected by the inclusion of this criterion.

5.2.10 Analysis of smoking status

We obtained the normalized Illumina 450k data by Liu et al. (n=689) [67] and by Hannum et al. (n=656) [94] from the Gene Expression Omnibus (GEO; accession numbers GSE42861 and GSE40279, respectively); smoking information is not available on the GEO record of the Hannum et al. data and can be obtained from the authors. We removed two samples with no smoking information from the Liu et al. data and 66 sample with no smoking information from the Hannum et al. data, and in both data sets we defined the smoking status as a categorical variable with three categories: never-smokers, ex-smokers, and current smokers (occasional smokers were considered as smokers).

In each of the two data sets, we tested the smoking status for each of the seven differentially methylated CpGS that were previously reported by Su et al. as either myeloid- or lymphoid-specific [158]. In our analysis, we accounted for rheumatoid arthritis status, gender, and age in the Liu et al. data, and for age, gender, and plate in the Hannum et al. data. In addition, in order to account for technical variation in the data, we considered a previously suggested approach by Lehne et al. for capturing technical variation in the Illumina 450k methylation arrays [97], wherein principal components (PCs) are calculated from control probes that are not expected to exhibit any biological signal. Specifically, in our case, we included in the analysis of each data set the top ten PCs calculated from a set of 1,000 sites that demonstrate the lowest variance in the data (and are therefore expected to exhibit no true biological variation). Finally, for evaluating genome-wide calibration, we tested all the methylation sites in the data for association with the smoking status, except for polymorphic probes, non-specific probes, and probes of sites that are on non-autosomal chromosomes as previously suggested [37].

Our evaluation included three methods: CellDMC, TCA under the assumption $X|Y$ using marginal conditional tests, and TCA under the assumption $X|Y$ using joint tests. All methods were executed under the assumption of two cell-types (myeloids and lymphoids). To that end, cell-type proportions of seven blood cell types were estimated using the reference-based method EpiDISH [157], and then aggregated within each of the lymphoid and myeloid compartments.

5.3 Results

5.3.1 Differential DNA methylation at a cell-type level: TCA in the context of a standard decomposition approach

Calling differential DNA methylation at a cell-type level from tissue-level bulk data has recently become a question of interest [133, 140, 151, 152, 153, 159], and thus far, two different approaches have been suggested for the task. The first approach employs a standard regression analysis while including and evaluating interaction terms (i.e. multiplicative terms) between cell-type proportions and the phenotype of interest; this approach reflects a generalization of a classical decomposition problem (see Subsection 5.2.2). The second approach, TCA, was introduced here in Chapter 4.

In Chapter 4, we focused on the application of TCA for detecting differential methylation at cell-type level under the assumption that methylation levels *affect* the phenotype of interest (or a mediating component thereof; denote this assumption by $Y|X$). However, the TCA framework can also properly accommodate the other direction - the assumption that methylation levels are *affected* by the phenotype of interest (or by a mediating component thereof; denote this assumption by $X|Y$). The latter is the assumption taken by the interaction model, and it is therefore of interest to investigate TCA and compare it with the interaction model under both assumptions.

Under the assumption $X|Y$, TCA is a more expressive model than the interaction model, as it makes more general assumptions about the variation of cell-type-specific methylation (see Subsection 5.2.4). Therefore, for large enough data (at least 60 samples as we later show), TCA is expected to be the better choice in general. To illustrate the difference between the methods, consider a case/control study design. In that case, the interaction model assumes a fixed effect between cases and controls as the only variation at the cell-type level; this corresponds to the unrealistic assumption that all individuals within a group (i.e. cases or controls) have the exact same methylome. TCA improves upon this by modeling the variation of cell-type-specific methylation across individuals (i.e. even within the same group). We show this theoretically by revealing the mathematical relation between TCA and the interaction model, which yields the latter as a degenerate case of the more general TCA model (see Subsection 5.2.4).

The interaction model cannot take the assumption $Y|X$, however, this model directionality is often of more interest than the assumption $X|Y$ (see Subsection 5.2.5). Our empirical results from Chapter 4 concerning the $Y|X$ case provide strong empirical evidence and mathematical intuition for the advance of TCA over alternative approaches (see Subsection 4.3.1). Here, we further provide asymptotic analysis, showing the theoretical merit of TCA over a standard regression model - the standard approach for addressing the assumption $Y|X$ - by revealing that TCA, unlike a linear regression analysis, allows an unbiased estimation of effect sizes (see Subsection 5.2.6).

In the next two subsections, we consider both the $X|Y$ and $Y|X$ assumptions, with a primary focus on the former (previously we only evaluated $Y|X$). We benchmark the performance of TCA and the interaction model through a set of experiments, including a thorough simulation study and analysis of methylation with smoking status using multiple methylation data sets. For the interaction model, we used CellDMC, which was recnetly suggested for the goal of detecting differential DNA methylation at the cell-type level [133].

5.3.2 Evaluation of TCA and the interaction model

In Chapter 4, we focused on two types of statistical tests in TCA: a marginal test, wherein the effect of each cell type is estimated and tested marginally (i.e. separately), irrespective of other cell types, and a joint test, wherein the effects of all cell types are estimated jointly and statistically tested for their combined effect. However, we have implemented other types of statistical tests as well, each allowing to test a different hypothesis about the relation between the phenotype of interest and cell-type-specific methylation (see Appendix B). Particularly, TCA allows to conduct marginal conditional tests, wherein the effects of all cell types are estimated jointly and then tested in each cell type. Marginal conditional tests are also the ones used by CellDMC.

We evaluated TCA and CellDMC using a simulation study, where we considered four different scenarios: unidirectional change in one cell type, unidirectional changes in two cell types, bidirectional changes in two cell types, and bidirectional changes in three cell types. Evaluating both methods under the same biological model (i.e. under the assumption $X|Y$) renders TCA as the (mildly) better performing method under all four scenarios and all three evaluation metrics that were considered (Figure 5.1). These results are not surprising given our theoretical results, which reveal CellDMC as a degenerate case of TCA (see Subsection 5.2.4). Repeating this analysis using different sample sizes shows a slight advantage for CellDMC over TCA in data with less than 60 individuals (Figure 5.2), thus providing insight into the sample size required in order to benefit from the generality of TCA over CellDMC.

Further simulating phenotypes to be statistically affected by methylation (i.e. setting $Y|X$ rather than $X|Y$ as the true model) results in a substantial decrease in specificity and precision for CellDMC (with the benefit of a mild increase in sensitivity) compared to TCA (Figure 5.3). This can be explained by the fact that TCA can properly accommodate the assumption $Y|X$, the true biological model in this case, whereas CellDMC is bound to assume $X|Y$. Notably, applying TCA under the wrong biological assumption in this case (i.e. assuming $X|Y$) performs better than CellDMC, reflecting better robustness of TCA to model misspecification (Figure 5.4).

Lastly, we evaluated TCA with marginal tests, which yielded a substantial increase in sensitivity, however, at the cost of a considerable decrease in specificity and precision (Figures 5.5 and 5.6); this result is expected given the inherent difference between marginal and marginal conditional tests (see Appendix B).

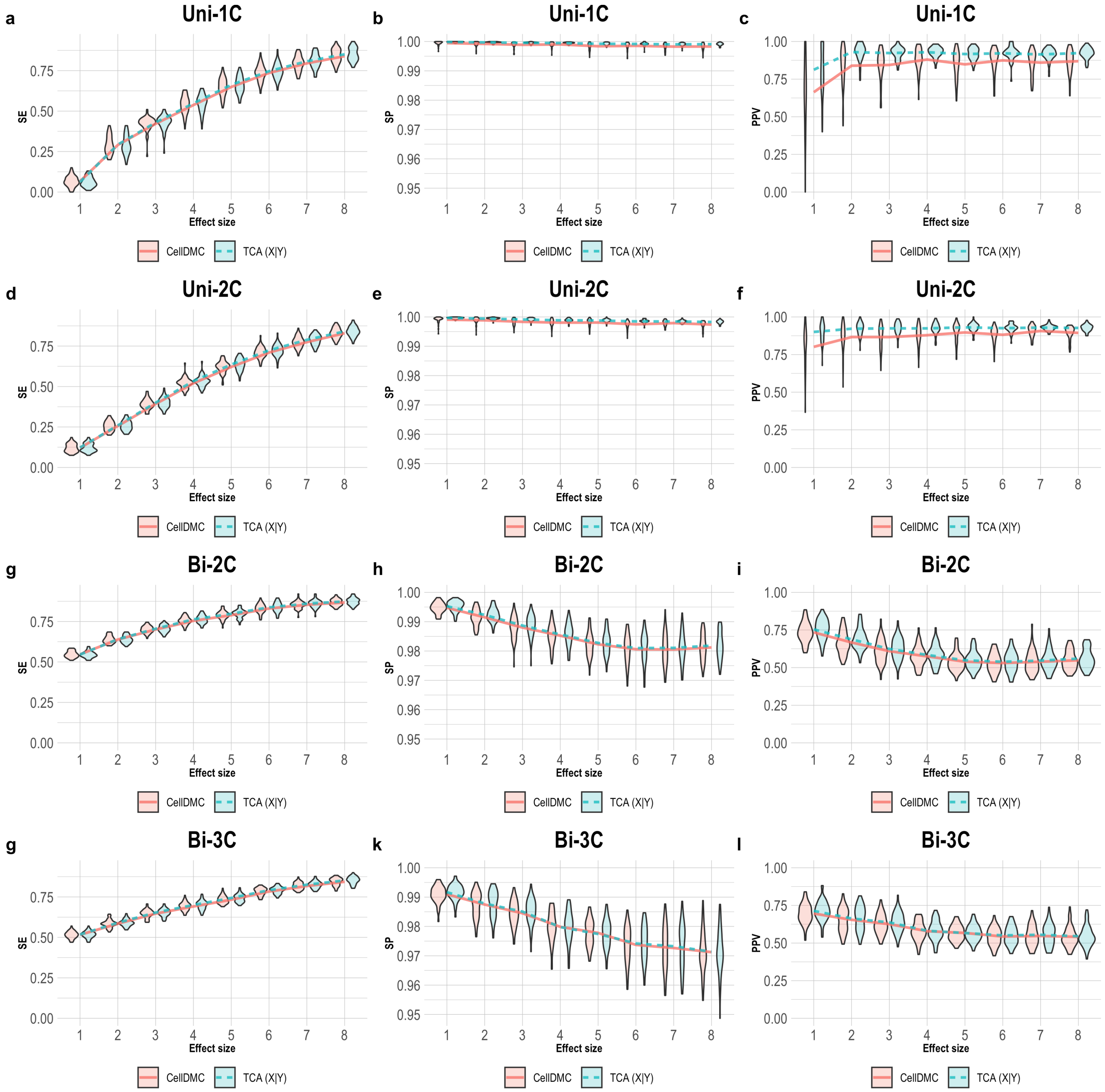

Figure 5.1: Evaluation of TCA and CellDMC in the case where the phenotype affects methylation ($X|Y$). (a)-(c) Comparison of the sensitivity (SE), specificity (SP), and precision (positive predictive value; PPV) to detect differentially methylated cell-types as a function of the association effect size, under the scenario where a single cell type out of 6 blood cell types is altered in cases versus controls (Uni-1C). (d)-(f) as in Uni-1C, only for the scenario where two cell types are altered in the same direction (Uni-2C). (g)-(i) as in Uni-2C, only for the scenario where the cell types are altered in opposite directions (Bi-2C). (j)-(l) as in Bi-2C, only for three cell types (Bi-3C). Results are shown across 50 simulated data sets using violin plots; solid lines represent median values. TCA was executed under the assumption $X|Y$ (TCA $X|Y$).

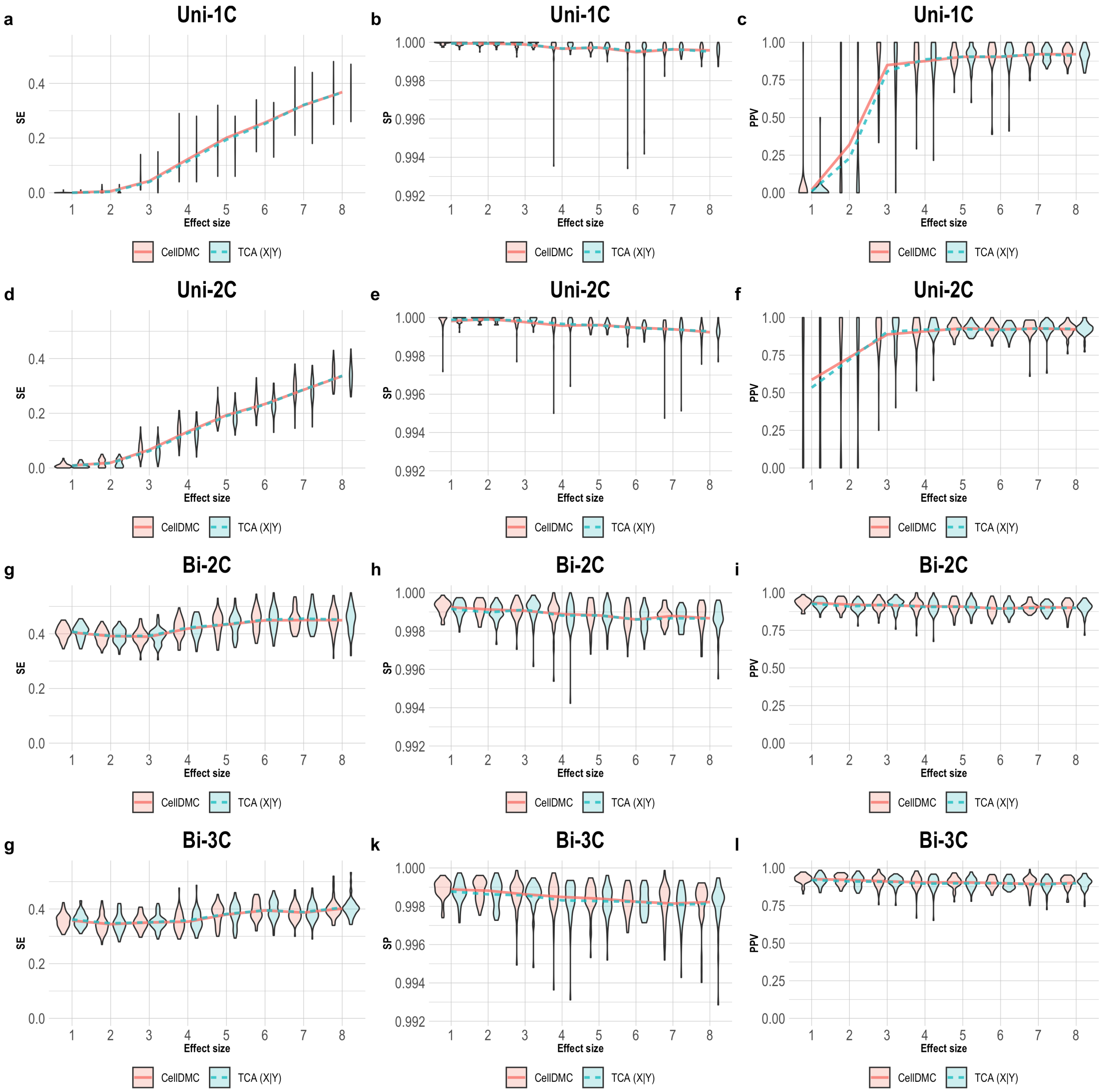

Figure 5.2: Evaluation of TCA and CellDMC in the case where the phenotype affects methylation $(X|Y)$ using small simulated data (n=60). (a)-(c) Comparison of the sensitivity (SE), specificity (SP), and precision (positive predictive value; PPV) to detect differentially methylated cell-types as a function of the association effect size, under the scenario where a single cell type out of 6 blood cell types is altered in cases versus controls (Uni-1C). (d)-(f) as in Uni-1C, only for the scenario where two cell types are altered in the same direction (Uni-2C). (g)-(i) as in Uni-2C, only for the scenario where the cell types are altered in opposite directions (Bi-2C). (j)-(l) as in Bi-2C, only for three cell types (Bi-3C). Results are shown across 50 simulated data sets using violin plots; solid lines represent median values. TCA was executed under the assumption $X|Y$ (TCA $X|Y$).

