Title
Identifying genomic regulatory patterns underlying complex phenotypes from heterogeneous data

Author
Jew, Brandon

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Identifying genomic regulatory patterns underlying complex phenotypes from heterogeneous data

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

by

Brandon Jew

2021

ABSTRACT OF THE DISSERTATION

Identifying genomic regulatory patterns
underlying complex phenotypes from heterogeneous data

by

Brandon Jew
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2021
Professor Eran Halperin, Co-Chair
Professor Jae Hoon Sul, Co-Chair

Large-scale transcriptomic datasets provide valuable opportunities to better understand the regulation of gene expression and its role in human health. However, these studies can be confounded by issues such as cell type heterogeneity. Furthermore, these datasets are growing extremely large with complex study designs, such as gene expression measured across a multitude of tissues, that must be accurately and efficiently modeled. Finally, better understanding of the mechanisms that influence gene regulation are required to integrate novel associations with biological understanding. In this dissertation, we introduce methods that address these issues in the analysis of tissue-level gene expression data. We present a method to accurately estimate cell type composition from these data by integrating single-cell information, as well as a scalable approach to model multi-tissue expression datasets and identify expression quantitative trait loci. We also present analyses of bulk expression data that support a hypothesized mechanism of gene regulation that occurs in the general female population through non-random X chromosome inactivation. The work presented in this dissertation allows researchers to perform efficient and accurate analyses of gene expression data and provides additional insight into the mechanisms that underlie associations between genetics, transcriptomics, and complex phenotypes.

The dissertation of Brandon Jew is approved.

Päivi Pajukanta

Sriram Sankararaman

Eran Halperin, Committee Co-Chair

Jae Hoon Sul, Committee Co-Chair

University of California, Los Angeles

2021

To my wife,

my constant source of happiness and inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I thank my advisors, Eran Halperin and Jae Hoon Sul. I am extremely grateful to have both as amazing mentors who provided me with exciting projects and opportunities to grow as a researcher. They both fostered welcoming and top-notch research groups that I am proud to be a part of. The *Halperin Lab* fleece jacket will always be a staple part of my wardrobe and the *Sul Lab* mug will always be featured on my desk.

I would also like to thank Eleazar Eskin, who roped me into bioinformatics in the first place and supported me through my PhD. Much of my work was done with Päivi Pajukanta, Noah Zaitlen, and Sriram Sankararaman. I am extremely grateful for their collaboration and mentorship. I thank Lana Martin and Robert Smith, who helped me immensely with funding and outreach projects during their time in Eleazar's lab. I thank Sim-Lin Lau and Margaret Chu for their amazing support.

Many thanks to my lab mates in Eran's lab (Elior, Mike, Leah, Johnson, Liat, Ulzee, Brian, Nadav, Jeff) and Jae Hoon's Lab (Albert, Sarah, Lingyu). Also, thanks to our shared lab space friends from Sriram's lab (Alec, Chris, Ruthie, Arun, Ariel, Ali). I really enjoyed talking too much during work hours and tracking down every opportunity for free food.

Finally, I would like to thank my family for their unwavering support. I thank my wife, the amazing Dr. Maegan Lu. Over the past 8 years, she's been a constant source of support and inspiration to move forward.

Chapter Two of this dissertation is a version of Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M. Garske, Jae Hoon Sul, Kirsi H. Pietiläinen, Päivi Pajukanta, Eran Halperin. "Accurate estimation of cell composition in bulk expression through robust integration of single-cell information". Nature Communications, 11(1), 1-11, 2020.

Chapter Three is a version of Brandon Jew, Jiajin Li, Sriram Sankararaman, Jae Hoon Sul. "An efficient linear mixed model framework for meta-analytic association studies across

multiple contexts". The 21st Workshop on Algorithms in Bioinformatics (WABI), LIPIcs 10:1 - 10:18, 2021, In press.

Chapter Four is a version of a manuscript in preparation by Brandon Jew, Jae Hoon Sul, Noah Zaitlen. "Selection contributes to skewed X chromosome inactivation across human tissues".

This work was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650604.

VITA

2017-2021	PhD, Bioinformatics, University of California, Los Angeles, CA, USA
2018-2021	Graduate Research Fellow, National Science Foundation
2013–2017	BS, Biochemistry (minor in Bioinformatics), University of California, Los Angeles, CA, USA
2014-2016	Research Assistant, Loo Laboratory, University of California, Los Angeles, CA, USA
2014	Research Intern, Evans Laboratory, Salk Institute for Biological Studies

PUBLICATIONS AND PRESENTATIONS

* Denotes equal contribution

Aditya Gorla, **Brandon Jew**, Luke Zhang, Jae Hoon Sul. xGAP: A python based efficient, modular, extensible and fault tolerant genomic analysis pipeline for variant discovery. *Bioinformatics*. 2021.

David Goodman-Meza, Akos Rudas, Jeffrey N Chiang, Paul C Adamson, Joseph Ebinger, Nancy Sun, Patrick Botting, Jennifer A Fulcher, Faysal G Saab, Rachel Brook, Eleazar Eskin, Ulzee An, Misagh Kordi, **Brandon Jew**, Brunilda Balliu, Zeyuan Chen, Brian L Hill, Elior Rahmani, Eran Halperin, Vladimir Manuel. A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity. *PLoS One*. 2020.

Zong Miao, Marcus Alvarez, Arthur Ko, Yash Bhagat, Elior Rahmani, **Brandon Jew**, Sini Heinonen, Karen L Mohlke, Markku Laakso, Kirsi H. Pietiläinen, Eran Halperin, Päivi Pajukanta. The causal effect of obesity on prediabetes and insulin resistance reveals the important role of adipose tissue in insulin resistance. *PLoS Genetics.* 2020.

Marcus Alvarez*, Elior Rahmani*, **Brandon Jew**, Kristina M Garske, Zong Miao, Jihane N Benhammou, Chun Jimmie Ye, Joseph R Pisegna, Kirsi H Pietiläinen, Eran Halperin, Päivi Pajukanta. Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *Scientific Reports.* 2020.

Brandon Jew*, Marcus Alvarez*, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M. Garske, Jae Hoon Sul, Kirsi H. Pietiläinen, Päivi Pajukanta, Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications.* 2020.

Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M. Garske, Jae Hoon Sul, Kirsi H. Pietiläinen, Päivi Pajukanta, Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Platform presented at the 69th Annual Meeting of the American Society of Human Genetics, October 19, 2019, Houston, Texas.

Jennifer Zou, Farhad Hormozdiari, **Brandon Jew**, Stephane E Castel, Tuuli Lappalainen, Jason Ernst, Jae Hoon Sul, Eleazar Eskin. Leveraging allelic imbalance to refine fine-mapping for eQTL studies. *PLoS Genetics.* 2019.

Brian L Hill, Robert Brown, Eilon Gabel, Nadav Rakocz, Christine Lee, Maxime Cannesson, Pierre Baldi, Loes Olde Loohuis, Ruth Johnson, **Brandon Jew**, Uri Maoz, Aman Mahajan, Sriram Sankararaman, Ira Hofer, Eran Halperin. An automated machine learning-

based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. *British Journal of Anaesthesia.* 2019.

Jiajin Li, **Brandon Jew**, Lingyu Zhan, Sungoo Hwang, Giovanni Coppola, Nelson B Freimer, Jae Hoon Sul. ForestQC: quality control on genetic variants from next-generation sequencing data using random forest. *PLoS Computational Biology.* 2019.

Brandon Jew, Jae Hoon Sul. Variant calling and quality control of large-scale human genome sequencing data. *Emerging Topics in Life Sciences.* 2019.

CHAPTER 1

Introduction

1.1 Scope of Research

Advances in sequencing technologies have paved the way for the generation of large-scale genomic datasets capturing genetic and transcriptomic variation across diverse sets of individuals [1, 2]. These data allow researchers to elucidate the relationship between genetics, gene expression, and complex phenotypes [3]. Genome-wide association studies have identified a vast number of genetic loci associated with different traits [4] and the identification of expression quantitative trait loci (eQTLs) bridges the gap between genotype and phenotype by elucidating the effects of genetic variation on gene expression [5]. These associations serve as stepping stones for better understanding of human health and potential therapeutic targets [6, 7]. However, studies involving tissue-level gene expression are hindered by confounding factors that can both decrease power to detect associations and cause spurious results.

Many tissues in the human body consist of several cell types. For example, brain tissue often contains an assortment of neuronal and glial cell populations, each with distinct functions and gene expression profiles. RNA sequencing (RNA-seq) is often performed on whole tissues and analyses of these data can be confounded by cell-type heterogeneity across samples [8]. Differences in cell type composition can be misinterpreted as differences in gene expression levels when measuring RNA across tissues. Furthermore, these mixtures can obscure cell-type-specific gene regulation that may be of interest [9, 10]. Single-cell approaches, such as single-cell RNA-seq (scRNA-seq) and single-nucleus RNA-seq (snRNA-seq), avoid these issues and provide insight into the gene expression of individual cells. However, these

experiments remain costly, noisy, and difficult to scale compared to bulk RNA-seq [11]. Given the continuing utility of tissue-level RNA-seq datasets, it is important to accurately model variability in cell type proportions in these samples.

Regulation of gene expression also varies significantly across tissues [2]. Moreover, several tissues can have distinct roles in biological systems underlying a complex phenotype. For example, analyses of Type II diabetes have observed transcriptomic changes in blood [12], adipose [13], and brain [14]. Therefore, it is imperative to measure gene expression across many tissues to fully understand gene regulation and its role in traits of interest. Analyzing gene expression across tissues, especially with growing sample sizes, requires approaches that can model this heterogeneity both accurately and efficiently.

Finally, the relationship between gene expression and genetic variation is highly complex [15] and remains to be fully understood. Many eQTLs have been identified and are thought to influence gene expression through interactions with regulatory regions of the genome, such as enhancers and promoters [16]. Alternative mechanisms of genetic influences on gene regulation have been hypothesized, specifically in the context of selective pressures causing non-random X chromosome inactivation [17]. To better understand human biology, it is important to validate these hypotheses and identify the extent of their influence on regulation of gene expression.

1.2 Contributions and Overview

In this dissertation, we introduce computational and statistical methods to accurately model tissue-level expression data with heterogeneity in cell composition and tissue identity. Furthermore, we present analyses of expression data that yields further insight into non-random X chromosome inactivation as an additional mechanism for gene regulation with several interesting implications in the context of X-linked phenotypes.

There is large interest in estimating cell type proportions from tissue-level RNA-seq data. These estimates allow researchers to account for this potential confounding factor in analy-

ses of gene expression data, such as eQTL and differential expression studies. Furthermore, cell type proportions can be used with additional methods to identify cell-type-specific associations from tissue-level data [10]. In Chapter 2, we describe Bisque, an approach for accurately estimating cell type proportions from bulk RNA-seq data. This method utilizes available single-cell data as a reference for cell-type-specific expression while accounting for technological biases between bulk and single-cell RNA-seq.

In Chapter 3, we present an efficient linear mixed model (LMM) for the analysis of massive datasets measuring multiple contexts across a set of individuals. This method, mcLMM, models all contexts jointly in a meta-analytic framework to improve power to detect associations. It can be applied to large expression datasets with measurements across several tissues, such as the GTEx dataset [2], to efficiently identify genetic variants that influence gene regulation in specific tissues or across several tissues. We further demonstrate the utility of this method by performing a multi-trait genome-wide association study across hundreds of thousands of individuals in the UK Biobank [1] using minimal computational resources.

In Chapter 4, we analyze the GTEx dataset [2] to support the hypothesis of selective pressures influencing non-random X chromosome inactivation [17] and quantify the extent of this effect in the female population. We estimate skewing in X inactivation from bulk RNA-seq data measured across non-diseased tissue samples and identify several genetic factors that are significantly associated with preferential inactivation of a haplotype, such as variation associated with increased deleteriousness and decreased proliferation. Furthermore, we identify common genetic variants in specific loci that contribute to skewed X chromosome inactivation. We highlight the implications of non-random skew in X chromosome inactivation, such as decreased penetrance or dampened effects of genetic variation on preferentially inactivated haplotypes.

CHAPTER 2

Accurate estimation of cell composition in bulk expression through robust integration of single-cell information

2.1 Background

Bulk RNA-seq experiments typically measure total gene expression from heterogeneous tissues, such as tumor and blood samples [18, 19]. Variability in cell type composition can significantly confound analyses of these data, such as in identification of expression quantitative trait loci (eQTLs) or differentially expressed genes [20]. Cell type heterogeneity may also be of interest in profiling changes in tissue composition associated with disease, such as cancer [21] or diabetes [22]. In addition, measures of cell composition can be leveraged to identify cell-specific eQTLs [9, 10] or differential expression [10] from bulk data.

Traditional methods for determining cell type composition, such as immunohistochemistry or flow cytometry, rely on a limited set of molecular markers and lack in scalability relative to the current rate of data generation [23]. Single-cell technologies provide a high-resolution view into cellular heterogeneity and cell type-specific expression [24, 25, 26]. However, these experiments remain costly and noisy compared to bulk RNA-seq [27]. Collection of bulk expression data remains an attractive approach for identifying population-level associations, such as differential expression regardless of cell type specificity. Moreover, many bulk RNA-seq studies that have been performed in recent years resulted in a large body of data that is available in public databases such as dbGAP and GEO. Given the wide availability of these bulk data, the estimation of cell type proportions, often termed decomposition,

can be used to extract large-scale cell type specific information.

There exist a number of methods for decomposing bulk expression, many of which are regression-based and leverage cell type-specific expression data as a reference profile [28]. CIBERSORT [29] is a SVM-regression based approach, originally designed for microarray data, that utilizes a reference generated from purified cell populations. A major limitation of this approach is the reliance on sorting cells to estimate a reference gene expression panel. BSEQ-sc [30] instead generates a reference profile from single-cell expression data that is used in the CIBERSORT model. MuSiC [31] also leverages single-cell expression as a reference, instead using a weighted non-negative least squares regression (NNLS) model for decomposition, with improved performance over BSEQ-sc in several datasets.

The distinct nature of the technologies used to generate bulk and single-cell sequencing data may present an issue for decomposition models that assume a direct proportional relationship between the single-cell-based reference and observed bulk mixture. For example, the capture of mRNA and chemistry of library preparation can differ significantly between bulk tissue and single-cell RNA-seq methods, as well as between different single-cell technologies [32, 33]. Moreover, some technologies may be measuring different parts of the transcriptome, such as nuclear pre-mRNA in single-nucleus RNA-seq (snRNA-seq) experiments as opposed to cellular and extra-cellular mRNA observed in traditional bulk RNA-seq experiments. As we show later, these differences may introduce gene-specific biases that break down the correlation between cell type-specific and bulk tissue measurements. Thus, while single-cell RNA-seq technologies have provided unprecedented resolution in identifying expression profiles of cell types in heterogeneous tissues, these profiles generally may not follow the direct proportionality assumptions of regression-based methods, as we demonstrate here.

We present Bisque, a highly efficient tool to measure cellular heterogeneity in bulk expression through robust integration of single-cell information, accounting for biases introduced in the single-cell sequencing protocols. The goal of Bisque is to integrate the different chemistries/technologies of single-cell and bulk tissue RNA-seq to estimate cell type propor-

tions from tissue-level gene expression measurements across a larger set of samples. Our reference-based model decomposes bulk samples using a single-cell-based reference profile and, while not required, can leverage single-cell and bulk measurements for the same samples for further improved decomposition accuracy. This approach employs gene-specific transformations of bulk expression to account for biases in sequencing technologies as described above. When a reference profile is not available, we propose BisqueMarker, a semi-supervised model that extracts trends in cellular composition from normalized bulk expression samples using only cell-specific marker genes that could be obtained using single cell data. We demonstrate using simulated and real datasets from brain and adipose tissue that our method is significantly more accurate than existing methods. Furthermore, it is extremely efficient, requiring seconds in cases where other methods require hours; thus, it is scalable to large genomic datasets that are now becoming available.

2.2 Methods

2.2.1 Processing bulk expression data

Paired-end reads were aligned with STAR v2.5.1 using default options. Gene counts were quantified using featureCounts v1.6.3. For featureCounts, fragments were counted at the gene-name level. Alignment and gene counts were generated against the GRCh38.p12 genome assembly. STAR v2.5.1 and GRCh38.p12 were included with CellRanger 3.0.2, which was used to process the single-nucleus data.

2.2.2 Processing single-nucleus expression data

Reads from single nuclei sequenced on the 10x Genomics Chromium platform were aligned and quantified using the CellRanger 3.0.2 count function against the GRCh38.p12 genome assembly. To account for reads aligning to both exonic and intronic regions, each gene transcript in this reference assembly was relabeled as an exon since CellRanger counts exonic reads only. We perform this additional step since snRNA-seq captures both mature mRNA

and pre-mRNA, the latter of which includes intronic regions.

After aggregating each single-nucleus sample with the CellRanger aggr function, the full dataset was processed using Seurat v3.0.0 [34]. The data were initially filtered for genes expressed in at least 3 cells and filtered for cells with reads quantified for between 200 and 2,500 genes. We further filtered for cells that had a percentage of counts coming from mitochondrial genes less than or equal to 5 percent. The data were normalized, scaled, and corrected for mitochondrial read percentages with sctransform v0.2.0 [35] using default options.

To identify clusters, Seurat employs a shared nearest neighbor approach. We identified clusters using the top 10 principal components of the processed expression data with resolution set at 0.2. The resolution parameter controls the number of clusters that will be identified, and suggested values vary depending on the size and quality of the dataset. We chose a value that produced 6 clusters in the adipose dataset and 13 clusters in the DLPFC dataset and visualized the clustering results with UMAP [36].

Marker genes were identified by determining the average log-fold change of expression of each cluster compared to the rest of the cells. We identified marker genes as those with an average log-fold change above 0.25. The significance of the differential expression of these genes was determined using a Wilcoxon rank sum test. Only genes that were detected in at least 25 percent of cells were considered. Clusters with many mitochondrial genes as markers (nine genes detected in both datasets) were removed from both datasets. In addition, a cluster with only three marker genes was removed from the DLPFC datasets. Finally, we remove mitochondrial genes from the list of marker genes for decomposition as we assume reads aligning to the mitochondrial genome originate from extra-nuclear RNA in the snRNA-seq dataset (targeting nuclear RNA).

Clusters were labeled by considering cell types associated with the identified marker genes. Marker genes were downloaded from PanglaoDB [37] and filtered for entries validated in human cells. For each gene, we count the possible cell type labels. Each cluster was labeled as the most frequent cell type across all of its marker genes, with each label associated with

a gene weighted by the average log-fold change. If multiple clusters share a cell type label, we consider each cluster a subtype of this label.

Exon-aligned reads were processed in the same exact procedure but snRNA-seq data was aligned to just exonic regions. Cluster names were manually changed for both datasets when aligned to exons to match the clusters from intronic reads as well. Specifically, for clusters identified in the exonic data not found in the full data, we relabeled as the label with the highest score found in the full data. These relabeled clusters were similar in proportion to the corresponding cluster in the full dataset.

2.2.3 Learning a single-cell based reference and bulk transformation for reference-based decomposition

We assume that only a subset of genes are relevant for estimating cell type composition. For the adipose and DLPFC datasets, we selected the marker genes identified by Seurat as described previously. Moreover, we filter out genes with zero variance in the single-cell data, unexpressed genes in the bulk expression, and mitochondrial genes. We convert the remaining gene counts to counts-per-million to account for variable sequencing depth. For m genes and k cell types, a reference profile $Z \in \mathbb{R}^{m \times k}$ is generated by averaging relative abundances within each cell type across the entire single-cell dataset.

Though there is a strong positive correlation between bulk and single-cell based pseudo-bulk (summed single-cell counts) expression data, we observe that the relationship is not one-to-one and varies between genes. This behavior indicates that the distribution of observed bulk expression may significantly differ from the distribution of the single-cell profile weighted by cell proportions. We propose transforming the bulk data to maximize the global linear relationship across all genes for improved decomposition. Our goal is to recover a one-to-one relationship between the transformed bulk and expected convolutions of the reference profile based on single-cell based estimates of cell proportions. This transformed bulk expression better satisfies the assumptions of regression-based approaches under sum-to-one constraints.

Cell type proportions $p \in \mathbb{R}^{k \times n'}$ are determined by counting the cells with each label in

the single-cell data for individuals. Given these proportions and the reference profile Z, we calculate the pseudo-bulk for the single-cell samples as

$$Y = Zp \tag{2.1}$$

where $Y \in \mathbb{R}^{m \times n'}$. For each gene j, our goal is to transform the observed bulk expression across all n bulk samples $X_j \in \mathbb{R}^n$ to match the mean and variance of $Y_j \in \mathbb{R}^{n'}$; hence, the transformation of X_j will be a linear transformation.