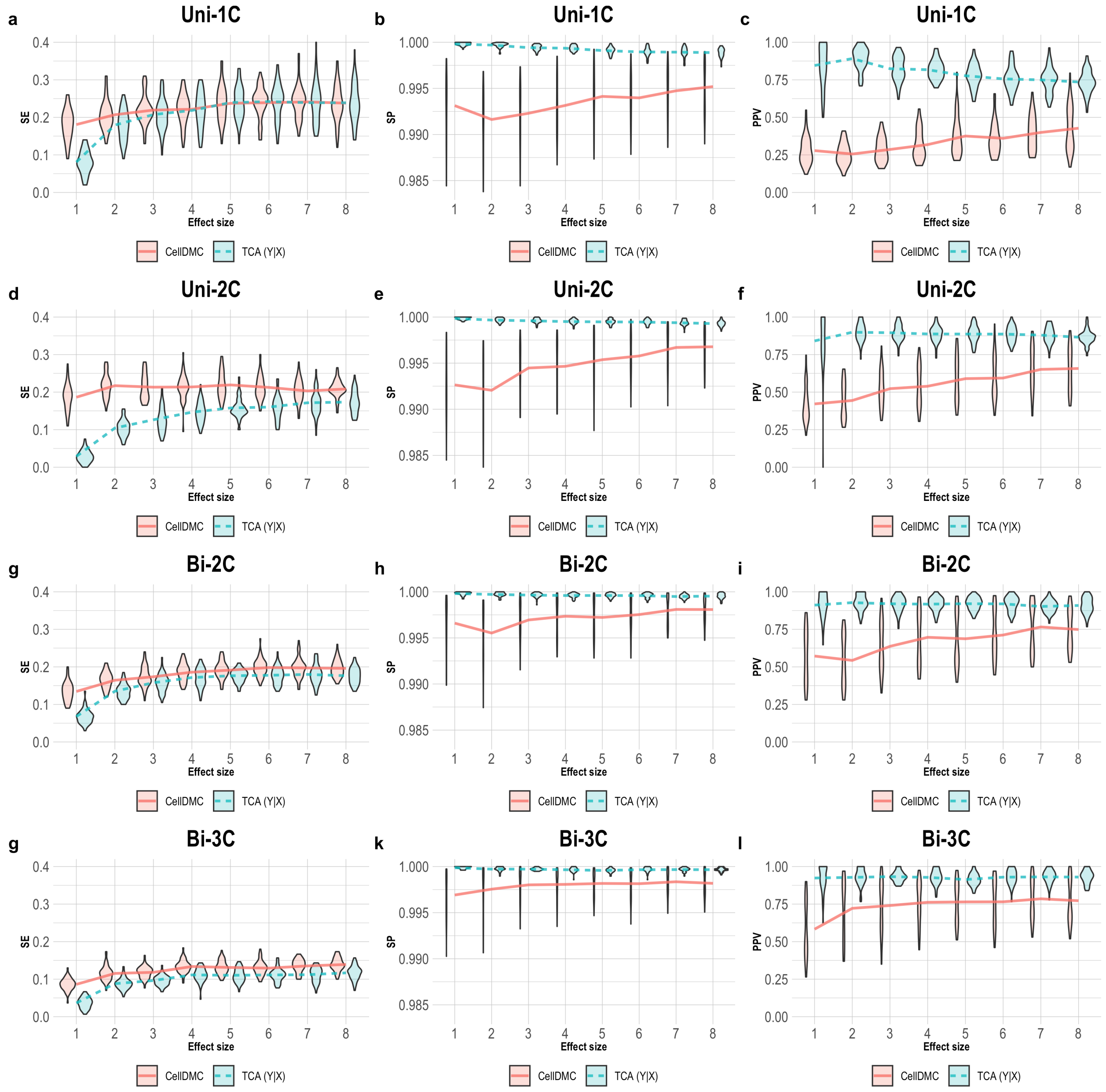

Figure 5.3: Evaluation of TCA and CellDMC in the case where the phenotype is affected by methylation ($Y|X$). (a)-(c) Comparison of the sensitivity (SE), specificity (SP), and precision (positive predictive value; PPV) to detect differentially methylated cell-types as a function of the association effect size, under the scenario where a single cell type out of 6 blood cell types is altered in cases versus controls (Uni-1C). (d)-(f) as in Uni-1C, only for the scenario where two cell types are altered in the same direction (Uni-2C). (g)-(i) as in Uni-2C, only for the scenario where the cell types are altered in opposite directions (Bi-2C). (j)-(l) as in Bi-2C, only for three cell types (Bi-3C). Results are shown across 50 simulated data sets using violin plots; solid lines represent median values. TCA was executed under the assumption $Y|X$ (TCA $Y|X$).

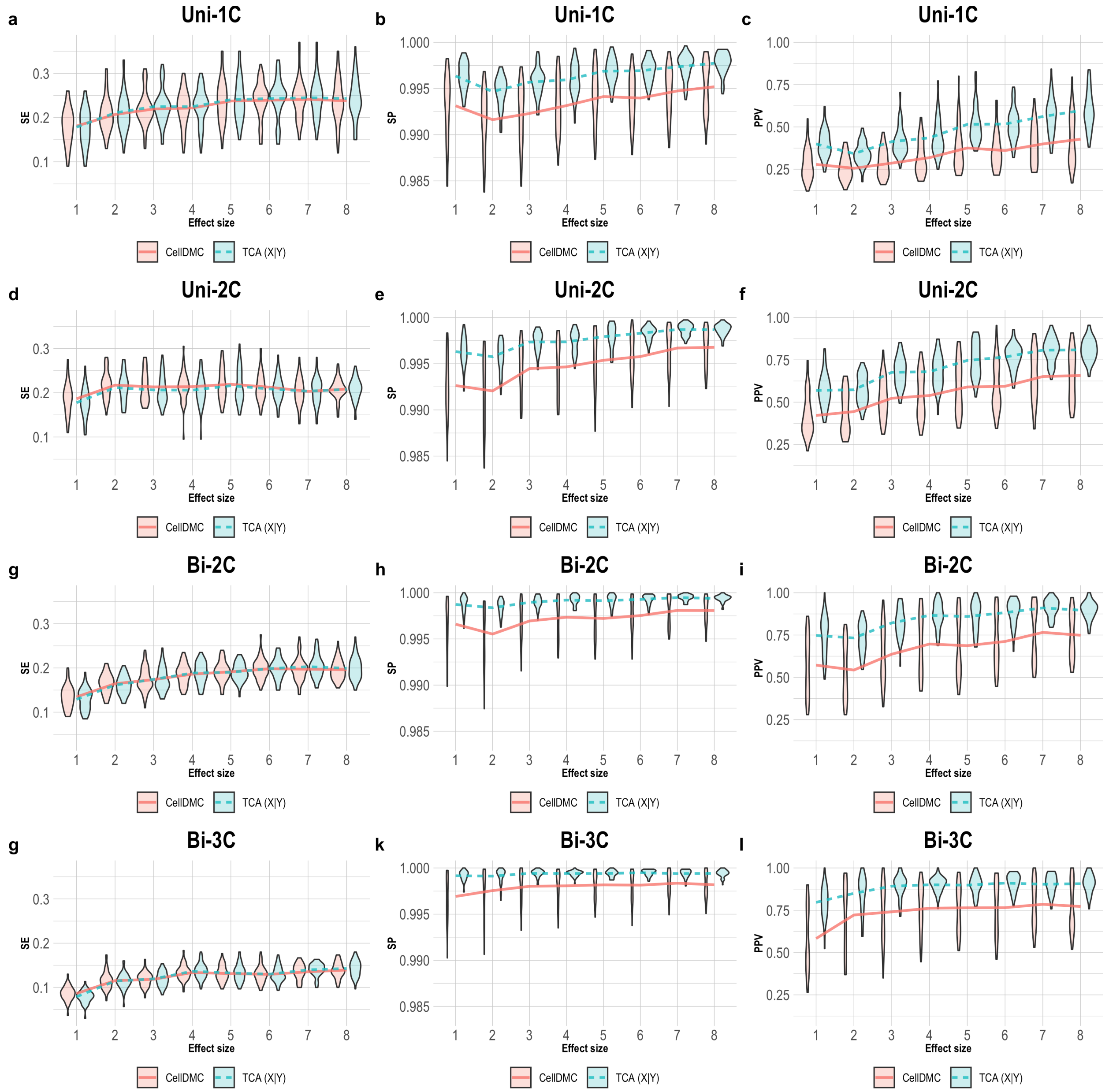

Figure 5.4: Evaluation of TCA and CellDMC in the case where the phenotype is affected by methylation $(Y|X)$, while executing TCA under the wrong assumption $X|Y$ (TCA $X|Y$). (a)-(c) Comparison of the sensitivity (SE), specificity (SP), and precision (positive predictive value; PPV) to detect differentially methylated cell-types as a function of the association effect size, under the scenario where a single cell type out of 6 blood cell types is altered in cases versus controls (Uni-1C). (d)-(f) as in Uni-1C, only for the scenario where two cell types are altered in the same direction (Uni-2C). (g)-(i) as in Uni-2C, only for the scenario where the cell types are altered in opposite directions (Bi-2C). (j)-(l) as in Bi-2C, only for three cell types (Bi-3C). Results are shown across 50 simulated data sets using violin plots; solid lines represent median values.

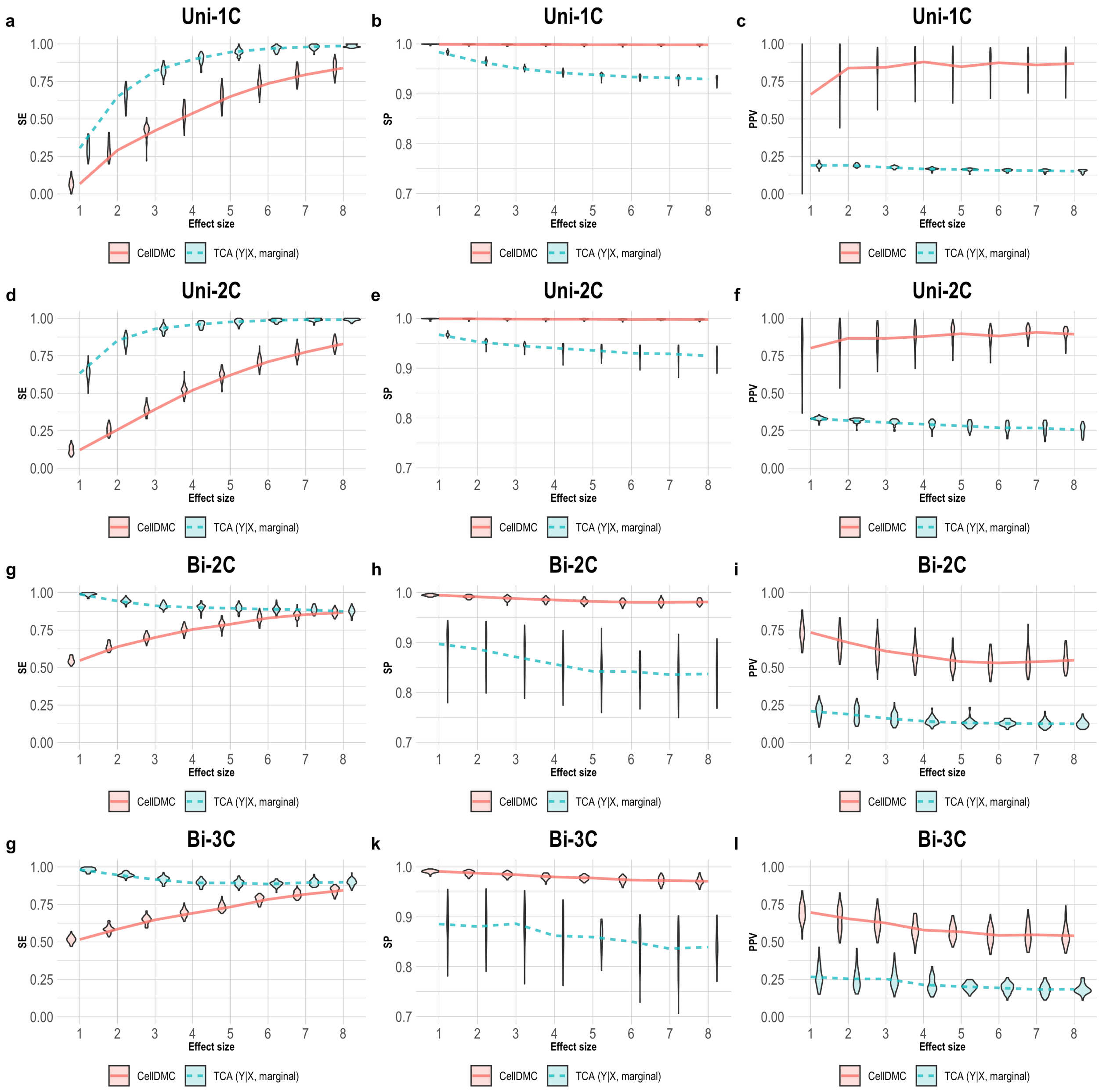

Figure 5.5: Evaluation of TCA and CellDMC in the case where the phenotype affects methylation $(X|Y)$, while executing TCA under the assumption $Y|X$ and using a marginal test. (a)-(c) Comparison of the sensitivity (SE), specificity (SP), and precision (positive predictive value; PPV) to detect differentially methylated cell-types as a function of the association effect size, under the scenario where a single cell type out of 6 blood cell types is altered in cases versus controls (Uni-1C). (d)-(f) as in Uni-1C, only for the scenario where two cell types are altered in the same direction (Uni-2C). (g)-(i) as in Uni-2C, only for the scenario where the cell types are altered in opposite directions (Bi-2C). (j)-(l) as in Bi-2C, only for three cell types (Bi-3C). Results are shown across 50 simulated data sets using violin plots; solid lines represent median values.