If individuals with both single-cell and bulk expression are available, we fit a linear regression model to learn this transformation. Let X'_j denote the expression values for these n' overlapping individuals. We fit the following model (with an intercept) and apply the model to the remaining bulk samples as our transformation:

$$Y_j = \beta_j X'_j + \epsilon_j \tag{2.2}$$

If there are no single-cell samples that have bulk expression available, we assume that the observed mean of Y_j is the true mean of our goal distribution for the transformed X_j. We further assume that the sample variance observed in Y_j is larger than the true variance of the goal distribution, since the number of single-cell samples is typically small. We use a shrinkage estimator of the sample variance of Y_j that minimizes the mean squared error and results in a smaller variance than the unbiased estimator:

$$\hat{\sigma}_j^2 = \frac{1}{n'+1} \sum_{i=1}^{n'} (Y_{i,j} - \bar{Y_j})^2 \tag{2.3}$$

We transform the remaining bulk as follows:

$$X_{j,transformed} = \frac{X_j - \bar{X}_j}{\sigma_{X_j}} \hat{\sigma}_j + \bar{Y}_j \tag{2.4}$$

where a bar indicates the mean value of the observed data and σ_{X_j} is the unbiased sample standard deviation of X_j.

To estimate cell type proportions, we apply non-negative least squares regression with an additional sum-to-one constraint to the transformed bulk data. For individual i, we minimize the following with respect to the cell proportion estimate p_i:

$$\|Zp_i - X_{i,transformed}\|_2 \text{ s.t. } p_i \geq 0, \sum p_i = 1 \tag{2.5}$$

2.2.4 Simulating bulk expression based on single-nucleus counts

We simulate the base bulk expression as the sum of all counts across cells/nuclei sequenced from an individual. To introduce gene-specific variation between the bulk and single-cell data, we sample a coefficient β_j and an intercept α_j from a half-normal (HN) distributions:

$$\beta_j \sim HN(1, \sigma) \tag{2.6}$$

$$\alpha_j \sim HN(0, \sigma) \tag{2.7}$$

where the variance of the HN distribution is $\sigma^2(1 - \frac{2}{\pi})$. At $\sigma = 0$, the base simulated bulk expression remains unchanged. We used a HN distribution to ensure coefficients and intercepts are positive. While our method can handle negative coefficients, this simulation model assumes expression levels have a positive correlation across technologies. We performed 10 replicates of this data-generating process at each σ in 0, 5, 10, 20. Decomposition performance on these data were measured in terms of global R and RMSD and plotted with 95% confidence intervals based on bootstrapping.

2.2.5 Determining significance of cell proportion associations with measured phenotypes

Reported associations were measured in terms of Spearman correlation. To determine the statistical significance of these associations while accounting for possible confounding factors, we applied two approaches. For the adipose dataset, which consisted entirely of twin pairs, we applied a linear mixed-effects model (R nlme package) with random effects accounting

for family. For the DLPFC dataset, we assumed individuals were unrelated and fit a simple linear model (R base package). In each model, we include cell type proportion, age, age-squared, and sex as covariates. We introduced an additional covariate for diabetes status when regressing the Matsuda index due to a known significant association between these two variables. We test whether the cell proportion effect estimates deviate significantly from 0 using a t-test. Each R method implements the described model fitting and significance testing.

2.2.6 Estimating relative cellular heterogeneity with a semi-supervised weighted PCA model

In order to estimate cell type proportions across individuals without the use of a cell-type-specific gene expression panel as reference, we use a weighted PCA approach. BisqueMarker requires a set of marker genes for each cell type as well as the specificity of each marker determined by the fold change from a differential expression analysis. Typical single-cell RNA-seq workflows calculate marker genes and provide both p-values and fold changes, as in Seurat [34]. For each cell type, we take statistically significant marker genes (FDR < 0.05) ranked by p-value. A weighted PCA is calculated on the expression matrix using a subset of the marker genes by first scaling the expression matrix and multiplying each gene column by its weight (the log fold-change) XW, where X is the sample by gene expression matrix and W is a diagonal matrix with entries equal to log fold-change of the corresponding gene. The bulk expression X should be corrected for global covariates so that the proportion estimates do not reflect this global variation. The first PC calculated from XW is used as the estimate of the cell type proportion. This allows cell type-specific genes to be prioritized over more broadly expressed genes. Alternatively, if weights are not available, PCA can be run on the matrix X and the first PC can be used.

In order to select marker genes, we iteratively run the above PCA procedure on a specified range of markers (from 25 to 200) and calculate the ratio of the first eigenvalue to the second. We then select the number of marker genes to use that maximizes this ratio. This procedure

is similar to other methods which select the number of markers to use by maximizing the condition number of the reference matrix [28].

2.3 Results

2.3.1 Overview of Bisque

A graphical overview of Bisque is presented in Figure 2.1. Our reference-based decomposition model requires bulk RNA-seq counts data and a reference dataset with read counts from single-cell RNA-seq. In addition, the single-cell data should be labeled with cell types to be quantified. A reference profile is generated by averaging read count abundances within each cell type in the single-cell data. Given the reference profile and cell proportions observed in the single-cell data, our method learns gene-specific transformations of the bulk data to account for technical biases between the sequencing technologies. Bisque can then estimate cell proportions from the bulk RNA-seq data using the reference and the transformed bulk expression data using non-negative least-squares (NNLS) regression.

2.3.2 Evaluation of decomposition performance in adipose tissue

We applied our method to 106 bulk RNA-seq subcutaneous adipose tissue samples collected from both lean and obese individuals, where 6 samples have both bulk RNA-seq and snRNA-seq data available (Table 2.1). Adipose tissue consists of several cell types, including adipocytes which are expected to be the most abundant population. Adipose tissue also contains structural cell types (i.e. fibroblasts and endothelial cells) and immune cells (i.e. macrophages and T cells) [38]. These 5 cell type populations were identified from the snRNA-seq data (Figure 2.2a).

We observed significant biases between the snRNA-seq and bulk RNA-seq data in samples that had both data available. We found that the linear relationship between the pseudo-bulk (summed snRNA-seq reads across cells) and the true bulk expression varied significantly by each gene (Figure 2.4a). Specifically, we observed best fit lines relating these expression

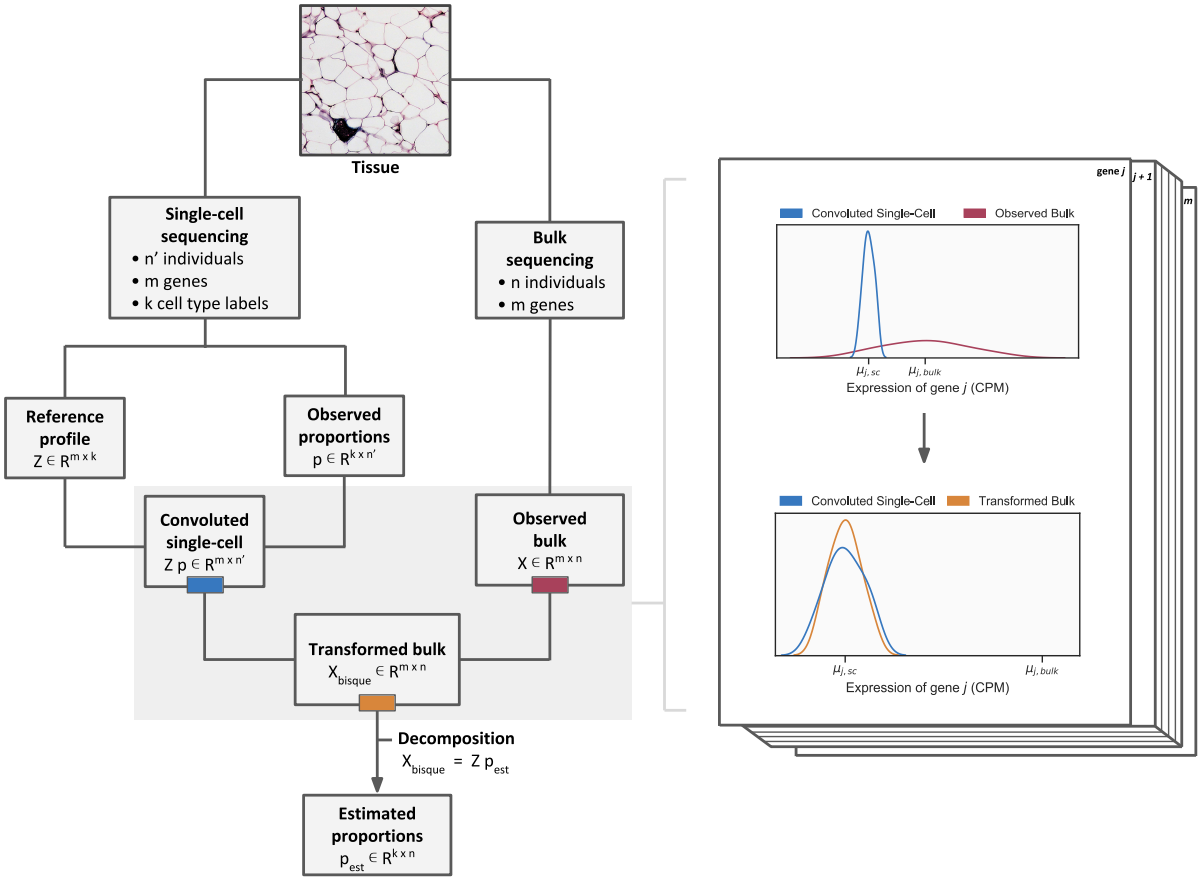

Figure 2.1: Graphical overview of the Bisque decomposition method. We integrate single-cell and bulk expression by learning gene-specific bulk transformations (pictured on right) that align the two datasets for accurate decomposition.

Tissue	Number of Samples	Bulk RNA-seq platform	snRNA-seq platform	snRNA-seq samples	Total nuclei	Average nuclei per individual	Number of cell types
Subcutaneous adipose	106	Illumina NovaSeq	10x Genomics Chromium	6	10,947	1,824	5
Dorsolateral prefrontal cortex	636	Illumina HiSeq	10x Genomics Chromium	8	68,028	8,503	11

Table 2.1: Summary of snRNA-seq and bulk expression datasets used for benchmarking Bisque and existing methods.