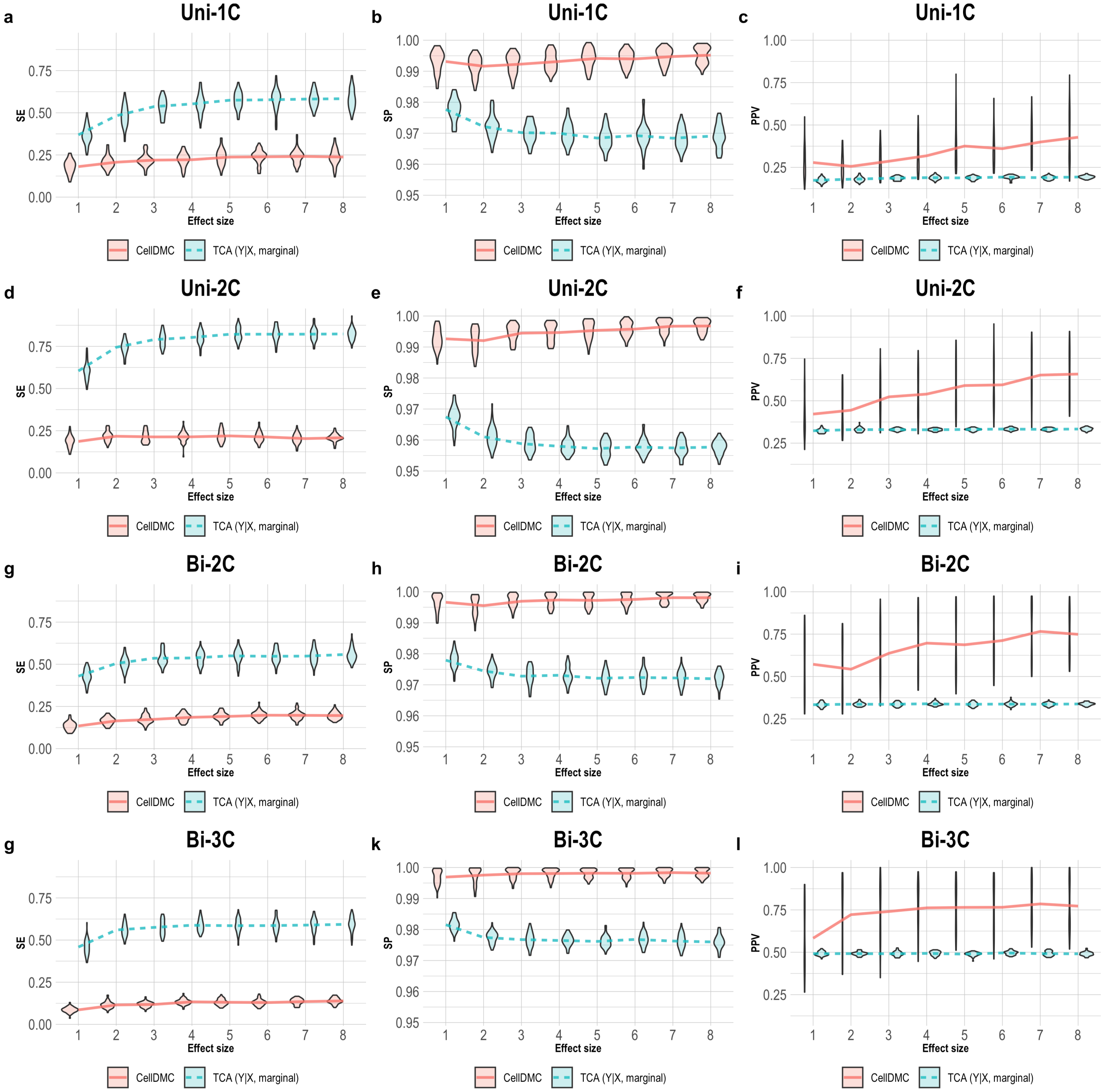

Figure 5.6: Evaluation of TCA and CellDMC in the case where the phenotype is affected by methylation $(Y|X)$, while executing TCA under the assumption $Y|X$ and using a marginal test. (a)-(c) Comparison of the sensitivity (SE), specificity (SP), and precision (positive predictive value; PPV) to detect differentially methylated cell-types as a function of the association effect size, under the scenario where a single cell type out of 6 blood cell types is altered in cases versus controls (Uni-1C). (d)-(f) as in Uni-1C, only for the scenario where two cell types are altered in the same direction (Uni-2C). (g)-(i) as in Uni-2C, only for the scenario where the cell types are altered in opposite directions (Bi-2C). (j)-(l) as in Bi-2C, only for three cell types (Bi-3C). Results are shown across 50 simulated data sets using violin plots; solid lines represent median values.

5.3.3 Analysis of methylation with smoking status

We next evaluated TCA and CellDMC in detecting differential methylation with smoking status in two large independent whole-blood methylation data sets [67, 94]. We formed a ground truth for evaluation by considering a set of 7 CpGs that were previously identified as exhibiting either myeloid-specific or lymphoid-specific changes in methylation [158].

Both TCA and CellDMC demonstrate an overall good performance in the detection of the 7 differentially methylated CpGs, with no clear difference in performance between the two methods (Figure 5.7a-b). Yet, evaluating specificity by applying the two methods on the entire data (i.e. rather than just on the 7 CpGs), shows that while TCA is well calibrated at the epigenome-wide level, CellDMC tends to suffer from a severe inflation in test statistic, thus indicating low specificity and precision for CellDMC (Figure 5.7d).

Notably, 7 out of the 14 tested CpGs across the two data sets did not achieve genome-wide significance, which would not have allowed de-novo detection of these associations in practice, presumably due to insufficient power. In order to address this, it is important to first appreciate that modeling the cell-type-specific nature of methylation is expected to benefit more types of analyses beyond calling for differentially methylated cell types. Particularly, compared to a standard regression analysis, TCA improves the detection of tissue-level associated CpGs via joint tests, wherein the effects of all cell types are tested jointly for their combined effect [140]. Such tissue-level tests can enable a powerful two-step approach of first detecting tissue-level associations followed by a post-hoc analysis of the associated CpGs at the cell-type level.

Indeed, combining a tissue-level test for each CpG with a cell-type level post-hoc analysis, as allowed by TCA, correctly detects 10 out of the 14 smoking associated CpGs and cell types at a genome-wide significance level (Figure 5.7c). This shows that the detection of tissue-level associations is of primary interest, and methods such as TCA and CellDMC should not be evaluated solely on their ability to directly capture differentially methylated cell types.

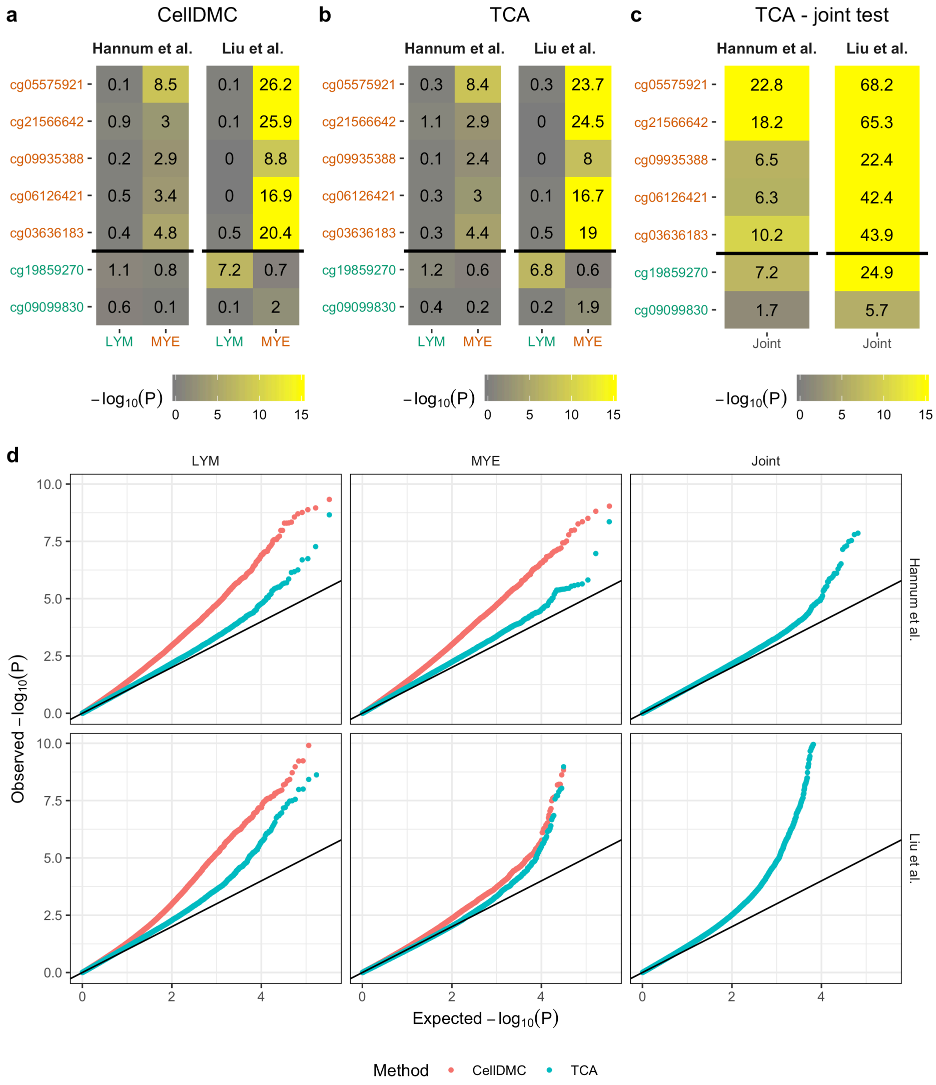

Figure 5.7: Evaluation of TCA and CellDMC in two independent whole-blood data sets with smoking. (a-c) Association tests were performed for each of 7 CpGs that were previously reported by Su et al. as exhibiting either myeloid-specific (in red) or lymphoid-specific (in green) associations with smoking status [158]. Results are displayed as heatmaps of the (negative-log transformed) p-values of the associations with myeloid cells (neutrophils and monocytes) and with lymphoid cells (T-cells, B-cells, and NK-cells) using (a) CellDMC, (b) TCA under the assumption $X|Y$ (using the `tca` function), and (c) TCA under the assumption $X|Y$, while using a joint test for tissue-level significance (also using the `tca` function). The latter achieves genome-wide significance (i.e. >6.98, assuming all 450K methylation array sites) in all but one CpG; calling the cell types that drive these associations using the results in (b) as a post-hoc analysis reveals the high-power of combining these two tests. (d) Results of an epigenome-wide analysis presented by quantile-quantile plots of the (negative-log transformed) p-values for the association tests in (a)-(c). Significant global deviation from the y=x line indicates an inflation arising from a badly specified model. Axes were truncated for visual purposes.

5.4 Discussion

We provide both empirical and theoretical evidence that for large enough sample sizes (at least 60), TCA is superior over the interaction model when it is applied under the assumptions taken in CellDMC, with the additional benefit of allowing to accommodate and therefore better handle different assumptions that are not allowed by CellDMC.

In light of the dramatic increase in sensitivity and decrease in specificity and precision observed in marginal tests compared with marginal conditional tests, we highly recommend to complement large data generation with small sets of sorted methylation data when possible. Such data can address the low precision limitation of the highly powerful marginal tests by providing a way to experimentally replicate associations at a cell-type-specific resolution.

In the absence of sorted data for validation, it is advised to use the less powerful yet more precise alternative tests provided in TCA. Particularly, a two-step approach, combining joint tests for the identification of tissue-level hits, followed by a cell-type level marginal conditional test for allowing a cell-type level resolution into the candidate CpGs, provides a powerful yet precise approach. That said, in some cases this two-step approach may not perform as well as a more standard one-step approach; for example, in scenarios where only a single cell type is expected to be related to the phenotype of interest, joint tests may be less powerful and therefore directly performing cell-type level tests are expected to be more powerful. A detailed discussion with practical guidelines for the selection of statistical tests and model assumptions in the application of TCA for detecting differential methylation is provided in Appendix B.

Appendix A

Tensor Composition Analysis: full model and optimization

A.0.1 The TCA model

Let Z_{hj}^i be the methylation level of individual $i \in \{1, ...n\}$ in cell type $h \in \{1, ...k\}$ at methylation site $j \in \{1, ...m\}$, and let $C^{(1)} \in \mathbb{R}^{p_1 \times n}$ be a matrix of p_1 covariates that may potentially affect methylation levels in a cell-type-specific manner. We assume:

$$Z_{hj}^i = \mu_{hj} + (c_i^{(1)})^T \gamma_h^j + \epsilon_{hj}^i \tag{A.1}$$

$$\epsilon_{hj}^i \sim N(0, \sigma_{hj}^2) \tag{A.2}$$

where $c_i^{(1)}$ is the i-th column of $C^{(1)}$ (corresponding to the p_1 covariates of the i-th individual), γ_h^j is a p_1-length vector of corresponding effects sizes for the p_1 covariates in the h-th cell type at site j, and e_{hj}^i is an i.i.d. component of variation.

We assume that observed methylation levels are convolved signals coming from k different cell-types. We denote $W \in \mathbb{R}^{k \times n}$ as a matrix of cell-type proportions of k cell types for each of the n individuals, and $C^{(2)} \in \mathbb{R}^{p_2 \times n}$ as a matrix of p_2 global covariates potentially affecting the observed methylation levels. Our model for X_{ij}, the observed methylation level of the i-th individual in cell type j, is as follows:

$$\begin{aligned} X_{ij} &= (c_i^{(2)})^T\delta_j + \sum_{h=1}^{k} w_{hi} Z_{hj}^i + \epsilon_{ij} && \text{(A.3)} \\ & \epsilon_{ij} \sim N(0, \tau^2) && \text{(A.4)} \\ \text{s.t.} \quad & \forall i : \sum_{h=1}^{k} w_{hi} = 1 && \text{(A.5)} \\ & \forall h, i : w_{hi} \geq 0 && \text{(A.6)} \end{aligned}$$

where $c_i^{(2)}$ is the i-th column of $C^{(2)}$ (corresponding to the p_2 covariates of the i-th individual), δ_j is a p_2-length vector of corresponding effects sizes of the p_2 covariates for the j-th site, and e_{ij} is a component of i.i.d. variation that models measurement noise.

A.0.2 Deriving the TCA estimator

Let $\Theta_j = (\mu_j, \sigma_j, w_i, \tau, \Gamma_j, \delta_j)$ be the set of the model's parameters for a particular site j, where Γ_j is a $p_1 \times k$ matrix with the vectors $\gamma_1^j, ..., \gamma_k^j$. Given the observed values, we are interested in the conditional distribution $Z_j^i | X_{ij} = x_{ij}$. Following the assumptions in (A.1) to (A.4), the conditional probability satisfies:

$$\begin{aligned}
Pr(Z_j^i = z_j^i | X_{ij} = x_{ij}, c_i^{(1)}, c_i^{(2)}, \Theta_j) &\propto Pr(Z_j^i = z_j^i | \mu_j, \sigma_j, c_i^{(1)}, \Gamma_j) Pr(X_{ij} = x_{ij} | Z_j^i = z_j^i, w_i, \tau, c_i^{(2)}, \delta_j) \\
&\propto exp\left(-\frac{1}{2}\left(z_j^i - \mu_j - \Gamma_j^T c_i^{(1)}\right)^T \Sigma_j^{-1} \left(z_j^i - \mu_j - \Gamma_j^T c_i^{(1)}\right)\right) \\
&\quad exp\left(-\frac{1}{2\tau^2}\left(x_{ij} - (z_j^i)^T w_i - (c_i^{(2)})^T \delta_j\right)^2\right) \\
&\propto exp\left(-\frac{1}{2}\left((z_j^i)^T \Sigma_j^{-1} z_j^i - 2(z_j^i)^T \Sigma_j^{-1}\left(\mu_j + \Gamma_j^T c_i^{(1)}\right)\right)\right) \\
&\quad exp\left(-\frac{1}{2\tau^2}\left((z_j^i)^T w_i w_i^T z_j^i - 2(z_j^i)^T w_i \left(x_{ij} - (c_i^{(2)})^T \delta_j\right)\right)\right) \\
&\propto exp\left(-\frac{1}{2}\left((z_j^i)^T\left(\Sigma_j^{-1} + \frac{w_i w_i^T}{\tau^2}\right) z_j^i\right)\right) \\
&\quad exp\left(-\frac{1}{2}\left(-2(z_j^i)^T\left(\Sigma_j^{-1}\left(\mu_j + \Gamma_j^T c_i^{(1)}\right) + w_i\left(\frac{x_{ij} - (c_i^{(2)})^T \delta_j}{\tau^2}\right)\right)\right)\right) \\
&\propto exp\left(-\frac{1}{2}(z_j^i - a_{ij})^T S_{ij}^{-1}(z_j^i - a_{ij})\right)
\end{aligned} \tag{A.7}$$

where

$$\Sigma_j = diag(\sigma_{1j}^2, ..., \sigma_{kj}^2) \tag{A.8}$$

$$S_{ij} = \left(\Sigma_j^{-1} + \frac{w_i w_i^T}{\tau^2}\right)^{-1} \tag{A.9}$$

$$a_{ij} = S_{ij}\left(\Sigma_j^{-1}\left(\mu_j + \Gamma_j^T c_i^{(1)}\right) + w_i\left(\frac{x_{ij} - (c_i^{(2)})^T \delta_j}{\tau^2}\right)\right) \tag{A.10}$$

The probability in (A.7) is maximized when z_j^i is the mode of the conditional distribution (which is the mean in this case). We therefore set the TCA estimator of z_j^i to be:

$$\hat{z}_j^i = a_{ij} = \left(\frac{w_i w_i^T}{\tau^2} + \Sigma_j^{-1}\right)^{-1}\left(\Sigma_j^{-1}\left(\mu_j + \Gamma_j^T c_i^{(1)}\right) + w_i\left(\frac{x_{ij} - (c_i^{(2)})^T \delta_j}{\tau^2}\right)\right) \tag{A.11}$$

A.0.3 Extracting underlying signals from convolved signals using TCA

In order to see why TCA can learn non-trivial information about the $\{z_{hj}^i\}$ values, note that [160]

$$Z_{hj}^i|X_{ij} \sim N\left(\tilde{\mu}_1 + \frac{\text{Cov}[Z_{hj}^i, X_{ij}]}{\tilde{\sigma}_2^2}(x_{ij} - \tilde{\mu}_2), \tilde{\sigma}_1^2 - \frac{\text{Cov}[Z_{hj}^i, X_{ij}]^2}{\tilde{\sigma}_2^2}\right) \tag{A.12}$$

where

$$\tilde{\mu}_1 = \text{E}[Z_{hj}^i], \tilde{\sigma}_1^2 = \text{V}[Z_{hj}^i] \tag{A.13}$$

$$\tilde{\mu}_2 = \text{E}[X_{ij}], \tilde{\sigma}_2^2 = \text{V}[X_{ij}] \tag{A.14}$$

Consider a simplified case where $\tau = 0$ and $\mu_{hj} = 0, \sigma_{hj} = 1$ for each h and some particular j. Assuming no covariates for simplicity, given the model of Z_{hj}^i and the model of X_{ij} in (A.1) to (A.4), we know that

$$\begin{aligned} \text{Cov}[Z_{hj}^i, X_{ij}] &= \text{E}[Z_{hj}^i X_{ij}] - \text{E}[Z_{hj}^i]\text{E}[X_{ij}] \\ &= \text{E}\left[Z_{hj}^i \sum_{l=1}^{k} w_{li} Z_{lj}^i\right] \\ &= w_{hi} \end{aligned} \tag{A.15}$$

Therefore, based on (A.12), in this case we get that

$$Z_{hj}^i|X_{ij} = x_{ij} \sim N\left(\frac{w_{hi}x_{ij}}{\sum_{l=1}^k w_{li}^2}, 1 - \frac{w_{hi}^2}{\sum_{l=1}^k w_{li}^2}\right) \tag{A.16}$$

This means that given the observed value x_{ij}, the conditional distribution of Z_{hj}^i has a lower variance compared with that of the marginal distribution of Z_{hj}^i ($\sigma_{hj}^2 = 1$), thus reducing the uncertainty and allowing us to provide a non-trivial estimate for the $\{z_{hj}^i\}$ values. This result is not specific for methylation but rather more general. In order to empirically verify this result and get an initial intuition as for the potential performance of TCA, we considered the following simplified general simulation.

We sampled three-dimensional source- and observation-specific values according to the model in (A.1)-(A.2) for every feature j, observation i and source h (i.e. for each of the $\{z_{hj}^{i}\}$ values) using $n = 250, m = 250, k = 3$ for the number of observations, features and sources, respectively. In this experiment, we sampled all the source- and observation-specific values, as well as the weights matrix (W), from a standard normal distribution. Eventually, we generated a matrix of observed mixtures (X) according to the model in (A.3)-(A.4) using the source- and observation-specific values, the weights matrix and an additional component of i.i.d. variation ($\tau = 0.01$). For performance evaluation, for each estimated vector $\hat{z}_{hj} = (\hat{z}_{hj}^{1}, ..., \hat{z}_{hj}^{n})^{T}$, we considered its linear correlation and mean squared error (MSE) with the true values in z_{hj}.

For simplicity, we assumed that all the parameters of the model are known, and applied TCA for estimating the $\{z_{hj}^{i}\}$ values. In order to form a baseline for comparison and to empirically verify that TCA can extract non-trivial information about the $\{z_{hj}^{i}\}$ values, we also applied TCA after permuting X (independent permutation of each row of the matrix). In addition, for each vector z_{hj}, we also measured to what extent its information can be captured by $x_j = (x_{1j}, ..., x_{nj})^{T}$, the observed levels in the j-th feature of X. We observed that TCA could effectively reconstruct a substantial portion of the information in the $\{z_{hj}\}$ vectors, far outperforming the baseline measurements (Figure A.1). We further verified the robustness of TCA by varying the parameters of the simulation across a wide range of values (Figure A.2).

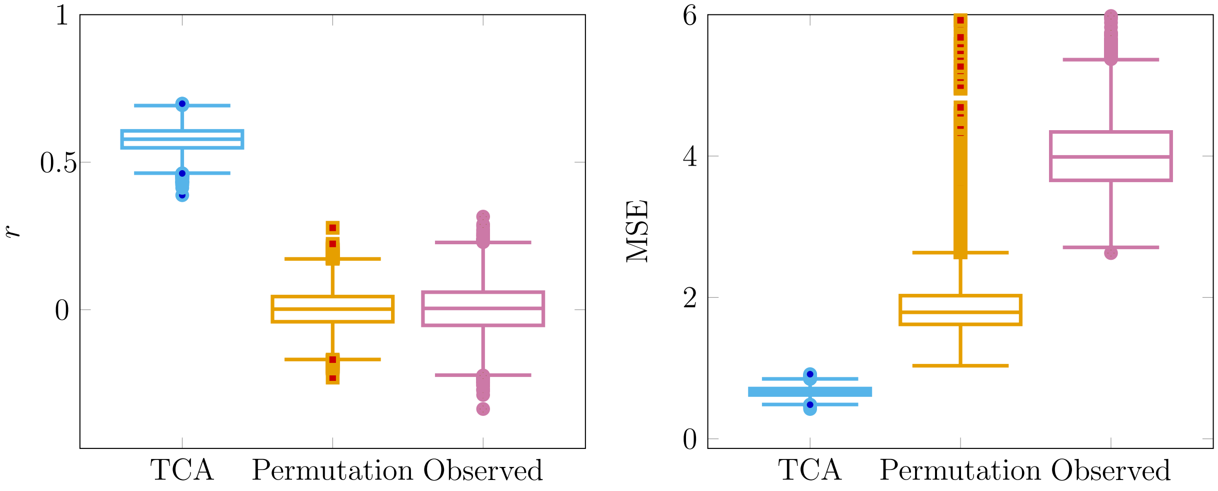

Figure A.1: Reconstructing three-dimensional observation- and source-specific values from two-dimensional input across ten simulated data sets ($n = 250, m = 250, k = 3, \tau = 0.01$). Three approaches were evaluated in capturing the observation-specific values for each feature j and source h (i.e. z_{hj}): TCA, TCA after permuting the observed two-dimensional data matrix ("Permutation") and directly using the observed data matrix ("Observed"). For each of the evaluated approaches, we present the distribution of the linear correlation between z_{hj} and its estimate $\hat{z}_{hj}$ across all h, j (in the left) and the distribution of the MSE between z_{hj} and its estimate $\hat{z}_{hj}$ across all h, j (in the right).

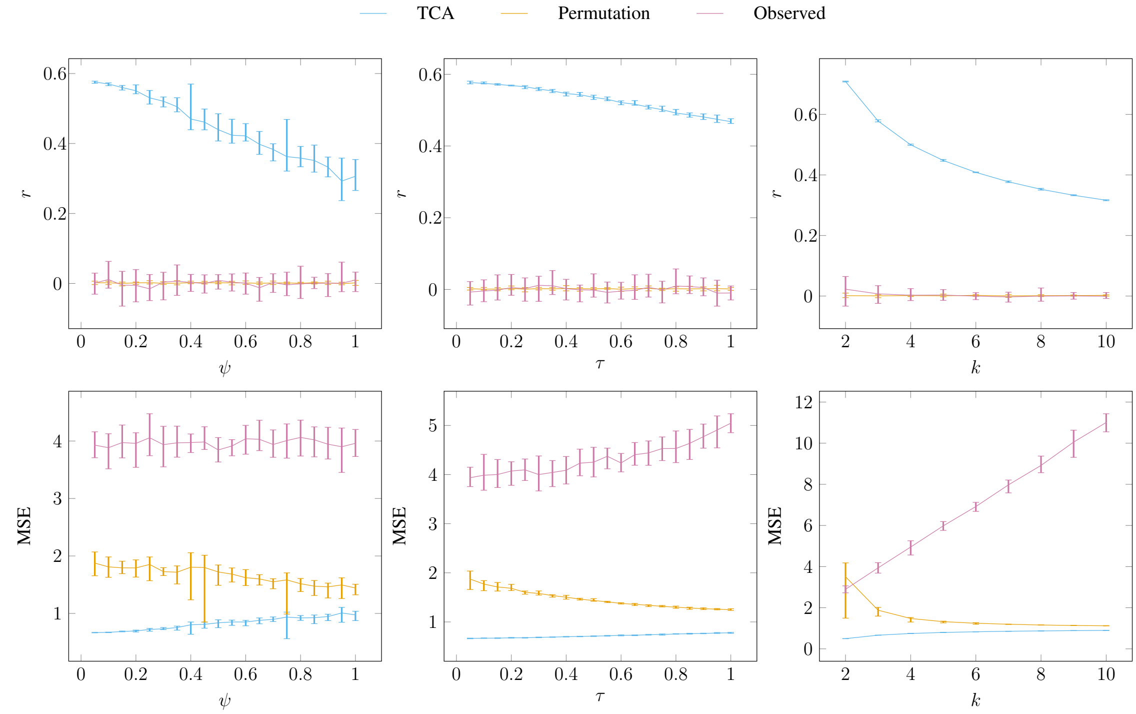

Figure A.2: Reconstructing three-dimensional observation- and source-specific values from two-dimensional input in simulated data ($n = 250, m = 250$) while varying the parameters of the simulation. Data was simulated under three scenarios: increasing level of i.i.d. noise added to W (ψ), increasing level of the i.i.d. component of variation added on top of X (τ) and increasing number of sources in the data (k). Three approaches were evaluated in capturing the observation-specific values for each feature j and source h (z_{hj}): TCA, TCA after permuting the observed data ("Permutation") and directly using the observed data ("Observed"). For each of the approaches and for each of the evaluated parameters, we present the median linear correlation between z_{hj} and its estimate $\hat{z}_{hj}$ across all h, j and across ten simulated data sets (top panel) and the median MSE between z_{hj} and its estimate $\hat{z}_{hj}$ across all h, j and across ten simulated data sets (bottom panel).

A.0.4 Inferring the parameters of the model

In order to estimate the $\{z_{hj}^i\}$ values, the TCA algorithm requires knowledge of the parameters in (A.1) to (A.4). Since X_{ij} is essentially a function of $Z_{1j}^i, ..., Z_{kj}^i$, we can use its assumed distribution for estimating all of the parameters in the model. More specifically, following the model in (A.3)-(A.4), we note that:

$$X_{ij} \sim N\left((c_i^{(2)})^T\delta_j + \sum_{h=1}^{k} w_{hi}\left(\mu_{hj} + (c_i^{(1)})^T\gamma_h^j\right), \sum_{h=1}^{k} w_{hi}^2\sigma_{hj}^2 + \tau^2\right) \tag{A.17}$$

We can therefore take an ML approach for estimating the parameters of the model from the observed data matrix X. In practice, we require an initial estimate of W as an input for the optimization. Such an estimate can be obtained by either using a reference-based approach [85] or a reference-free semi-supervised approach [134]. Given an estimate of W, we can then estimate the rest of the parameters in the model, and given estimates for the rest of the parameters in the model, we can update the estimate of W. We perform this alternating optimization procedure until convergence. Since we assume that different individuals are independent, updating W requires us to solve a set of n relatively easy optimization problems, each with k parameters, while satisfying the constraints in (A.5) and (A.6); we solve this numerically using a standard non-linear optimization procedure. Below, we describe the optimization of the rest of the parameters of the model given W (or an estimate of W).