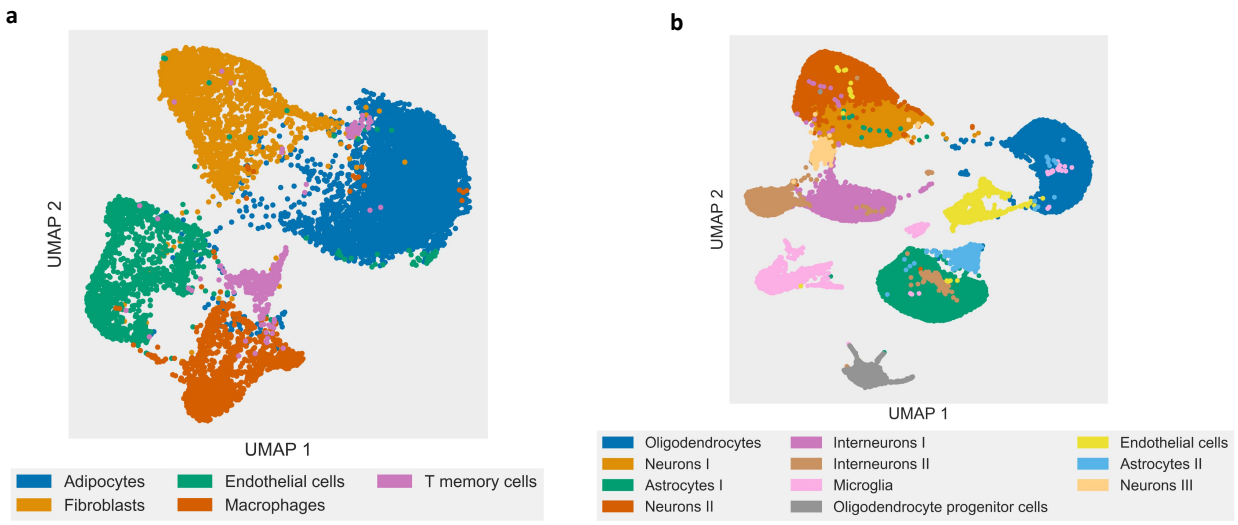

Figure 2.2: Cell types quantified in snRNA-seq experiments. (**a**) UMAP projection of adipose snRNA-seq data with 5 identified cell type clusters labeled. (**b**) UMAP projection of cortex snRNA-seq data with 11 identified clusters.

levels between technologies with a mean slope of roughly 0.30 and a variance in slope of 5.67. In our model, a slope of 1 would indicate no bias between technologies. We further investigated whether gene expression differences between the bulk and snRNA-seq were the same across individuals and experiments. Comparing log-ratios of RNA-seq to snRNA-seq expression levels, we found that the majority of gene biases were preserved across individuals, tissues, and experiments (R=0.75 across experiments) (Figure 2.3), providing evidence that technological differences drive consistent gene expression differences across bulk and snRNA-seq methods.

We performed simulations based on the adipose snRNA-seq data to demonstrate the effect of technology-based biases between the reference profile and bulk expression on decomposition performance. In these analyses, we benchmarked Bisque and three existing methods (MuSiC, BSEQ-sc, and CIBERSORT). Briefly, we simulated bulk expression for 6 individuals by summing the observed snRNA-seq read counts. To model discordance between the reference and bulk, we applied gene-specific linear transformations of the simulated bulk expression. For each gene, the coefficient and intercept of the linear transformation were sam-

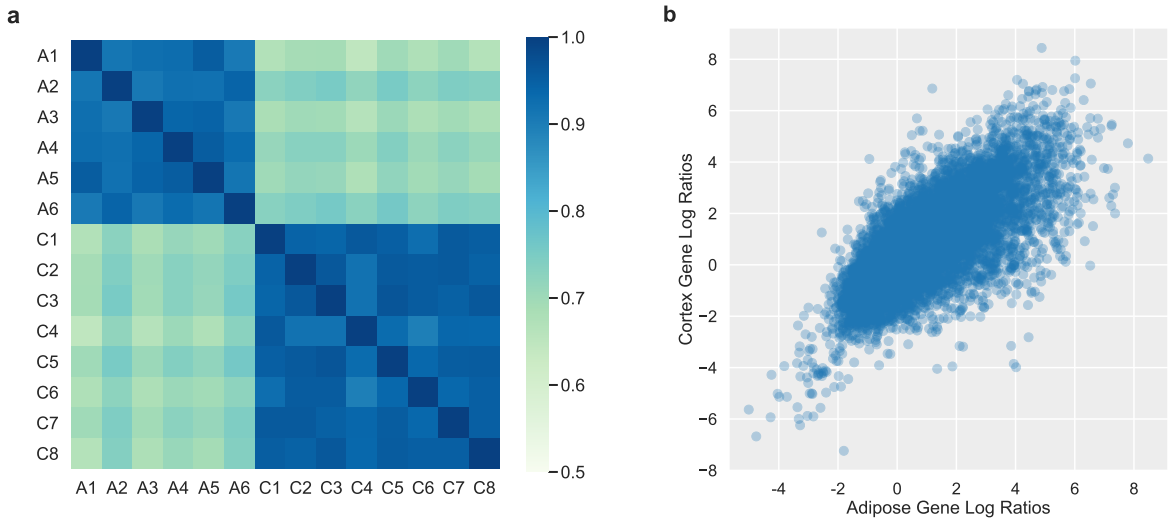

Figure 2.3: Consistency of snRNA-seq to bulk RNA-seq expression log-ratios across individuals, tissues, and experiments. (**a**) Heatmap depicting Pearson correlation between pairs of individual's log-ratios of snRNA-seq expression to bulk RNA-seq gene expression measured in counts per million (CPM). A sample prefix of 'A' indicates an individual from the adipose dataset and 'C' indicates an individual from the cortex dataset. Correlation is high between individuals within experiments as well as between experiments/tissues, indicating the same genes are over/under-expressed in snRNA-seq when compared to bulk RNA-seq. (**b**) Scatterplot of average snRNA-seq to bulk RNA-seq gene expression log-ratios across individuals in adipose dataset (x-axis) and cortex dataset (y-axis). Each point corresponds to a gene detected in both experiments, depicting the average ratio across all individuals for that tissue. The snRNA-seq to bulk RNA-seq ratios vary across genes and correlate (R=0.747) between these two experiments.

pled from half-normal distributions with increasing variance. In this model, a higher variance corresponds to a larger bias between sequencing experiments. While these transformations closely mirrored the Bisque decomposition model, they utilized the true snRNA-seq counts for each individual whereas Bisque learned these transformations using the reference profile generated from averaging these counts across all cells. Hence, this simulation framework introduced additional noise that Bisque does not entirely model. We evaluated decomposition performance by comparing proportion estimates to the proportions observed in the snRNA-seq data in terms of global Pearson correlation (R) and root mean squared deviation (RMSD). Due to the small number of samples, we applied leave-one-out cross-validation to predict the cell composition of each individual using the remaining snRNA-seq samples as training data for each method. In these simulations, Bisque remained robust ($R \approx 0.85$, $RMSD \approx 0.07$) at higher levels of simulated bias between the bulk and snRNA-seq-based reference (Figure 2.4b).

Next, we performed this cross-validation benchmark on the observed bulk RNA-seq data for these 6 individuals and found that Bisque ($R = 0.923$, $RMSD = 0.074$) provided significantly improved global accuracy in detecting each cell type over existing methods (Table 2.2). MuSiC ($R = -0.111$, $RMSD = 0.427$), BSEQ-sc ($R = -0.113$, $RMSD = 0.432$), and CIBERSORT ($R = -0.131$, $RMSD = 0.416$) severely underestimated the proportion of adipocytes (the most abundant population in adipose tissue) while overestimating the endothelial cell fraction. We also benchmarked CIBERSORTx [39], which employs a batch correction mode to account for biases in sequencing technologies. While CIBERSORTx ($R = 0.687$, $RMSD = 0.099$) outperformed existing methods, Bisque provided improved accuracy. It should be noted that cell-specific accuracy is more informative than global R and RMSD; however, these small sample sizes did not provide robust measures of within-cell-type performance in this cross-validation framework. We were able to slightly improve the number of detected cell populations by MuSiC, BSEQ-sc, and CIBERSORT when we considered only snRNA-seq reads aligning to exonic regions of the transcriptome, indicating that intronic reads introduced increasing discrepancy between snRNA-seq and bulk RNA-seq in the con-

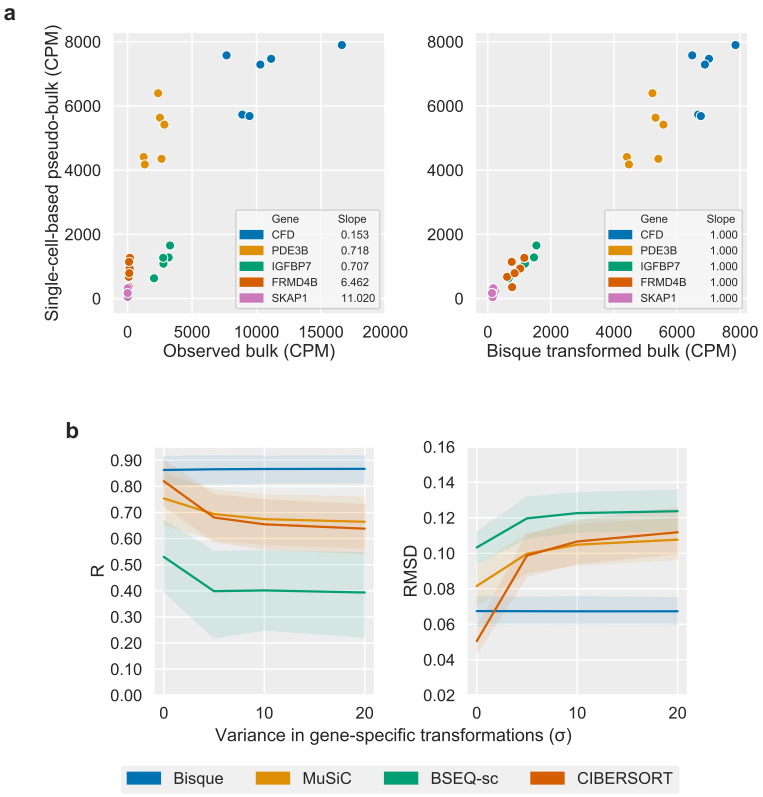

Figure 2.4: The effect of discrepancies between a single-cell based reference and bulk expression on decomposition. (**a**) Observed discrepancies in real data between single-nucleus and bulk expression for selected marker genes (left) for six individuals. Each color corresponds to a gene. On the left, observed bulk expression on the x-axis is plotted against the pseudo-bulk expression on the y-axis, where pseudo-bulk expression is calculated by summing the single-cell based reference with cell proportions as weights. On the right, the Bisque transformation of bulk expression is on the x-axis. Bisque recovers a one-to-one relationship by transforming the bulk expression for improved decomposition accuracy (right). (**b**) Simulation of bulk expression for six individuals based on true adipose snRNA-seq data with increasing gene-specific differences. These differences are modeled as a linear transformation of the summed snRNA-seq counts with coefficient and intercept sampled from Half-Normal distributions with parameter as indicated on the x-axis. At $\sigma = 0$, the simulated bulk is simply the sum of the observed single-cell read counts. Performance on y-axis measured in global Pearson correlation (R) (left) and root mean squared deviation (RMSD) (right). Shaded regions indicate 95% confidence intervals based on bootstrapping with central lines indicate the mean observed value. Bisque remains robust to increasing gene-specific variation between single-cell and bulk expression levels.

Method	R	RMSD
Bisque	**0.923 ± 0.064**	**0.074 ± 0.034**
CIBERSORTx	0.687 ± 0.450	0.099 ± 0.046
MuSiC	-0.111 ± 0.182	0.427 ± 0.058
BSEQ-sc	-0.113 ± 0.180	0.432 ± 0.058
CIBERSORT	-0.131 ± 0.176	0.416 ± 0.059

Table 2.2: Leave-one-out cross-validation in subcutaneous adipose using 6 samples with snRNA-seq and bulk RNA-seq data available. Proportions based on snRNA-seq were used as a proxy for the true proportions. Performance measured in Pearson correlation (R) and root-mean-square deviation (RMSD) across all 5 identified cell types in each sample. Reported values were averaged across the 6 samples with standard deviation indicated.

text of decomposition. However, given that a significant portion of the nuclear transcriptome consists of pre-mRNA, this filtering process removed over 40 percent of cells detected in the snRNA-seq data. Moreover, Bisque provided improved accuracy over existing methods using this exonic subset of the snRNA-seq data.

We then applied these decomposition methods to the remaining 100 bulk samples and found that the distribution of cell proportion estimates produced by Bisque were most concordant with the expected distribution inferred from the limited number of snRNA-seq samples and previously reported proportions [40, 41] (Figure 2.5a). While these benchmarks provided a measure of calibration (i.e. the ability to detect cell populations in expected ranges), they did not provide measurements of cell-specific proportion accuracy across individuals. In order to evaluate cell-specific accuracy, we replicated previously reported associations between cell proportions and measured phenotypes. Specifically, we compared cell proportion estimates from each method to body mass index (BMI) and Matsuda index, a measure of insulin resistance. We measured the significance of these association accounting for age, age-squared, sex, and relatedness.

Obesity is associated with adipocyte hypertrophy, the expansion of the volume of fat cells [42]; thus, we expected a negative association between adipocyte proportion and BMI. Bisque, MuSiC and CIBERSORTx produced adipocyte proportion estimates that replicate this behavior, while BSEQ-sc and CIBERSORT were unable to detect this cell population (Figure 2.5b). The adipocyte proportion estimates produced by Bisque ($p = 0.030$) and

Method	Spearman Correlation	Spearman p-value	Effect Estimate	Effect Standard Error	Effect t-value	Effect p-value
Bisque	-0.178	0.090	-0.282	0.126	-2.240	**0.030**
MuSiC	0.038	0.719	-0.081	0.108	-0.754	0.455
BSEQ-sc	-	-	-	-	-	-
CIBERSORT	-	-	-	-	-	-
CIBERSORTx	-0.300	**0.004**	-0.361	0.100	-3.624	**0.001**
BisqueMarker	-0.227	**0.030**	-0.304	0.096	-3.154	**0.003**

Table 2.3: Association of adipocyte proportion with BMI. A negative association was expected.

Method	Spearman Correlation	Spearman p-value	Effect Estimate	Effect Standard Error	Effect t-value	Effect p-value
Bisque	0.389	$<$ **0.001**	0.460	0.099	4.671	$<$ **0.001**
MuSiC	0.065	0.540	0.034	0.110	0.308	0.760
BSEQ-sc	0.238	**0.022**	0.278	0.092	3.013	**0.004**
CIBERSORT	0.239	**0.022**	0.162	0.102	1.597	0.118
CIBERSORTx	0.273	**0.009**	0.224	0.102	2.192	**0.034**
BisqueMarker	0.296	**0.004**	0.253	0.103	2.465	**0.018**

Table 2.4: Association of macrophage proportion with BMI. A positive association was expected.

CIBERSORTx ($p = 0.001$) had a significant negative association with BMI (Table 2.3). In addition, macrophage abundance has been shown to increase in adipose tissue with higher levels of obesity, concomitant with a state of low grade inflammation [43]. Each method detected macrophage populations that positively associated with BMI; however, only Bisque ($p < 0.001$), BSEQ-sc ($p = 0.004$) and CIBERSORTx ($p = 0.049$) reached significance (Table 2.4).

T cells were the least abundant cell type population identified from the snRNA-seq data, constituting around 4 percent of all sequenced nuclei. The abundance of T cells has been observed to positively correlate with insulin resistance [44]. Thus, we compared decomposition estimates for T cell proportions to Matsuda index. As a lower Matsuda index indicates higher insulin resistance, we expect a negative association between T cell proportion and Matsuda index. Proportion estimates produced by Bisque and CIBERSORTx followed this

Method	Spearman Correlation	Spearman p-value	Effect Estimate	Effect Standard Error	Effect t-value	Effect p-value
Bisque	-0.195	0.075	-0.387	0.116	-3.328	**0.002**
MuSiC	-	-	-	-	-	-
BSEQ-sc	-	-	-	-	-	-
CIBERSORT	-	-	-	-	-	-
CIBERSORTx	-0.317	**0.003**	-0.230	0.111	-2.068	**0.046**
BisqueMarker	-0.294	**0.007**	-0.188	0.100	-1.874	0.069

Table 2.5: Association of T cell proportion with Matusda index, a measure of insulin resistance. A negative association was expected. An additional covariate accounting for diabetes status was added to the LMM due to previously reported significant associations with Matsuda index.

trend while the remaining existing methods did not identify T cells in the bulk samples (Figure 2.5c). We found this association significant for Bisque ($p < 0.001$) and CIBERSORTx ($p = 0.047$) (Table 2.5) after correcting for diabetes status, since Matsuda index may not be informative in these individuals [45].

2.3.3 Evaluation of decomposition performance in cortex tissue

We also benchmarked these decomposition methods using expression data collected from the dorsolateral prefrontal cortex (DLPFC). This dataset was generated by the Rush Alzheimer's Disease (AD) Center [46] and includes 636 postmortem bulk RNA-seq samples. The Religious Orders Study and Rush Memory and Aging Project were approved by an IRB of Rush University Medical Center. Both bulk RNA-seq and snRNA-seq data were collected from 8 of the individuals (Table 2.1). Using the same pipeline we used to process the adipose dataset, we identified 11 clusters: 3 neuronal subtypes, 2 interneuronal subtypes, 2 astrocyte subtypes, oligodendrocytes, oligodendrocyte progenitor cells, and microglia (Figure 2.2b). We observed a higher overlap in marker genes for these clusters than in those identified in the adipose dataset (average of 10% of marker genes shared between clusters in DLPFC compared to 3% in adipose).

We again applied leave-one-out cross-validation on the 8 individuals with both RNA-

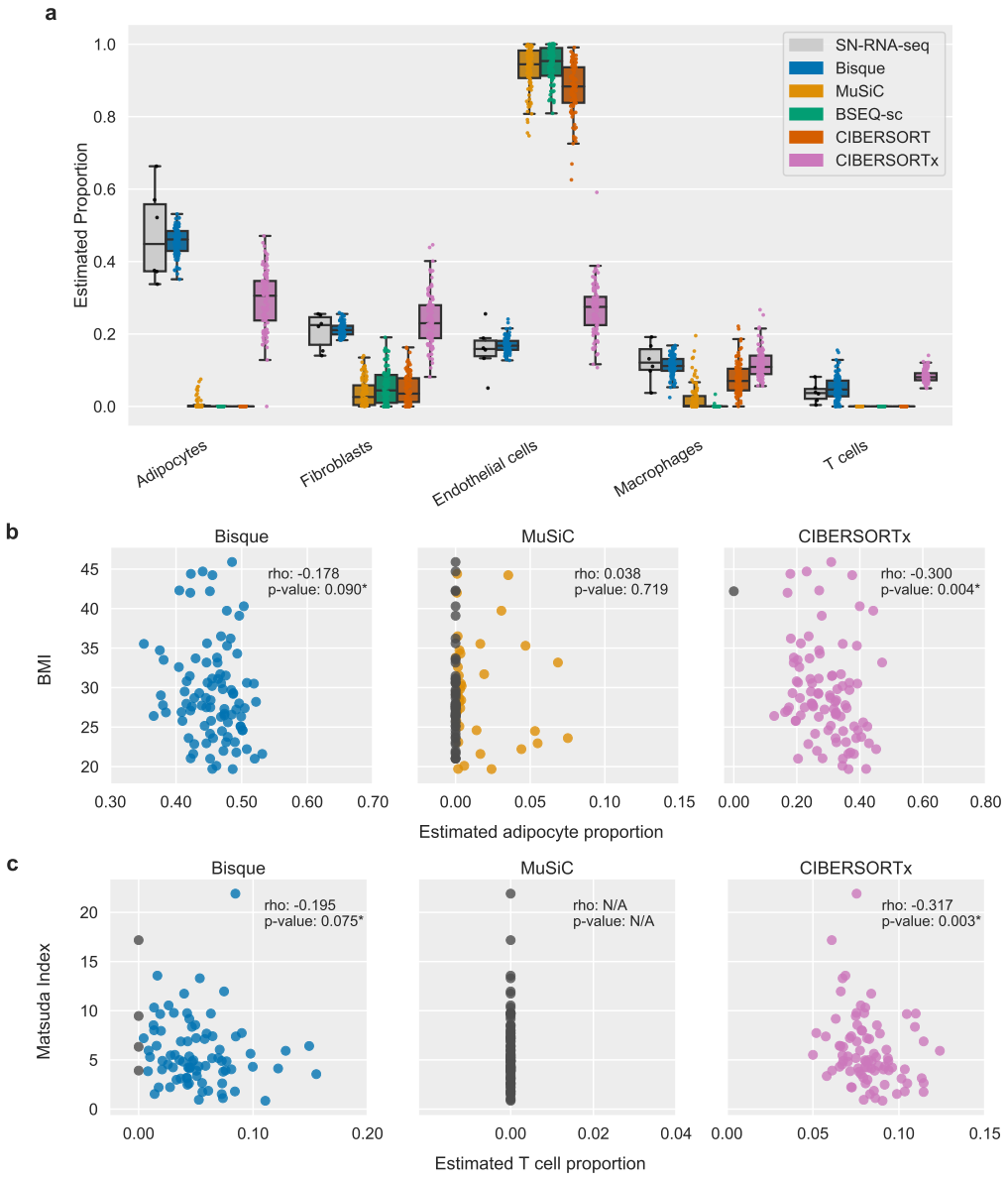

Figure 2.5: Decomposition benchmark in human subcutaneous adipose tissue. (**a**) Comparison of decomposition estimates from 100 individuals with estimates from 6 individuals with snRNA-seq data available. (**b-c**) Scatterplots comparing decomposition estimates with measured phenotypes in 100 individuals. Reported 'rho' corresponds to Spearman correlation and p-values indicate the significance of these correlations, with an asterisk denoting significance after correction for covariates (sex, age, age-squared, and relatedness). CIBERSORT and BSEQ-sc are not shown since they did not detect these cell populations. (**b**) Adipocyte proportion has been observed to negatively correlate with BMI so we expected a negative correlation. (**c**) T cell proportion has previously been reported to positively correlate with insulin resistance. Matsuda index decreases with higher insulin resistance so we expected a negative correlation.