Given W and the variances $\tau, \sigma_j = (\sigma_{1j}, ..., \sigma_{kj})^T$, ML solution for $\mu_j = (\mu_{1j}, ..., \mu_{kj})^T$, δ_j, $\{\gamma_h^j\}_{h=1}^k$ for feature j is given by solving the following constrained regression problem:

$$\hat{\mu}_j, \hat{\delta}_j, \{\hat{\gamma}_h^j\}_{h=1}^k = \underset{\mu_j,\delta_j,\{\gamma_h^j\}_{h=1}^k}{\operatorname{argmin}} \quad \sum_{i=1}^{n}\left(\tilde{x}_{ij} - \sum_{h=1}^{k}\tilde{w}_{hi}\mu_{hj} - \sum_{l=1}^{p_2}\tilde{c}_{li}^{(2)}\delta_{jl} - \sum_{l=1}^{p_1}\sum_{h=1}^{k}\tilde{c}_{lih}^{(1)}\gamma_h^j\right)^2 \tag{A.18}$$

$$\text{s.t.} \quad \forall 1 \leq j \leq m : \mu_{hj} \in [0,1] \tag{A.19}$$

where

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\sum_{l=1}^{k} w_{li}^2 \sigma_{lj}^2 + \tau^2}} \tag{A.20}$$

$$\tilde{w}_{ih} = \frac{w_{hi}}{\sqrt{\sum_{l=1}^{k} w_{li}^2 \sigma_{lj}^2 + \tau^2}} \tag{A.21}$$

$$\tilde{c}_{lih}^{(1)} = \frac{w_{ih} c_{li}^{(1)}}{\sqrt{\sum_{d=1}^{k} w_{di}^2 \sigma_{dj}^2 + \tau^2}} \tag{A.22}$$

$$\tilde{c}_{li}^{(2)} = \frac{c_{li}^{(2)}}{\sqrt{\sum_{h=1}^{k} w_{hi}^2 \sigma_{hj}^2 + \tau^2}} \tag{A.23}$$

and where δ_{jl} is the l-th entry of the vector δ_j, and $c_{li}^{(1)}, c_{li}^{(2)}$ are the l-th covariate of individual i in $C^{(1)}$ and in $C^{(2)}$, respectively. The constraints in (A.19) reflect the fact that methylation levels are bounded to the range $[0, 1]$, which means the mean levels should be also bounded to that range. We note that in principle we should also constrain the effects contributed by $\delta^j, \{\gamma_h^j\}_{h=1}^k$, in order to make sure that the total estimated methylation levels do not fall out of the range $[0, 1]$. In practice, in real data, these additional constraints may result with less accurate estimates. This problem can be solved efficiently using quadratic programming.

Since τ, σ_j are typically unknown, we perform an alternative optimization procedure as follows. We start by finding initial estimates for $\delta^j, \{\gamma_h^j\}_{h=1}^k$, by assuming that $\sigma_{1j} = ... = \sigma_{kj}, \tau = 0$. Under these conditions, the solution to the optimization problem in (A.18) is now independent of σ_j, τ. Specifically, for obtaining an initial estimate of $\mu_j, \delta^j, \{\gamma_h^j\}_{h=1}^k$, we solve the problem in (A.18) while setting

$$\tilde{x}_{ij} = \frac{x_{ij}}{\|w_i\|_2} \tag{A.24}$$

$$\tilde{w}_{hi} = \frac{w_{hi}}{\|w_i\|_2} \tag{A.25}$$

$$\tilde{c}_{lih}^{(1)} = \frac{w_{hi} c_{lih}^{(1)}}{\|w_i\|_2} \tag{A.26}$$

$$\tilde{c}_{li}^{(2)} = \frac{c_{li}^{(2)}}{\|w_i\|_2} \tag{A.27}$$

Once we obtain $\hat{\mu}_j, \hat{\delta}_j, \{\hat{\gamma}_h^j\}_{h=1}^k$, we can fix them and estimate σ_j, τ using any hill climbing algorithm (and then repeat until convergence). In practice, for learning σ_j, τ we perform another alternating optimization procedure as follows. We first assume τ to be unique for each site and estimate for each site j separately initial estimates of σ_j, τ. Then, we re-estimate τ using the entire data and the estimates of $\{\sigma_j\}$ from all sites, and finally, we re-estimate σ_j for each site j using the updated estimate of τ.

Notably, the number of parameters we need to estimate in our model is very large compared with the number of data points available for inference. However, for each set of constant number of parameters that we estimate, we use n data points. For instance, for estimating the parameters $\mu_j, \delta_j, \{\gamma_h^j\}_{h=1}^k$ for site j (a constant set of $k(p_1+1)+p_2$ parameters), we use n data points.

A.0.5 Testing a phenotype for cell-type-specific associations

TCA allows us to estimate cell-type-specific methylation levels for each individual in the data. In principle, such estimates can then be used for running a cell-type-specific EWAS by testing the estimates of a particular cell type for association with a phenotype of interest (or for running a joint test for several cell types by using their estimated cell-type-specific methylation levels jointly). However, for the application of association testing, we suggest an alternative one-step approach instead of the more straightforward two-steps approach.

We model the phenotype of interest as potentially affected by cell-type-specific methylation levels, and use the conditional distribution of the phenotype given the observed data in X. Effectively, this allows us to integrate over all the potential values of the $\{z_{hj}^i\}$ individual and cell-type-specific levels. In addition to taking into account covariates that may affect the methylation levels, as described in (A.1) and in (A.3), we also consider potential direct effects of other (or the same) covariates on the phenotype.

A.0.6 Joint test for effect sizes in all cell types

Let $Y \in \mathbb{R}^{n\times 1}$ be a quantitative phenotype of interest, where Y_i corresponds to the phenotypic level of sample i, and let $C^{(3)} \in \mathbb{R}^{p_3\times n}$ be a matrix of p_3 covariates potentially affecting the phenotype (may also include an intercept term), we assume the following model:

$$Y_i = (c_i^{(3)})^T\alpha + \sum_{h=1}^{k}\beta_{hj}Z_{hj}^i + e_i \tag{A.28}$$

$$e_i \sim N(0, \phi^2) \tag{A.29}$$

where $\beta_{1j}, ..., \beta_{kj}$ are the effect sizes of the k different cell types in site j. Recall the model of X_{ij} in (A.3)-(A.4), using (A.12) we get

$$Y_i|X_{ij} \sim N\left(\tilde{\mu}_1 + \frac{\mathrm{Cov}[X_{ij}, Y_i]}{\tilde{\sigma}_2^2}(x_{ij} - \tilde{\mu}_2), \tilde{\sigma}_1^2 - \frac{\mathrm{Cov}[X_{ij}, Y_i]^2}{\tilde{\sigma}_2^2}\right) \tag{A.30}$$

where

$$\tilde{\mu}_1 = \mathrm{E}[Y_i], \tilde{\sigma}_1^2 = \mathrm{V}[Y_i] \tag{A.31}$$

$$\tilde{\mu}_2 = \mathrm{E}[X_{ij}], \tilde{\sigma}_2^2 = \mathrm{V}[X_{ij}] \tag{A.32}$$

Different individuals are assumed to be independent (both in their phenotypic and their cell-type-specific methylation levels), and $\mathrm{Cov}[y_i, X_{tj}] = 0$ for any $t \neq i$.

Note that

$$
\begin{aligned}
\mathrm{Cov}[X_{ij}, Y_i] &= \mathrm{E}[Y_i X_{ij}] - \mathrm{E}[Y_i]\mathrm{E}[X_{ij}] \\
&= \mathrm{E}\left[\left((c_i^{(3)})^T\alpha + \sum_{h=1}^{k} \beta_{hj} Z_{hj}^i + e\right)\left((c_i^{(2)})^T\delta^j + \sum_{h=1}^{k} w_{hi} Z_{hj}^i + \epsilon\right)\right] \\
&- \left((c_i^{(3)})^T\alpha + \sum_{h=1}^{k} \beta_{hj}\mathrm{E}[Z_{hj}^i]\right)\left(c_i^{(2)}\delta^j + \sum_{h=1}^{k} w_{hi}\mathrm{E}[Z_{hj}^i]\right) \\
&= (c_i^{(3)})^T\alpha \sum_{h=1}^{k} w_{hi}\mathrm{E}[Z_{hj}^i] + (c_i^{(2)})^T\delta_j \sum_{h=1}^{k} \beta_{hj}\mathrm{E}[Z_{hj}^i] + \mathrm{E}\left[\sum_{h=1}^{k} \beta_{hj} Z_{hj}^i \sum_{h=1}^{k} w_{hi} Z_{hj}^i\right] \\
&- (c_i^{(3)})^T\alpha \sum_{h=1}^{k} w_{hi}\mathrm{E}[Z_{hj}^i] - (c_i^{(2)})^T\delta_j \sum_{h=1}^{k} \beta_{hj}\mathrm{E}[Z_{hj}^i] - \sum_{h=1}^{k} \beta_{hj}\mathrm{E}[Z_{hj}^i] \sum_{h=1}^{k} w_{hi}\mathrm{E}[Z_{hj}^i] \\
&= \sum_{h=1}^{k} w_{hi}\beta_{hj}\mathrm{E}[(Z_{hj}^i)^2] - \sum_{h=1}^{k} w_{hi}\beta_{hj}\mathrm{E}[Z_{hj}^i]^2 \\
&= \sum_{h=1}^{k} w_{hi}\beta_{hj}\sigma_{hj}^2
\end{aligned}
\tag{A.33}
$$

Therefore, we get

$$
Y_i | X_{ij} = x_{ij} \sim N\left(\tilde{\mu}_{ij}, \tilde{\sigma}_{ij}^2\right) \tag{A.34}
$$

where

$$
\tilde{\mu}_{ij} = (c_i^{(3)})^T\alpha + \sum_{h=1}^{k} \beta_{hj}\left(\mu_{hj} + (c_i^{(1)})^T\gamma_h^j + \frac{w_{hi}\sigma_{hj}^2 \tilde{x}_{ij}}{\tau^2 + \sum_{l=1}^{k} w_{li}^2\sigma_{lj}^2}\right) \tag{A.35}
$$

$$
\tilde{x}_{ij} = x_{ij} - (c_i^{(2)})^T\delta_j - \sum_{l=1}^{k} w_{li}(\mu_{lj} + (c_i^{(1)})^T\gamma_l^j) \tag{A.36}
$$

$$
\tilde{\sigma}_{ij}^2 = \phi^2 + \sum_{h=1}^{k} \beta_{hj}^2\sigma_{hj}^2 - \frac{\left(\sum_{h=1}^{k} \beta_{hj} w_{hi}\sigma_{hj}^2\right)^2}{\tau^2 + \sum_{h=1}^{k} w_{hi}^2\sigma_{hj}^2} \tag{A.37}
$$

Using the distributions $Y_i | X_{ij} = x_{ij}$ for each individual i, we can now consider the following hypothesis testing for site j:

$$H_0 : \beta_{1j} = ... = \beta_{kj} = 0 \tag{A.38}$$

$$H_1 : \exists h.\beta_{hj} \neq 0 \tag{A.39}$$

This formulation essentially tests the particular site under test j for association with the phenotype by considering the joint contribution of all cell-type-specific effects. Alternatively, we can look for cell-type-specific effects of a subset of the cell types.

A.0.7 Marginal test for the effect size of a particular cell type

Consider the following model:

$$Y_i = (c_i^{(3)})^T \alpha + \beta_{hj} Z_{hj}^i + e_i \tag{A.40}$$

$$e_i \sim N(0, \phi^2) \tag{A.41}$$

where β_{hj} is the effect size of a particular cell type h. Similarly as before, we get:

$$Y_i | X_{ij} = x_{ij} \sim N\left(\tilde{\mu}_{ij}, \tilde{\sigma}_{ij}^2\right) \tag{A.42}$$

where

$$\tilde{\mu}_{ij} = (c_i^{(3)})^T \alpha + \beta_{hj} \left(\mu_{hj} + (c_i^{(1)})^T \gamma_h^j + \frac{w_{hi} \sigma_{hj}^2 \tilde{x}_{ij}}{\tau^2 + \sum_{l=1}^k w_{li}^2 \sigma_{lj}^2} \right) \tag{A.43}$$

$$\tilde{x}_{ij} = x_{ij} - (c_i^{(2)})^T \delta_j - \sum_{l=1}^k w_{li} (\mu_{lj} + (c_i^{(1)})^T \gamma_l^j) \tag{A.44}$$

$$\tilde{\sigma}_{ij}^2 = \phi^2 + \beta_{hj}^2 \left(\sigma_{hj}^2 - \frac{w_{hi}^2 \sigma_{hj}^4}{\tau^2 + \sum_{l=1}^k w_{li}^2 \sigma_{lj}^2} \right) \tag{A.45}$$

Using the distributions $Y_i | X_{ij} = x_{ij}$ for each individual i, we can now consider the following hypothesis testing for site j:

$$H_0 : \beta_{hj} = 0 \tag{A.46}$$

$$H_1 : \beta_{hj} \neq 0 \tag{A.47}$$

We calculate p-values for both the joint test and the marginal test using a generalized likelihood-ratio test. The null model can be fitted using standard ML estimators. For the alternative model, given the estimates for a particular site j, $\Theta_j = (\mu_j, \sigma_j, W, \tau, \Gamma_j, \delta_j)$, and given the observed data $Y, X_j, C^{(1)}, C^{(2)}, C^{(3)}$, the parameters $\alpha = (\alpha_1, ..., \alpha_p), \phi$ and $\beta_j = (\beta_{1j}, ..., \beta_{kj})$ (in a marginal test for cell type h only the estimate of β_{hj} is needed) can be estimated using ML. In practice, we do that by numerically maximizing the log likelihood of the conditional distribution using a standard non-linear optimization procedure.

Throughout our experiments, we observed that TCA, albeit powerful, resulted in a deflation in the test statistic under the null, leading it to be an over-conservative test. This behavior may be explained by the optimization procedure we apply. Specifically, an appropriate application of the generalized-likelihood ratio test we use relies upon using ML estimates of the parameters in the TCA model. In our case, we achieve ML estimates under the null model, however, in general, we do not achieve ML estimates under the alternative model for two reasons. First, our optimization procedure involves a non-convex optimization, which is not guaranteed to yield global optimum, and second, for computational convenience, we leverage only the bulk methylation data (X) in learning the parameters of the TCA model. The latter is not optimal since in principle the phenotypic data (Y) provides more information about the parameters of the model. As a result, the estimates under the alternative hypothesis are not ML estimates, which leads to a lower likelihood of the alternative model and therefore to a deflation in the test statistic of the generalized-likelihood ratio test (and thus the test is over-conservative).

Appendix B

Tensor Composition Analysis: A practical guide

B.0.1 Statistical testing within the TCA framework

The TCA framework allows us to run several different types of statistical tests on a phenotype of interest, each of which can test a different hypothesis about the statistical relation of the phenotype to the methylation levels under test. In this section, we briefly describe the statistical tests we implemented as part of the TCA R package (`TCA` on CRAN).

Much like in standard regression analysis, where we can test different hypotheses about the coefficients of the independent variables, both the model in (5.7)-(5.8) and in (5.15) can be used to test several different hypotheses about the effect sizes $\beta_{1j}, ..., \beta_{kj}$. Below, we provide a list of the tests we implemented in the TCA R package for testing the statistical association of a given phenotype with each particular methylation site j in the data.

Tests under the biological assumption $Y|X$, using the model in (5.15):

- *marginal conditional* - fits the parameters $\beta_{1j}, ..., \beta_{kj}$ for all cell types jointly and tests for the significance of the effect of each cell type separately.

- *marginal* - for each cell type h, fits the parameter β_{hj} and tests for the significance of its effect, while assuming $\forall l \neq h : \beta_{lj} = 0$.

- *joint* - fits the parameters $\beta_{1j}, ..., \beta_{kj}$ for all cell types jointly and tests for the significance of the overall effect across all cell types (i.e. a tissue-level test).

- *single effect size* - fits the parameters $\beta_{1j}, ..., \beta_{kj}$ under the assumption $\beta_{1j} = ... = \beta_{kj}$ and tests for the significance of the overall effect across all cell types.

- *custom* - compares and tests any two nested models, each representing a subset of the parameters $\beta_{1j}, ..., \beta_{kj}$, and tests for the significance of the overall effect across all cell types in the alternative model (i.e. the larger model) that are not in the null model (i.e. the smaller model).

Tests under the biological assumption $X|Y$, using the model in (5.7)-(5.8):

- *marginal conditional* - same as the analogous model for $X|Y$ above, only under the assumption $X|Y$.
- *joint* - same as the analogous model for $X|Y$ above, only under the assumption $X|Y$.

Of note, the interaction model in (5.3) can in principle define more statistical tests on the cell-type-specific effect sizes under the assumption $X|Y$ (i.e. note just a marginal conditional test), similarly to the tests above allowed in TCA. However, these are expected to be inferior to their analogous ones in TCA given that the interaction model is a degenerate case of TCA (see Subsection 5.2.4). Notably, there is no clear way to accommodate the assumption $Y|X$ in the innteraction model, as the method relies on the interaction between cell-type proportions and a phenotype, which is the explained variable to be modeled under $Y|X$.

B.0.2 Selecting appropriate statistical tests for differential methylation at cell-type resolution

The first important decision one should make when testing for differential methylation at cell-type resolution is how to set the model directionality. This decision clearly should be context- and phenotype-dependent, however, admittedly, it may often not be clear how to make an informed decision. Yet, in some cases, the selection of a biological assumption is natural. For example, when looking for associations with demographic factors such as age or ancestry, it makes no sense to assume $Y|X$, as these demographics cannot be altered by methylation.

Based on the results from our simulation study (see Subsection 5.3.2), the consistency between TCA and the interaction model may provide useful evidence as for the true underlying model. Specifically, high consistency in the predicted associations between TCA and the interaction model while applying TCA under the assumption $X|Y$ provides evidence that the assumption $X|Y$ holds (Figure 5.1). In contrast, limited consistency between the two methods owing to lower specificity and precision of the interaction model - which is expected to result in more predicted associations for the interaction model over TCA and likely demonstrate an inflation in test statistic - provides evidence that the assumption $Y|X$ holds; either when applying TCA under the assumption $X|Y$ (Figure 5.4) or when applying TCA under the assumption $Y|X$, which is expected to provide even worse inconsistency based on our simulations (Figure 5.3). That said, in practice, the extent to which a given phenotype will demonstrate patterns of consistency that are similar to those revealed by simulations is still unclear; particularly, a phenotype may be both affected by some methylation sites (or by a component captured by methylation) and affect some other methylation sites (or a mediating component thereof).

In cases of association studies that aim at de-novo detection of differential methylation, we recommend to apply a joint test for an initial screening for tissue-level associations. Then, marginal conditional tests should be performed as a post-hoc analysis for calling differentially methylated cell types in the CpGs that passed multiple testing correction in the first tissue-level screening.

As per our previous suggestion, we recommend that future studies include small replication data sets from sorted or single cells, in which case users may opt to replace marginal conditional tests with the much more powerful, yet less precise marginal tests; in such cases, pre-screening for tissue-level associations using joint tests may be less powerful.

In cases where only a single cell type (or a small subset of cell types) is associated with the phenotype, joint tests are expected to be less powerful than marginal and marginal conditional tests (owing to the unnecessarily higher degrees of freedom in a joint test). While this is typically unknown a-priori, this rationale can be applied to cases where only a particular cell type (or a small subset of cell types) are of interest, in which case, an initial screening step using a joint test should be avoided.

Lastly, it is important to understand the limitations of methods such as TCA and the interaction model. Particularly, there are limitations that rise due to inherent properties of these models: first, the proportions of different cell types are correlated, owing to the fact that fractions sum up to 1 and thus depend on each other, and second, the higher the abundance of a cell type is, the higher the variance that it accounts for in the observed mixture data. Consequently, the estimated cell-type-specific methylation in TCA and the cell-type-phenotype interactions in the interaction model, both of which directly rely on the cell-type proportions, are expected to be correlated between different cell types. For that reason, these two models are expected to be bounded in their precision to detect truly differentially methylated cell types.

In order to see that, consider an example where we have cell type A with accurate estimates of the cell-type proportions and cell type B with less accurate estimates of the cell-type proportions (e.g., granulocytes and monocytes in whole-blood). Further assume that cell type B is truly differentially methylated at some particular CpG under test. In that case, if the proportions of both cell types are highly correlated (and they typically are), cell type A may capture some of the cell-type-specific methylation of cell type B that was not captured by directly using the proportions of cell type B (due to the limited accuracy of the estimated proportions of cell type B and the correlation between the proportions of A and B). Not only that, these correlations between the proportions of different cell types may also introduce high multicollinearity between estimated methylation of different cell types and their effects. As a result of these, cell type A in our example may be called as differentially methylated, even though the true signal is coming from cell type B.

The above limitations are also the reason for the particularly low precision of the marginal test in TCA, where the cell type under test tends to capture true signal coming from other cell types as well (Figures 5.5 and 5.6), much like what one would observe when applying marginal tests under standard linear regression in the case of having multiple highly correlated features (where only some of them are truly statistically related to the dependent variable). Importantly, applying a marginal conditional test instead of a marginal test mitigates this limitation (although not completely, as explained above) by accounting for the other cell types.

[1] Randy L Jirtle and Michael K Skinner. Environmental epigenomics and disease susceptibility. *Nature reviews genetics*, 8(4):253–262, 2007. 1

[2] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G Bragi Walters, Steinunn Gunnarsdottir, et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428, 2008. 1, 3

[3] Greg Gibson. The environmental contribution to gene expression profiles. *Nature reviews genetics*, 9(8):575–581, 2008. 1

[4] William Cookson, Liming Liang, Gonçalo Abecasis, Miriam Moffatt, and Mark Lathrop. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194, 2009. 1

[5] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015. 1

[6] Antigone S Dimas, Samuel Deutsch, Barbara E Stranger, Stephen B Montgomery, Christelle Borel, Homa Attar-Cohen, Catherine Ingle, Claude Beazley, Maria Gutierrez Arcelus, Magdalena Sekowska, et al. Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science*, 325(5945):1246–1250, 2009. 1

[7] Eulàlia De Nadal, Gustav Ammerer, and Francesc Posas. Controlling gene expression in response to stress. *Nature Reviews Genetics*, 12(12):833–845, 2011. 1

[8] Julien Bryois, Alfonso Buil, Pedro G Ferreira, Nikolaos I Panousis, Andrew A Brown, Ana Viñuela, Alexandra Planchon, Deborah Bielser, Kerrin Small, Tim Spector, et al. Time-dependent genetic effects on gene expression implicate aging processes. *Genome research*, 27(4):545–552, 2017. 1

[9] Jennifer C Long and Javier F Caceres. The sr protein family of splicing factors: master regulators of gene expression. *Biochemical Journal*, 417(1):15–27, 2009. 1

[10] Ino D Karemaker and Michiel Vermeulen. Single-cell dna methylation profiling: technologies and biological applications. *Trends in biotechnology*, 36(9):952–965, 2018. 2

[11] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002. 2, 79

[12] J Raphael Gibbs, Marcel P van der Brug, Dena G Hernandez, Bryan J Traynor, Michael A Nalls, Shiao-Lin Lai, Sampath Arepalli, Allissa Dillman, Ian P Rafferty, Juan Troncoso, et al. Abundant quantitative trait loci exist for dna methylation and gene expression in human brain. *PLoS Genet*, 6(5):e1000952, 2010. 3, 5, 7, 17

[13] Shai S Shen-Orr and Renaud Gaujoux. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current opinion in immunology*, 25(5):571–578, 2013. 3

[14] Alexander J Titus, Rachel M Gallimore, Lucas A Salas, and Brock C Christensen. Cell-type deconvolution from dna methylation: a review of recent applications. *Human molecular genetics*, 26(R2):R216–R224, 2017. 3

[15] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008. 5, 21

[16] Alkes L Price, Johannah Butler, Nick Patterson, Cristian Capelli, Vincenzo L Pascali, Francesca Scarnicci, Andres Ruiz-Linares, Leif Groop, Angelica A Saetta, Penelope Korkolopoulou, et al. Discerning the ancestry of european americans in genetic association studies. *PLoS Genet*, 4(1):e236, 2008. 5

[17] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000. 5

[18] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006. 5, 7, 9, 18

[19] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009. 5, 15, 18, 23, 25

[20] Wen-Yun Yang, John Novembre, Eleazar Eskin, and Eran Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nature genetics*, 44(6):725–731, 2012. 5

[21] Eran Elhaik, Tatiana Tatarinova, Dmitri Chebotarev, Ignazio S Piras, Carla Maria Calò, Antonella De Montis, Manuela Atzori, Monica Marini, Sergio Tofanelli, Paolo Francalacci, et al. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nature communications*, 5, 2014. 5

[22] Bogdan Pasaniuc, Noah Zaitlen, Guillaume Lettre, Gary K Chen, Arti Tandon, WH Linda Kao, Ingo Ruczinski, Myriam Fornage, David S Siscovick, Xiaofeng Zhu, et al. Enhanced statistical tests for gwas in admixed populations: assessment using african americans from care and a breast cancer consortium. *PLoS Genet*, 7(4):e1001371, 2011. 5

[23] Dandan Zhang, Lijun Cheng, Judith A Badner, Chao Chen, Qi Chen, Wei Luo, David W Craig, Margot Redman, Elliot S Gershon, and Chunyu Liu. Genetic control of individual differences in gene-specific methylation in human brain. *The American Journal of Human Genetics*, 86(3):411–419, 2010. 5, 7, 10, 17

[24] Jordana T Bell, Athma A Pai, Joseph K Pickrell, Daniel J Gaffney, Roger Pique-Regi, Jacob F Degner, Yoav Gilad, Jonathan K Pritchard, et al. Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biol*, 12(1):R10, 2011. 5, 7, 10, 17

[25] Degui Zhi, Stella Aslibekyan, Marguerite R Irvin, Steven A Claas, Ingrid B Borecki, Jose M Ordovas, Devin M Absher, and Donna K Arnett. Snps located at cpg sites modulate genome-epigenome interaction. *Epigenetics*, 8(8):802–806, 2013. 5, 7, 10, 13

[26] Joshua M Galanter, Christopher R Gignoux, Sam S Oh, Dara Torgerson, Maria Pino-Yanes, Neeta Thakur, Celeste Eng, Donglei Hu, Scott Huntsmann, Harold J Farber, et al. Methylation analysis reveals fundamental differences between ethnicity and genetic ancestry. *bioRxiv*, page 036822, 2016. 5, 12

[27] Hunter B Fraser, Lucia L Lam, Sarah M Neumann, and Michael S Kobor. Population-specificity of human dna methylation. *Genome Biol*, 13(2):R8, 2012. 5

[28] Erika L Moen, Xu Zhang, Wenbo Mu, Shannon M Delaney, Claudia Wing, Jennifer McQuade, Jamie Myers, Lucy A Godley, M Eileen Dolan, and Wei Zhang. Genome-wide variation of cytosine modifications between european and african populations and the implications for complex traits. *Genetics*, 194(4):987–996, 2013. 5, 6

[29] Alicia K Smith, Varun Kilaru, Mehmet Kocak, Lynn M Almli, Kristina B Mercer, Kerry J Ressler, Frances A Tylavsky, and Karen N Conneely. Methylation quantitative trait loci (meqtls) are consistently detected across ancestry, developmental stage, and tissue type. *BMC genomics*, 15(1):145, 2014. 5

[30] Marco P Boks, Eske M Derks, Daniel J Weisenberger, Erik Strengman, Esther Janson, Iris E Sommer, Rene S Kahn, and Roel A Ophoff. The relationship of dna methylation with age, gender and genotype in twins and healthy controls. *PloS one*, 4(8):e6767, 2009. 5

[31] Kristi Kerkel, Alexandra Spadola, Eric Yuan, Jolanta Kosek, Le Jiang, Eldad Hod, Kerry Li, Vundavalli V Murty, Nicole Schupf, Eric Vilain, et al. Genomic surveys by methylation-sensitive snp analysis identify sequence-dependent allele-specific dna methylation. *Nature genetics*, 40(7):904–908, 2008. 5

[32] Leonard C Schalkwyk, Emma L Meaburn, Rebecca Smith, Emma L Dempster, Aaron R Jeffries, Matthew N Davies, Robert Plomin, and Jonathan Mill. Allelic skewing of dna methylation is widespread across the genome. *The American Journal of Human Genetics*, 86(2):196–212, 2010. 5

[33] Nicholas E Banovich, Xun Lan, Graham McVicker, Bryce Van de Geijn, Jacob F Degner, John D Blischak, Julien Roux, Jonathan K Pritchard, and Yoav Gilad. Methylation qtls are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet*, 10(9):e1004663, 2014. 5

[34] Richard T Barfield, Lynn M Almli, Varun Kilaru, Alicia K Smith, Kristina B Mercer, Richard Duncan, Torsten Klengel, Divya Mehta, Elisabeth B Binder, Michael P Epstein, et al. Accounting for population stratification in dna methylation studies. *Genetic epidemiology*, 38(3):231–241, 2014. x, xi, xii, 6, 16, 18, 21, 22, 25, 26

[35] Devin C Koestler, Brock C Christensen, Margaret R Karagas, Carmen J Marsit, Scott M Langevin, Karl T Kelsey, John K Wiencke, and E Andres Houseman. Blood-based profiles of dna methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics*, 8(8):816–826, 2013. 6, 22, 48, 74

[36] Elior Rahmani, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, Esteban G Burchard, Eleazar Eskin, James Zou, et al. Sparse pca corrects for cell type heterogeneity in epigenome-wide association studies. *Nature methods*, 13(5):443–445, 2016. xx, 6, 12, 16, 22, 23, 32, 33, 37, 47, 48, 50, 76, 84, 89, 95, 105

[37] Yi-an Chen, Mathieu Lemire, Sanaa Choufani, Darci T Butcher, Daria Grafodatskaya, Brent W Zanke, Steven Gallinger, Thomas J Hudson, and Rosanna Weksberg. Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray. *Epigenetics*, 8(2):203–209, 2013. 6, 13, 15, 22, 45, 92, 130

[38] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 8

[39] Steve Horvath. Dna methylation age of human tissues and cell types. *Genome biology*, 14(10):R115, 2013. 8, 21, 33, 64, 76, 83

[40] Paula Singmann, Doron Shem-Tov, Simone Wahl, Harald Grallert, Giovanni Fiorito, So-Youn Shin, Katharina Schramm, Petra Wolf, Sonja Kunze, Yael Baran, et al. Characterization of whole-genome autosomal differences of dna methylation between men and women. *Epigenetics & chromatin*, 8(1):1–13, 2015. 8, 21, 33, 64, 76, 83

[41] Sonja Zeilinger, Brigitte Kühnel, Norman Klopp, Hansjörg Baurecht, Anja Kleinschmidt, Christian Gieger, Stephan Weidinger, Eva Lattka, Jerzy Adamski, Annette Peters, et al. Tobacco smoking leads to extensive genome-wide changes in dna methylation. *PloS one*, 8(5):e63812, 2013. 8