Method	R	RMSD
Bisque	**0.924 ± 0.062**	**0.029 ± 0.010**
CIBERSORTx	0.671 ± 0.153	0.070 ± 0.019
MuSiC	-0.192 ± 0.107	0.173 ± 0.013
BSEQ-sc	0.098 ± 0.216	0.120 ± 0.023
CIBERSORT	-0.281 ± 0.049	0.197 ± 0.012

Table 2.6: Leave-one-out cross-validation in dorsolateral prefrontal cortex using 8 samples with snRNA-seq and bulk RNA-seq data available. Proportions based on snRNA-seq were used as a proxy for the true proportions. Performance measured in Pearson correlation (R) and root-mean-square deviation across all 11 identified cell types in each sample. Reported values were averaged across the 8 samples with standard deviation indicated.

seq and snRNA-seq data available. In this example, randomly sampled 25% of the nuclei in the snRNA-seq data to accommodate CIBERSORTx (which is currently web-based and restricts the size of files that can be processed). Bisque was able to detect each cell population identified from the snRNA-seq data with high global accuracy (R = 0.924, RMSD = 0.029) while MuSiC (R = -0.192, RMSD = 0.173), BSEQ-sc (R = 0.098, RMSD = 0.120), and CIBERSORT (R = -0.281, RMSD = 0.197) did not detect a number of cell populations (Table 2.6). Bisque also provided higher accuracy than CIBERSORTx (R = 0.671, RMSD = 0.070). However, we found that the performance of the existing methods improved when estimates with subtypes were summed together. While each method was able to quantify major cell populations after merging subtypes, Bisque was able to distinguish between these closely related cell populations. Interestingly, we found that in both adipose and DLPFC, endothelial cell proportions were overestimated by each of the existing methods.

We applied these decomposition methods to the remaining 628 individuals and compared the distribution of estimates to the proportions observed in the 8 snRNA-seq samples. We found that Bisque was able to detect each cell population and produced estimates that were closest in mean to the snRNA-seq observations (Figure 2.6a). The increased accuracy of Bisque over existing methods persisted when we merged closely related subtypes. Moreover, immunohistochemistry (IHC) analyses on 70 of these samples found similar proportions of major cell populations [47], confirming the relative accuracy of snRNA-seq based estimates of cell proportions.

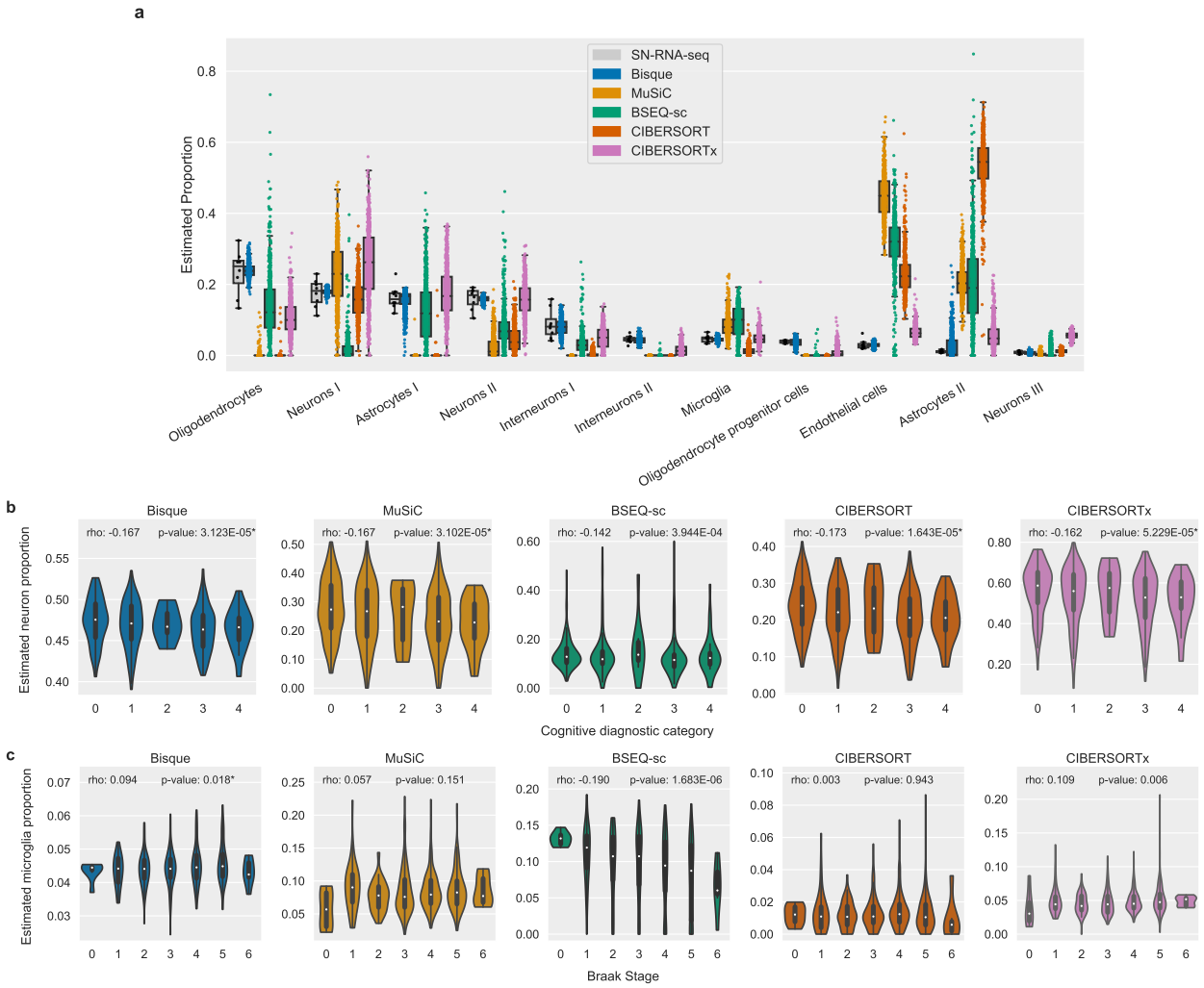

Figure 2.6: Decomposition benchmark in human dorsolateral prefrontal cortex tissue. (**a**) Comparison of decomposition estimates from 628 individuals with estimates from 8 individuals with snRNA-seq data available. (**b-c**) Violin plots depicting association of decomposition estimates aggregated into major cell types with measured phenotypes in 628 individuals. Reported 'rho' corresponds to Spearman correlation and p-values indicate the significance of these correlations, with an asterisk denoting both an expected effect direction and significance after correction for covariates. (**b**) Neuronal degeneration has been observed in patients diagnosed with Alzheimer's disease (AD). Cognitive diagnostic category measures a physician's diagnosis of cognitive impairment (CI), with 0 indicating no CI and 4 indicating a confident AD diagnosis. We expected a negative correlation between neuron proportion and cognitive diagnostic category. (**c**) Microglia proportion has been observed to positively correlate with increased severity of AD symptoms, such as neurofibrillary tangles. Braak stage provides a semiquantitative measure of tangle severity, so we expected an overall positive correlation between microglia proportion and Braak stage.

Method	Spearman Correlation	Spearman p-value	Effect Estimate	Effect Standard Error	Effect t-value	Effect p-value
Bisque	-0.167	< **0.001**	-0.145	0.039	-3.705	< **0.001**
MuSiC	-0.167	< **0.001**	-0.147	0.039	-3.742	< **0.001**
BSEQ-sc	-0.142	< **0.001**	-0.053	0.039	-1.341	0.180
CIBERSORT	-0.173	< **0.001**	-0.155	0.039	-3.971	< **0.001**
CIBERSORTx	-0.162	< **0.001**	-0.127	0.039	-3.237	**0.001**
BisqueMarker	-0.141	< **0.001**	-0.142	0.039	-3.645	< **0.001**

Table 2.7: Association of neuron proportion with cognitive diagnosis category. A negative association was expected.

Method	Spearman Correlation	Spearman p-value	Effect Estimate	Effect Standard Error	Effect t-value	Effect p-value
Bisque	0.094	**0.018**	0.118	0.037	3.220	**0.001**
MuSiC	0.057	0.151	0.019	0.037	0.509	0.611
BSEQ-sc	-0.190	< 0.001	-0.166	0.037	-4.525	< 0.001
CIBERSORT	0.003	0.943	-0.005	0.037	-0.137	0.891
CIBERSORTx	0.109	**0.006**	0.056	0.037	1.517	0.130
BisqueMarker	0.092	**0.021**	0.054	0.037	1.444	0.149

Table 2.8: Association of microglia proportion with Braak stage, a measure of neurofibrillary tangles. A positive association was expected.

Again, to determine cell-specific decomposition accuracy, we replicated known associations between cell type proportions and measured phenotypes in the 628 individuals. For these analyses, we compared cell proportion estimates to each individual's Braak stage and physician cognitive diagnostic category at time of death. Braak stage is a semiquantitative measure of neurofibrillary tangles, ranging in value from 0 to 5 with increasing severity. The cognitive diagnostic category provides a semiquantitative measure of dementia severity, where a code of 1 indicates no cognitive impairment and 5 indicates a confident diagnosis of AD by physicians. We determined the significance of these associations based on t-values estimated by a linear regression model that accounted for age, age-squared, and sex.