[42] H Wichmann, C Gieger, T Illig, et al. Kora-gen-resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*, 67:S26, 2005. 12, 16

[43] Liliane Pfeifferm, Simone Wahl, Luke C Pilling, Eva Reischl, Johanna K Sandling, Sonja Kunze, Lesca M Holdt, Anja Kretschmer, Katharina Schramm, Jerzy Adamski, et al. Dna methylation of lipid-related genes affects blood lipid levels. *Circulation: Genomic and Precision Medicine*, pages CIRCGENETICS–114, 2015. 12, 79

[44] Andrew E Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics*, 29(2):189–196, 2013. 12, 13

[45] Ruth Pidsley, Chloe CY Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C Schalkwyk. A data-driven approach to preprocessing illumina 450k methylation array data. *BMC genomics*, 14(1):293, 2013. 12

[46] Melanie Kolz, Toby Johnson, Serena Sanna, Alexander Teumer, Veronique Vitart, Markus Perola, Massimo Mangino, Eva Albrecht, Chris Wallace, Martin Farrall, et al. Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genet*, 5(6):e1000504, 2009. 12

[47] Joshua M Galanter, Christopher R Gignoux, Dara G Torgerson, Lindsey A Roth, Celeste Eng, Sam S Oh, Elizabeth A Nguyen, Katherine A Drake, Scott Huntsman, Donglei Hu, et al. Genome-wide association study and admixture mapping identify different asthma-associated loci in latinos: The genes-environments & admixture in latino americans study. *Journal of Allergy and Clinical Immunology*, 134(2):295–305, 2014. 12, 26

[48] Thomas J Hoffmann, Yiping Zhan, Mark N Kvale, Stephanie E Hesselson, Jeremy Gollub, Carlos Iribarren, Yontao Lu, Gangwu Mei, Matthew M Purdy, Charles Quesenberry, et al. Design and coverage of high throughput genotyping arrays optimized for individuals of east asian, african american, and latino race/ethnicity using imputation and a novel hybrid snp selection algorithm. *Genomics*, 98(6):422–430, 2011. 12

[49] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014. 13, 16, 45, 91

[50] Jovana Maksimovic, Lavinia Gordon, Alicia Oshlack, et al. Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. *Genome Biol*, 13(6):R44, 2012. 13

[51] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007. 13

[52] Brenda Eskenazi, Asa Bradman, Eleanor A Gladstone, Selene Jaramillo, Kelly Birch,

and Nina Holland. Chamacos, a longitudinal birth cohort study: lessons from the fields. *Journal of Children's Health*, 1(1):3–27, 2003. 13, 25

[53] Paul Yousefi, Karen Huen, Raul Aguilar Schall, Anna Decker, Emon Elboudwarej, Hong Quach, Lisa Barcellos, and Nina Holland. Considerations for normalization of dna methylation data by illumina 450k beadchip assay in population studies. *Epigenetics*, 8(11):1141–1152, 2013. 13

[54] Hui-Ju Tsai, Shweta Choudhry, Mariam Naqvi, William Rodriguez-Cintron, Esteban González Burchard, and Elad Ziv. Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. *Human genetics*, 118(3-4):424–433, 2005. 13

[55] Karen Huen, Kim Harley, Kenneth Beckman, Brenda Eskenazi, and Nina Holland. Associations of pon1 and genetic ancestry with obesity in early childhood. *PloS one*, 8(5):e62565, 2013. 13, 25

[56] Shweta Choudhry, Margaret Taub, Rui Mei, José Rodriguez-Santana, William Rodriguez-Cintron, Mark D Shriver, Elad Ziv, Neil J Risch, and Esteban González Burchard. Genome-wide screen for asthma in puerto ricans: evidence for association with 5q23 region. *Human genetics*, 123(5):455–468, 2008. 13

[57] Laura Fejerman, Isabelle Romieu, Esther M John, Eduardo Lazcano-Ponce, Scott Huntsman, Kenneth B Beckman, Eliseo J Pérez-Stable, Esteban González Burchard, Elad Ziv, and Gabriela Torres-Mejía. European ancestry is positively associated with breast cancer risk in mexican women. *Cancer Epidemiology Biomarkers & Prevention*, 19(4):1074–1082, 2010. 13

[58] Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, et al. High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, 2011. 14

[59] Adrian Bird, Mary Taggart, Marianne Frommer, Orlando J Miller, and Donald Macleod. A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. *Cell*, 40(1):91–99, 1985. 14

[60] Rafael A Irizarry, Christine Ladd-Acosta, Bo Wen, Zhijin Wu, Carolina Montano, Patrick Onyango, Hengmi Cui, Kevin Gabo, Michael Rongione, Maree Webster, et al. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific cpg island shores. *Nature genetics*, 41(2):178–186, 2009. 14

[61] Louise Laurent, Eleanor Wong, Guoliang Li, Tien Huynh, Aristotelis Tsirigos, Chin Thing Ong, Hwee Meng Low, Ken Wing Kin Sung, Isidore Rigoutsos, Jeanne Loring, et al. Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–331, 2010. 14

[62] Serge Saxonov, Paul Berg, and Douglas L Brutlag. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5):1412–1417, 2006. 14

[63] Madeleine P Ball, Jin Billy Li, Yuan Gao, Je-Hyuk Lee, Emily M LeProust, In-Hyun Park, Bin Xie, George Q Daley, and George M Church. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology*, 27(4):361–368, 2009. 14

[64] Bonnie R Joubert, Siri E Håberg, Roy M Nilsen, Xuting Wang, Stein E Vollset, Susan K Murphy, Zhiqing Huang, Cathrine Hoyo, Øivind Midttun, Lea A Cupul-Uicab, et al. 450k epigenome-wide scan identifies differential dna methylation in newborns related to maternal smoking during pregnancy. *Environmental health perspectives*, 120(10):1425, 2012. 14

[65] Wei Jie Seow, Molly L Kile, Andrea A Baccarelli, Wen-Chi Pan, Hyang-Min Byun, Golam Mostofa, Quazi Quamruzzaman, Mahmuder Rahman, Xihong Lin, and David C Christiani. Epigenome-wide dna methylation changes with development of arsenic-

induced skin lesions in bangladesh: A case–control follow-up study. *Environmental and molecular mutagenesis*, 55(6):449–456, 2014. 14

[66] Katherine J Dick, Christopher P Nelson, Loukia Tsaprouni, Johanna K Sandling, Dylan Aïssi, Simone Wahl, Eshwar Meduri, Pierre-Emmanuel Morange, France Gagnon, Harald Grallert, et al. Dna methylation and body-mass index: a genome-wide analysis. *The Lancet*, 383(9933):1990–1998, 2014. 14

[67] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, et al. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology*, 31(2):142–147, 2013. xix, 14, 44, 45, 47, 53, 54, 70, 72, 90, 91, 94, 95, 104, 105, 120, 130, 141

[68] Alex T Adams, Nicholas A Kennedy, Richard Hansen, Nicholas T Ventham, Kate R O'Leary, Hazel E Drummond, Colin L Noble, Emad El-Omar, Richard K Russell, David C Wilson, et al. Two-stage genome-wide methylation profiling in childhood-onset crohn's disease implicates epigenetic alterations at the vmp1/mir21 and hla loci. *Inflammatory bowel diseases*, 20(10):1784–1793, 2014. 14

[69] Jian-Bing Fan, Kevin L Gunderson, Marina Bibikova, Joanne M Yeakley, Jing Chen, Eliza Wickham Garcia, Lori L Lebruska, Marc Laurent, Richard Shen, and David Barker. [3] illumina universal bead arrays. *Methods in enzymology*, 410:57–73, 2006. 14

[70] Elior Rahmani, Liat Shenhav, Regev Schweiger, Paul Yousefi, Karen Huen, Brenda Eskenazi, Celeste Eng, Scott Huntsman, Donglei Hu, Joshua Galanter, et al. Genome-wide methylation data mirror ancestry information. *Epigenetics & chromatin*, 10(1):1, 2017. 14, 17, 33, 64, 76, 94, 120

[71] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. 15

[72] Andrew E Jaffe and Rafael A Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*, 15(2):R31, 2014. 16, 32, 80, 94, 102

[73] Eugene Andres Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13(1):1, 2012. 16, 23

[74] Lovisa E Reinius, Nathalie Acevedo, Maaike Joerink, Göran Pershagen, Sven-Erik Dahlén, Dario Greco, Cilla Söderhäll, Annika Scheynius, and Juha Kere. Differential dna methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS one*, 7(7):e41361, 2012. 16, 33, 45, 46, 47, 63, 76, 89, 90, 91, 97, 127

[75] Maria Pino-Yanes, Neeta Thakur, Christopher R Gignoux, Joshua M Galanter, Lindsey A Roth, Celeste Eng, Katherine K Nishimura, Sam S Oh, Hita Vora, Scott Huntsman, et al. Genetic ancestry influences asthma susceptibility and lung function among latinos. *Journal of Allergy and Clinical Immunology*, 135(1):228–235, 2015. 18

[76] Joshua M Galanter, Dara Torgerson, Christopher R Gignoux, Saunak Sen, Lindsey A Roth, Marc Via, Melinda C Aldrich, Celeste Eng, Scott Huntsman, Jose Rodriguez-Santana, et al. Cosmopolitan and ethnic-specific replication of genetic risk factors for asthma in 2 latino populations. *Journal of Allergy and Clinical Immunology*, 128(1):37–43, 2011. 20

[77] Paul Yousefi, Karen Huen, Veronica Davé, Lisa Barcellos, Brenda Eskenazi, and Nina Holland. Sex differences in dna methylation assessed by 450 k beadchip in newborns. *BMC genomics*, 16(1):1, 2015. 21, 33, 64, 76

[78] Kelly M Bakulski, Jason I Feinberg, Shan V Andrews, Jack Yang, Shannon Brown, Stephanie L McKenney, Frank Witter, Jeremy Walston, Andrew P Feinberg, and

M Daniele Fallin. Dna methylation of cord blood cell types: Applications for mixed cell birth studies. *Epigenetics*, pages 1–9, 2016. 22

[79] Paul Yousefi, Karen Huen, Hong Quach, Girish Motwani, Alan Hubbard, Brenda Eskenazi, and Nina Holland. Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies. *Environmental and molecular mutagenesis*, 56(9):751–758, 2015. 23, 33, 48, 50, 74

[80] Brenda Eskenazi, Katherine Kogut, Karen Huen, Kim G Harley, Maryse Bouchard, Asa Bradman, Dana Boyd-Barr, Caroline Johnson, and Nina Holland. Organophosphate pesticide exposure, pon1, and neurodevelopment in school-age children from the chamacos study. *Environmental research*, 134:149–157, 2014. 25

[81] Laura Fejerman, Esther M John, Scott Huntsman, Kenny Beckman, Shweta Choudhry, Eliseo Perez-Stable, Esteban González Burchard, and Elad Ziv. Genetic ancestry and risk of breast cancer among us latinas. *Cancer research*, 68(23):9723–9728, 2008. 25

[82] Marcus W Koch, Luanne M Metz, and Olga Kovalchuk. Epigenetic changes in patients with multiple sclerosis. *Nature Reviews Neurology*, 9(1):35–43, 2013. 31

[83] Tempei Ikegame, Miki Bundo, Fumiko Sunaga, Tatsuro Asai, Fumichika Nishimura, Akane Yoshikawa, Yoshiya Kawamura, Hiroyuki Hibino, Mamoru Tochigi, Chihiro Kakiuchi, et al. Dna methylation analysis of bdnf gene promoters in peripheral blood cells of schizophrenia patients. *Neuroscience research*, 77(4):208–214, 2013. 31

[84] Gidon Toperoff, Dvir Aran, Jeremy D Kark, Michael Rosenberg, Tatyana Dubnikov, Batel Nissan, Julio Wainstein, Yechiel Friedlander, Ephrat Levy-Lahad, Benjamin Glaser, et al. Genome-wide survey reveals predisposing diabetes type 2-related dna methylation variations in human peripheral blood. *Human molecular genetics*, 21(2):371–383, 2012. 31

[85] Eugene Andres Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey.

DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 2012. xx, 32, 35, 44, 48, 74, 84, 87, 89, 90, 109, 110, 113, 151

[86] Eugene Andres Houseman, John Molitor, and Carmen J Marsit. Reference-free cell mixture adjustments in analysis of dna methylation data. *Bioinformatics*, 30(10):1431–1439, 2014. 32

[87] James Zou, Christoph Lippert, David Heckerman, Martin Aryee, and Jennifer Listgarten. Epigenome-wide association studies without the need for cell-type composition. *Nature methods*, 11(3):309–11, March 2014. 32, 105

[88] E Andres Houseman, Molly L Kile, David C Christiani, Tan A Ince, Karl T Kelsey, and Carmen J Marsit. Reference-free deconvolution of dna methylation data and mediation by cell composition effects. *BMC bioinformatics*, 17(1):259, 2016. 32, 33, 36, 44, 48

[89] Pavlo Lutsik, Martin Slawski, Gilles Gasparoni, Nikita Vedeneev, Matthias Hein, and Jörn Walter. Medecom: discovery and quantification of latent components of heterogeneous methylomes. *Genome biology*, 18(1):55, 2017. 32, 33, 36, 44, 48

[90] Andrew E Teschendorff, Yang Gao, Allison Jones, Matthias Ruebner, Matthias W Beckmann, David L Wachter, Peter A Fasching, and Martin Widschwendter. Dna methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nature communications*, 7, 2016. 33

[91] Jerry Guintivano, Martin J Aryee, and Zachary A Kaminsky. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*, 8(3):290–302, 2013. 33

[92] Elior Rahmani, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, Esteban G Burchard, Eleazar Eskin, James Zou, et al. Correcting for cell-type heterogeneity in dna methylation: a comprehensive evaluation. *Nature methods*, 14(3):218, 2017. 33, 44, 75, 76, 95

[93] Elior Rahmani, Reut Yedidim, Liat Shenhav, Regev Schweiger, Omer Weissbrod, Noah Zaitlen, and Eran Halperin. Glint: a user-friendly toolset for the analysis of high-throughput dna-methylation array data. *Bioinformatics*, page btx059, 2017. 44, 45, 76, 89, 92, 94, 95

[94] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367, 2013. xix, 44, 45, 47, 53, 54, 70, 72, 90, 91, 94, 105, 127, 130, 141

[95] Eilis Hannon, Emma Dempster, Joana Viana, Joe Burrage, Adam R Smith, Ruby Macdonald, David St Clair, Colette Mustard, Gerome Breen, Sebastian Therman, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential dna methylation. *Genome biology*, 17(1):176, 2016. xix, 44, 47, 53, 54, 70, 72

[96] Devin C Koestler, Meaghan J Jones, Joseph Usset, Brock C Christensen, Rondi A Butler, Michael S Kobor, John K Wiencke, and Karl T Kelsey. Improving cell mixture deconvolution by id entifying optimal dna methylation libraries (idol). *BMC bioinformatics*, 17(1):1, 2016. 44, 48, 74, 89

[97] Benjamin Lehne, Alexander W Drong, Marie Loh, Weihua Zhang, William R Scott, Sian-Tsung Tan, Uzma Afzal, James Scott, Marjo-Riitta Jarvelin, Paul Elliott, et al. A coherent approach for analysis of the illumina humanmethylation450 beadchip improves data quality and performance in epigenome-wide association studies. *Genome biology*, 16(1):37, 2015. 45, 91, 94, 130

[98] Ira S Hofer, Eilon Gabel, Michael Pfeffer, Mohammed Mahbouba, and Aman Mahajan. A systematic approach to creation of a perioperative data warehouse. *Anesthesia & Analgesia*, 122(6):1880–1884, 2016. 46

[99] Andres Cardenas, Catherine Allard, Myriam Doyon, E Andres Houseman, Kelly M

Bakulski, Patrice Perron, Luigi Bouchard, and Marie-France Hivert. Validation of a dna methylation reference panel for the estimation of nucleated cells types in cord blood. *Epigenetics*, 11(11):773–779, 2016. 50

[100] Chi-Yu Lai, Elizabeth Scarr, Madhara Udawela, Ian Everall, Wei J Chen, and Brian Dean. Biomarkers in schizophrenia: a focus on blood based diagnostics and theranostics. *World journal of psychiatry*, 6(1):102, 2016. 67

[101] İbrahim Tekeoğlu, Gönül Gürol, Halil Harman, Engin Karakeçe, and İhsan Hakkı Çiftçi. Overlooked hematological markers of disease activity in rheumatoid arthritis. *International journal of rheumatic diseases*, 19(11):1078–1082, 2016. 67

[102] R Solana, MC Alonso, and J Pena. Natural killer cells in healthy aging. *Experimental gerontology*, 34(3):435–443, 1999. 68

[103] Rafael Solana and Erminia Mariani. Nk and nk/t cells in human senescence. *Vaccine*, 18(16):1613–1620, 2000. 68

[104] Norikuni Kawanaka, Masahiro Yamamura, Tetsushi Aita, Yoshitaka Morita, Akira Okamoto, Masanori Kawashima, Mitsuhiro Iwahashi, Akiko Ueno, Yasukazu Ohmoto, and Hirofumi Makino. Cd14+, cd16+ blood monocytes and joint inflammation in rheumatoid arthritis. *Arthritis & Rheumatology*, 46(10):2578–2586, 2002. 68

[105] S Wijngaarden, JAG Van Roon, JWJ Bijlsma, JGJ Van De Winkel, and FPJG Lafeber. Fcγ receptor expression levels on monocytes are elevated in rheumatoid arthritis patients with high erythrocyte sedimentation rate who do not use anti-rheumatic drugs. *Rheumatology*, 42(5):681–688, 2003. 68

[106] Mitsuhiro Iwahashi, Masahiro Yamamura, Tetsushi Aita, Akira Okamoto, Akiko Ueno, Norio Ogawa, Sachiko Akashi, Kensuke Miyake, Paul J Godowski, and Hirofumi Makino. Expression of toll-like receptor 2 on cd16+ blood monocytes and synovial tissue macrophages in rheumatoid arthritis. *Arthritis & Rheumatology*, 50(5):1457–1467, 2004. 68

[107] Frederico AC Azevedo, Carlos H Andrade-Moraes, Marco R Curado, Ana V Oliveira-Pinto, Daniel M Guimarães, Diego Szczupak, Bruna V Gomes, Ana TL Alho, Livia Polichiso, Edilaine Tampellini, et al. Automatic isotropic fractionation for large-scale quantitative cell analysis of nervous tissue. *Journal of neuroscience methods*, 212(1):72–78, 2013. 74

[108] Alexander R Pinto, Alexei Ilinykh, Malina J Ivey, Jill T Kuwabara, Michelle L DAntoni, Ryan Debuque, Anjana Chandran, Lina Wang, Komal Arora, Nadia A Rosenthal, et al. Revisiting cardiac cellular composition. *Circulation research*, 118(3):400–409, 2016. 74

[109] Adeline Divoux, Joan Tordjman, Danièle Lacasa, Nicolas Veyrie, Danielle Hugol, Abdelhalim Aissat, Arnaud Basdevant, Michèle Guerre-Millo, Christine Poitou, Jean-Daniel Zucker, et al. Fibrosis in human adipose tissue: composition, distribution, and link with lipid metabolism and fat mass loss. *Diabetes*, 59(11):2817–2825, 2010. 74

[110] Peng Lu, Aleksey Nakorchevskiy, and Edward M Marcotte. Expression deconvolution: a reinterpretation of dna microarray data reveals dynamic changes in cell populations. *Proceedings of the National Academy of Sciences*, 100(18):10370–10375, 2003. 76

[111] Alexander R Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS one*, 4(7):e6098, 2009. 76

[112] Alexandre Kuhn, Doris Thu, Henry J Waldvogel, Richard LM Faull, and Ruth Luthi-Carter. Population-specific expression analysis (psea) reveals molecular changes in diseased brain. *Nature methods*, 8(11):945–947, 2011. 76

[113] Neta S Zuckerman, Yair Noam, Andrea J Goldsmith, and Peter P Lee. A self-directed method for cell-type identification and separation of gene expression microarrays. *PLoS Comput Biol*, 9(8):e1003189, 2013. 76

[114] Yael Steuerman and Irit Gat-Viks. Exploiting gene-expression deconvolution to probe the genetics of the immune system. *PLoS Comput Biol*, 12(4):e1004856, 2016. 76

[115] Yoshinori Fukazawa, Haesun Park, Mark J Cameron, Francois Lefebvre, Richard Lum, Noel Coombes, Eisa Mahyari, Shoko I Hagen, Jin Young Bae, Marcelo Delos Reyes III, et al. Lymph node t cell responses predict the efficacy of live attenuated siv vaccines. *Nature medicine*, 18(11):1673, 2012. 78

[116] Amy M Becker, Kathryn H Dao, Bobby Kwanghoon Han, Roger Kornu, Shuchi Lakhanpal, Angela B Mobley, Quan-Zhen Li, Yun Lian, Tianfu Wu, Andreas M Reimold, et al. Sle peripheral blood b cell, t cell and myeloid cell transcriptomes display unique profiles and each subset contributes to the interferon signature. *PloS one*, 8(6):e67003, 2013. 78

[117] George Plitas, Catherine Konopacki, Kenmin Wu, Paula D Bos, Monica Morrow, Ekaterina V Putintseva, Dmitriy M Chudakov, and Alexander Y Rudensky. Regulatory t cells exhibit distinct features in human breast cancer. *Immunity*, 45(5):1122–1134, 2016. 78

[118] Adrian Schwarzer, Stephan Emmrich, Franziska Schmidt, Dominik Beck, Michelle Ng, Christina Reimer, Felix Ferdinand Adams, Sarah Grasedieck, Damian Witte, Sebastian Käbler, et al. The non-coding rna landscape of human hematopoiesis and leukemia. *Nature Communications*, 8, 2017. 78

[119] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486, 2015. 78

[120] Blue B Lake, Rizi Ai, Gwendolyn E Kaeser, Neeraj S Salathia, Yun C Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, et al. Neuronal subtypes

and diversity revealed by single-nucleus rna sequencing of the human brain. *Science*, 352(6293):1586–1590, 2016. 78

[121] Itay Tirosh, Andrew S Venteicher, Christine Hebert, Leah E Escalante, Anoop P Patel, Keren Yizhak, Jonathan M Fisher, Christopher Rodman, Christopher Mount, Mariella G Filbin, et al. Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*, 539(7628):309, 2016. 78

[122] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016. 78

[123] Melina Claussnitzer, Simon N Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S Sousa, Jacqueline L Beaudry, Vijitha Puviindran, et al. Fto obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine*, 373(10):895–907, 2015. 78

[124] Sara Mostafavi, Hideyuki Yoshida, Devapregasan Moodley, Hugo LeBoité, Katherine Rothamel, Towfique Raj, Chun Jimmie Ye, Nicolas Chevrier, Shen-Ying Zhang, Ting Feng, et al. Parsing the interferon transcriptional network and its disease associations. *Cell*, 164(3):564–578, 2016. 78

[125] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, et al. Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome research*, 24(1):14–24, 2014. 79

[126] Fred A Wright, Patrick F Sullivan, Andrew I Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, et al. Heritability and genomics of gene expression in peripheral blood. *Nature genetics*, 46(5):430–437, 2014. 79

[127] Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817, 2014. 79

[128] Omer Schwartzman and Amos Tanay. Single-cell epigenomics: techniques and emerging applications. *Nature Reviews Genetics*, 16(12):716, 2015. 79

[129] Stephen J Clark, Heather J Lee, Sébastien A Smallwood, Gavin Kelsey, and Wolf Reik. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome biology*, 17(1):72, 2016. 79

[130] Christof Angermueller, Stephen J Clark, Heather J Lee, Iain C Macaulay, Mabel J Teng, Tim Xiaoming Hu, Felix Krueger, Sébastien A Smallwood, Chris P Ponting, Thierry Voet, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods*, 13(3):229, 2016. 79

[131] Shai S Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L Bodian, Frank Staedtler, Nicholas M Perry, Trevor Hastie, Minnie M Sarwal, Mark M Davis, and Atul J Butte. Cell type–specific gene expression differences in complex tissues. *Nature methods*, 7(4):287, 2010. 81, 82, 111, 113

[132] Harm-Jan Westra, Danny Arends, Tõnu Esko, Marjolein J Peters, Claudia Schurmann, Katharina Schramm, Johannes Kettunen, Hanieh Yaghootkar, Benjamin P Fairfax, Anand Kumar Andiappan, et al. Cell specific eqtl analysis without sorting cells. *PLoS genetics*, 11(5):e1005223, 2015. 81, 111, 114

[133] Shijie C Zheng, Charles E Breeze, Stephan Beck, and Andrew E Teschendorff. Identification of differentially methylated cell types in epigenome-wide association studies. *Nature methods*, 15(12):1059, 2018. 82, 96, 98, 111, 115, 131, 132

[134] Elior Rahmani, Regev Schweiger, Liat Shenhav, Theodora Wingert, Ira Hofer, Eilon Gabel, Eleazar Eskin, and Eran Halperin. Bayescce: a bayesian framework for esti-

mating cell-type composition from dna methylation without the need for methylation reference. *Genome biology*, 19(1):141, 2018. 84, 87, 90, 109, 110, 113, 151

[135] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, 2010. 85

[136] Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33:245, 2003. 89

[137] Shicheng Guo, Qi Zhu, Ting Jiang, Rongsheng Wang, Yi Shen, Xiao Zhu, Yan Wang, Fengmin Bai, Qin Ding, Xiaodong Zhou, et al. Genome-wide dna methylation patterns in cd4+ t cells from chinese han patients with rheumatoid arthritis. *Modern rheumatology*, 27(3):441–447, 2017. 91, 107

[138] Brooke Rhead, Calliope Holingue, Michael Cole, Xiaorong Shao, Hong L Quach, Diana Quach, Khooshbu Shah, Elizabeth Sinclair, John Graf, Thomas Link, et al. Rheumatoid arthritis naive t cells share hypermethylation sites with synoviocytes. *Arthritis & Rheumatology*, 69(3):550–559, 2017. 91, 107

[139] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012. 94

[140] Elior Rahmani, Regev Schweiger, Brooke Rhead, Lindsey A Criswell, Lisa F Barcellos, Eleazar Eskin, Saharon Rosset, Sriram Sankararaman, and Eran Halperin. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature communications*, 10(1):1–11, 2019. 94, 96, 105, 106, 107, 108, 116, 117, 121, 125, 131, 141

[141] Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, et al. The genecards

suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics*, 54(1):1–30, 2016. 94, 96

[142] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901, 2016. 94, 96

[143] Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, et al. Genehancer: genome-wide integration of enhancers and target genes in genecards. *Database*, 2017(1), 2017. 94, 96

[144] Belinda Phipson, Jovana Maksimovic, and Alicia Oshlack. missmethyl: an r package for analyzing data from illuminas humanmethylation450 platform. *Bioinformatics*, 32(2):286–288, 2015. 95

[145] Xiangyu Luo, Can Yang, and Yingying Wei. Detection of cell-type-specific risk-cpg sites in epigenome-wide association studies. *Preprint at https://www.biorxiv.org/content/10.1101/415109v1*, 2018. 96, 109, 110

[146] Tibor T Glant, Katalin Mikecz, and Tibor A Rauch. Epigenetics in the pathogenesis of rheumatoid arthritis. *BMC medicine*, 12(1):35, 2014. 105

[147] Adam Cribbs, Marc Feldmann, and Udo Oppermann. Towards an understanding of the role of dna methylation in rheumatoid arthritis: therapeutic and diagnostic implications. *Therapeutic advances in musculoskeletal disease*, 7(5):206–219, 2015. 105

[148] María C de Andres, Eva Perez-Pampin, Manuel Calaza, Francisco J Santaclara, Ignacio Ortea, Juan J Gomez-Reino, and Antonio Gonzalez. Assessment of global dna methylation in peripheral blood cell subpopulations of early rheumatoid arthritis before and after methotrexate. *Arthritis research & therapy*, 17(1):233, 2015. 107

[149] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228, 2015. 109

[150] Xingjie Hao, Ping Zeng, Shujun Zhang, and Xiang Zhou. Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS genetics*, 14(1):e1007186, 2018. 109

[151] Isabel Mendizabal, Stefano Berto, Noriyoshi Usui, Kazuya Toriumi, Paramita Chatterjee, Connor Douglas, Iksoo Huh, Hyeonsoo Jeong, Thomas Layman, Carol A Tamminga, et al. Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome biology*, 20(1):135, 2019. 111, 115, 131

[152] Ziyi Li, Zhijin Wu, Peng Jin, and Hao Wu. Dissecting differential signals in high-throughput data from complex tissues. *Bioinformatics*, 35(20):3898–3905, 2019. 111, 115, 131

[153] Xiangyu Luo, Can Yang, and Yingying Wei. Detection of cell-type-specific risk-cpg sites in epigenome-wide association studies. *Nature communications*, 10(1):1–12, 2019. 112, 117, 131

[154] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982. 126

[155] Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996. 126

[156] Han Jing, Shijie C Zheng, Charles E Breeze, Stephan Beck, and Andrew E Teschendorff. Calling differential dna methylation at cell-type resolution: an objective status-quo. *BioRxiv*, page 822940, 2019. 127

[157] Andrew E Teschendorff, Charles E Breeze, Shijie C Zheng, and Stephan Beck. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. *BMC bioinformatics*, 18(1):105, 2017. 127, 131

[158] Dan Su, Xuting Wang, Michelle R Campbell, Devin K Porter, Gary S Pittman, Brian D Bennett, Ma Wan, Neal A Englert, Christopher L Crowl, Ryan N Gimple, et al. Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. *PloS one*, 11(12):e0166486, 2016. xxviii, 130, 141, 142

[159] Mario Bauer. Cell-type-specific disturbance of dna methylation pattern: a chance to get more benefit from and to minimize cohorts for epigenome-wide association studies. *International Journal of Epidemiology*, 47(3):917–927, 2018. 131

[160] Jerry L Jensen. *Statistics for petroleum engineers and geoscientists*, volume 2. Gulf Professional Publishing, 2000. 147