Neuronal death is a hallmark symptom of AD [48]. Therefore, we expected to find a negative association between cognitive diagnosis and neuron proportion. We found that each decomposition method provides estimates of total neuron proportion that tend to decrease

with cognitive diagnostic category (Figure 2.6b). Each method generates proportions with negative association with cognitive diagnosis. Each method, excluding BSEQ-sc, reached significance in this model ($p \leq 0.003$ for each method) (Table 2.7). As another example, we compared each individual's Braak stage to their estimated proportion of microglia, a relatively small cell population that constituted roughly 5 percent of the sequenced nuclei. Microglia activation has been observed to increase with AD severity [49]. We used Braak stage as a proxy for AD severity and expected a positive association between microglia proportion and Braak stage. Bisque and MuSiC provided estimates that follow this expected trend (Figure 2.6c). Only Bisque produced estimates with a significant positive association ($p = 0.001$) (Table 2.8). Interestingly, we observe a decrease in microglia proportions estimated by Bisque in Braak stage 6 individuals which has been previously observed in AD patients [50].

2.3.4 Runtime comparisons of reference-based decomposition methods

Given the large amounts of transcriptomic data that are becoming available, we also benchmarked these decomposition methods in terms of runtime. In the subcutaneous adipose dataset, which included 100 bulk RNA-seq samples and 6 snRNA-seq samples with about 1,800 nuclei sequenced per individual, Bisque was able to estimate cell proportions efficiently compared to existing methods. Bisque (1 second) and MuSiC (1 second) provided decomposition estimates faster than BSEQ-sc (26 seconds), CIBERSORT (27 seconds), and CIBERSORTx (389 seconds) (Figure 2.7a). Bisque also provided improved efficiency in processing the reduced DLPFC dataset, which included 628 bulk RNA-seq samples and 8 snRNA-seq samples with around 2,125 nuclei per individual. Bisque (4 seconds) and MuSiC (10 seconds) estimated cell proportions relatively quickly compared to BSEQ-sc (273 seconds), CIBERSORT (298 seconds), and CIBERSORTx (6,566 seconds) (Figure 2.7b).

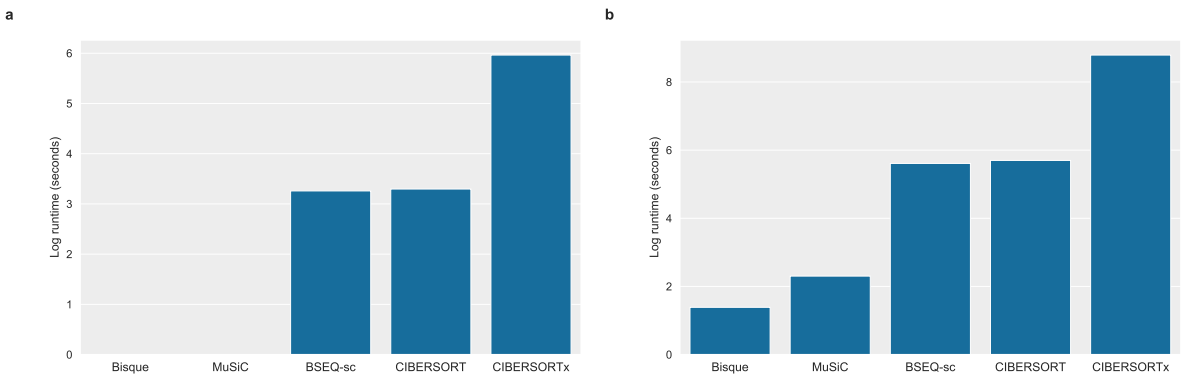

Figure 2.7: Runtime comparisons in log-transformed seconds for benchmarked reference-based decomposition methods. (**a**) Runtime for subcutaneous adipose dataset, which included 100 RNA-seq samples and 6 snRNA-seq samples with around 1,800 nuclei per individual. (**b**) Runtime for dorsolateral prefrontal cortex dataset, which included 628 RNA-seq samples and 8 snRNA-seq samples. We benchmarked each method using around 2,125 nuclei per snRNA-seq sample.

2.3.5 Robustness of the reference-based decomposition model

Our reference-based decomposition method is based on the assumption that cell populations are equally represented in single-cell and bulk RNA sequencing of the same tissue samples. Since this assumption may be violated [51], we explored the performance of our model as we relaxed this assumption in simulations. First, we simulated snRNA-seq data where cell proportions were increasingly biased. Using the DLPFC snRNA-seq data, we downsampled or upsampled the cells identified as microglia at varying levels and performed decomposition. Indeed, the absolute estimates produced by Bisque propagated these shifts in snRNA-seq proportions. However, we found that our estimates maintained their expected positive association with Braak stage, evidence for the correlation between these estimates and the true microglia proportions (Figure 2.8a). Given these results, we suggest that users take note of this behavior if both the mean abundances are important for downstream analysis and the single-cell reference data is known to be significantly biased against specific cell populations of interest.

Next, we simulated a situation where an unknown cell population contributes to bulk expression but is not represented in the snRNA-seq reference data. For situations where

this unknown contribution varies across the bulk dataset, we simulated bulk expression by mixing the observed bulk expression for the DLPFC dataset with increasing amounts of expression observed in the adipose dataset. To determine the effect of unknown cell populations on our model, we analyzed the distribution of residual norms produced by the method. These residual norms provide a measure of the difference between the vector of observed bulk and expression reference weighted by the estimated proportions across all genes for each individual. As we increased the contribution from unknown cell types, the residual norm values tend to increase (Figure 2.8b). In our simulation framework, this variability in unknown cell type contribution could be qualitatively identified by the presence of a multimodal residual norm distribution.

Given that single-cell datasets still remain relatively small compared to bulk datasets, we also explored the impact of sample size in the reference single-cell data on the performance of Bisque. In the DLPFC dataset, we saw a drop in performance when using less than four randomly selected snRNA-seq samples (Figure 2.8c). This threshold is likely to differ between experiments, though we recommend at least three single-cell samples to generate reference data.

Finally, since marker gene selection can vary between studies, we were interested in the performance of Bisque as we varied the number of marker genes. Again, we measured cell type proportion estimation performance for microglia in the DLPFC dataset by correlating the estimates with Braak stage, which is known to have a positive association. We recalculated this correlation as we removed marker genes for this cell type. We removed marker genes in order of both decreasing and increasing log-fold-change, which provides a measure of the importance of marker genes for identifying this cell type. In both procedures, we observe that as we remove an increasing percentage of the 102 identified marker genes, performance remains stable until a shared drop off point around 75% (Figure 2.8d). Since we observed this trend in both marker gene removal schemes, we assume that a relatively few number of marker genes, regardless of their log-fold-change magnitude, can be used to accurately estimate cell type proportions. These results suggest that as long as a core set of marker genes are present,

variations in less important marker genes will have little effect on downstream analyses.

2.3.6 Marker-based decomposition using known cell type marker genes

While a reference profile from snRNA-seq can help to decompose bulk level gene expression, it may not be available for the same data set. The majority of bulk RNA-seq data sets do not have corresponding snRNA-seq data in the same set of individuals. However, marker gene information from prior experiments can still be applied to distinct expression data sets of the same tissue. The basis of most decomposition methods relies on the logic that as the proportion of a cell type varies across individuals, the expression of its marker genes will tend to correlate in the same direction as its cell type proportion. This linear co-variation can be captured in a principal components analysis (PCA). Under the same argument, the more cell type-specific a marker gene is, the more its expression will reflect its cell type proportion. These observations form the basis for BisqueMarker, a weighted PCA-based (wPCA) decomposition approach. Genes that are more specifically expressed within a cell type will provide more information than genes with shared expression across cell types. To estimate cell type proportions without the use of cell type-specific gene expression information, we applied wPCA to bulk-level adipose tissue expression.

For each cell type, we extracted the first PC from a wPCA of the expression matrix of its markers. The expression matrix was corrected for the first global expression PC as a covariate so that wPCA estimates would not reflect technical variation. We first confirmed that these genes were distinct across cell types. If 2 cell types share a high proportion of marker genes, the wPCA estimates from bulk RNA-seq will correlate highly. We then investigated whether the second or third PC could have represented cell type proportions. The percent of variance explained by the first PC was typically 30-60% across adipose cell types, and additionally, over 90% of the markers correlated in the same direction as the first PC. In contrast, roughly 50-70% of markers correlated in the same direction as the second or third PC. As performed for reference-based decomposition, we correlated phenotypes with cell type proportions estimated by BisqueMarker. We identified the same associations

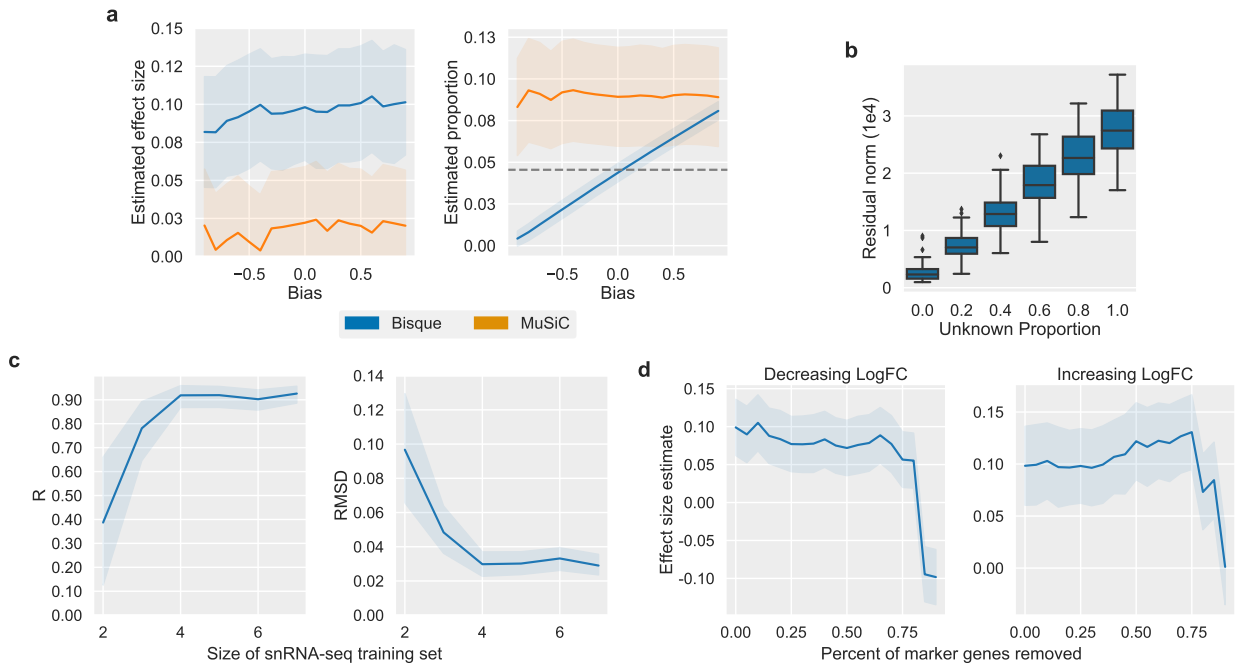

Figure 2.8: Robustness of the reference-based decomposition model. (**a**) Microglia cells in the DLPFC snRNA-seq data were upsampled or downsampled at various percentages, denoted as bias on the x-axis. Decomposition performance, measured as the estimated effect size of microglia proportion on Braak stage (which is expected to be positive) on the y-axis was consistent for each method as the bias in the snRNA-seq reference varied (left). The simulated bias propagates to the estimated proportions for Bisque(right). Shaded regions indicate standard error of estimates. (**b**) In order to model the severity of the sample discordance due to unknown cell fractions, we compared the amount of adipose contamination, denoted as unknown proportion on the x-axis, to the residuals from the Bisque model (y-axis). (**c**) Leave-one-out cross-validation performance in the DLPFC dataset after utilizing random subsamples of the snRNA-seq data as a reference. Performance measured in terms of Pearson correlation (left) and RMSD (right). Shaded regions indicate 95% confidence interval. (**d**) At each amount of marker genes removed (x-axis), performance was measured as the effect size of the estimated microglia proportion on Braak stage (y-axis). Genes were removed in order of decreasing (left) or increasing (right) log-fold-change. Shaded regions indicate standard error of estimates.

as with reference-based decomposition, demonstrating its validity when a reference is not available (Tables 2.3, 2.4, 2.5). Similarly, we observed the same trends between estimated cell type abundances and phenotypes as we did using our reference-based method in the DLPFC cohort (Tables 2.7, 2.8).

2.4 Discussion

Bisque effectively leverages single-cell information to decompose bulk expression samples, outperforming existing methods in datasets with snRNA-seq data available. In simulations, we demonstrated that the decomposition accuracy of Bisque is robust to increasing variation between the generation of the reference profile and bulk expression, which is a significant issue when comparing snRNA-seq and bulk RNA-seq data. In observed bulk expression, our reference-based method accurately estimates cell proportions that are consistent with previously reported distributions and reliably detects rare cell types. We found that these estimates consistently follow expected trends with measured phenotypes, suggesting that cell-specific estimates of proportion are sufficiently accurate to extract relevant biological signals. In addition, differences in tissue structure can lead to significant differences in the quality of single-cell expression data [52]. We demonstrated the improved performance of our method in adipose and DLPFC, two distinct tissues, suggesting that Bisque is robust across different tissue types.

The cell type proportion estimates determined by Bisque may be utilized to effectively identify cell-type-specific interactions, such as expression quantitative trait loci (eQTLs), and adjust for confounding effects from variability in cell populations. With this reference-based approach, single-cell sequencing of a subset of samples from large-scale bulk expression cohorts can provide high power to detect cell-specific associations in complex phenotypes and diseases.

However, we note that there are limitations to this reference-based method that users should consider. First, if the number of individuals with single-cell data available is small,

the reference profile and gene-specific transformations may become unreliable. In addition, a key assumption of our transformation framework is that single-cell based estimates of cell proportions accurately reflect the true proportions we wish to estimate. As a result of this assumption, Bisque provides estimates of cell proportions reported by the single-cell technology used to generate the reference data. Given that snRNA-seq can provide less bias in isolating specific cell types compared to scRNA-seq [53, 54], we expect these estimates to be useful for downstream analyses such as those previously discussed. Nevertheless, the accuracy of Bisque may decrease if the proportion of cell types captured by single-cell experiments differs significantly from the true physiological distributions. Therefore, we advise users to take caution if there is a known significant bias in the single-cell measurements of a tissue, such as severe underrepresentation of a cell type of interest, that can affect downstream analysis. Our results demonstrate that even with these limitations, Bisque can be used to provide cell-type specific biological insight in relevant datasets.

In cases where these described issues may be significant, BisqueMarker provides cell type abundance estimations using only known marker genes. While this reference-free method may be less accurate than reference-based methods, it does not depend on single-cell based estimates of cell proportions or expression profiles, but rather on the fact that the expression in certain genes differs across different cell types; moreover, this method also does not model explicitly the expression level, and it is thus robust to biases in the single cell sequencing protocol. We found that BisqueMarker estimates followed expected trends with measured phenotypes; however, it should be noted that this method estimates relative differences in abundances that cannot be compared across cell types. Also, given the semi-supervised nature of this method, these cell type abundance estimates may include signals from technical or other biological variation in the data. Therefore, we highly suggest applying this method to data that is properly normalized with sources of undesired variation removed. Bisque is available as an R package on CRAN (BisqueRNA) and at `https://github.com/cozygene/bisque`.

CHAPTER 3

An efficient linear mixed model framework for meta-analytic association studies across multiple contexts

3.1 Background

Over the last decade, the scale of genomic datasets has steadily increased. These datasets have grown to the size of hundreds of thousands of individuals [1] with millions soon to come [55]. Similarly, datasets for transcriptomics and epigenomics are growing to thousands of samples [56, 2, 57]. These studies provide valuable insight into the relationship between our genome and complex phenotypes [4].

Identifying these associations requires statistical models that can account for biases in study design that can negatively influence results through false positives or decreased power. Linear mixed models (LMMs) have been a popular choice for controlling these biases in genomic studies, utilizing variance components to account for issues such as population stratification [58]. These models can also be used to analyze studies with repeated measurements from individuals, such as replicates or measurements across different contexts. Meta-Tissue [59] is a method that applies this model in the context of identifying expression quantitative trait loci (eQTLs) across multiple tissues. In this framework, gene expression is measured in several tissues from the same individuals and the LMM is utilized to test the association between these values and genotypes. A meta-analytic approach is used to combined effects across multiple tissues to increase the power of detecting eQTLs. This approach has also been applied to increase power in genome-wide association studies (GWAS) by testing the

association between genotypes and multiple related phenotypes [60].

However, these approaches are computationally intensive. Existing approaches for fitting these models are cubic in time complexity with respect to the number of samples across all contexts [58, 61]. Here, we present an ultra-fast LMM framework specifically for multiple-context studies. Our method, mcLMM, is linear in complexity with respect to the number of individuals and allows for statistical tests in a manner of hours rather than days or years with existing approaches. To illustrate the computational efficiency of mcLMM, we compare the runtime and memory usage of our method with EMMA and GEMMA [58, 61], two popular approaches for fitting these models. We further apply mcLMM to identify a large number of eQTLs in the Genotype-Tissue Expression (GTEx) dataset [2] and compare our results from METASOFT [62], which performs the meta-analysis of the mcLMM output, to a recent meta-analytic approach known as `mash` [63]. Finally, to demonstrate the practicality of mcLMM on modern datasets, we perform a multiple-phenotype GWAS combining over a million observations sampled from hundreds of thousands of individuals in the UK Biobank [1] within hours.

3.2 Methods

3.2.1 Linear Mixed Model

For multi-context experiments with n individuals, t contexts, and c covariates, we fit the following linear mixed model

$$\mathbf{y} = X\beta + \mathbf{u} + \mathbf{e} \tag{3.1}$$

where $\mathbf{u} \sim N(0, \sigma_g^2 K)$, $\mathbf{e} \sim N(0, \sigma_e^2 I)$, $\mathbf{y} \in R^{nt}$ is a vectorized representation of the responses, $X \in R^{nt \times tc}$ is the matrix of covariates, $\beta \in R^{tc}$ is the vector of estimated coefficients, $K \in R^{nt \times nt}$ is a binary matrix where $K_{i,j} = 1$ indicates that sample i and sample j in Y come from the same individual, and $I \in R^{nt \times nt}$ is an identity matrix. X is structured

such that both an intercept and the covariate effects are fit within each context. For sake of simplicity, dimensions of nt assume that there is no missing data; however, this is not a requirement for the model. We note that this definition of K models within-individual variability as a random-effect, while within-context or across-individual variability is not included.

The full and restricted log-likelihood functions for this model are

$$l_F(\mathbf{y};\beta,\sigma_g,\delta) = \frac{1}{2}\left[-N\log\left(2\pi\sigma_g^2\right) - \log(|H|) - \frac{1}{\sigma_g^2}(\mathbf{y}-X\beta)^T H^{-1}(\mathbf{y}-X\beta)\right] \tag{3.2}$$

$$l_R(\mathbf{y};\beta,\sigma_g,\delta) = l_F(\mathbf{y};\beta,\sigma_g,\sigma_e) + \frac{1}{2}\left[tc\log(2\pi\sigma_g^2) + \log\left(|X^TX|\right) - \log\left(|X^TH^{-1}X|\right)\right] \tag{3.3}$$

where N is the total number of measurements made across the individuals and contexts, $\delta = \frac{\sigma_e^2}{\sigma_g^2}$, and $H = K + \delta I$ [64]. These likelihood functions are maximized with the generalized least squares estimator $\hat{\beta} = (X^TH^{-1}X)^{-1}X^TH^{-1}\mathbf{y}$ and $\hat{\sigma_g^2} = \frac{R}{N}$ in the full log-likelihood and $\hat{\sigma_g^2} = \frac{R}{N-tc}$ in the restricted log-likelihood, where $R = (\mathbf{y}-X\hat{\beta})^TH^{-1}(\mathbf{y}-X\hat{\beta})$. Our goal is to maximize these likelihood functions to estimate the optimal $\hat{\delta}$.

3.2.2 Likelihood refactoring in the general case

The EMMA algorithm optimizes these likelihoods for δ by refactoring them in terms of constants calculated from eigendecompositions of H and SHS, where $S = I - X(X^TX)^{-1}X^T$, that allow linear complexity optimization iterations with respect to the number of individuals [58]. The GEMMA algorithm further increases efficiency by replacing the SHS eigendecomposition with a matrix-vector multiplication [61]. Both approaches require the eigendecomposition of at least one N by N matrix which is typically cubic in complexity. Here, we show that our specific definition of K as a binary indicator matrix allows us to refactor these likelihood functions without any eigendecomposition steps. It should be noted that EMMA and GEMMA can fit this model for any positive semidefinite K, while mcLMM is restricted to the definition described above.

We note that previous work has shown similar speedups when the matrix K is low rank and has a block structure as described here [65]. This work, FaST-LMM, shows that the likelihood functions can be computed in linear time with respect to the number of individuals after singular value decomposition of a matrix with complexity that is also linear with respect to the number of individuals. We improve upon these methods by recognizing that the eigenvalues of the K matrix are known beforehand, which allows for further efficiency in fitting this model. Furthermore, the FaST-LMM model assumes that all individuals within each context share additional covariance while mcLMM assumes that all contexts observed within an individual share additional covariance.

First, note that $H = K + \delta I$ is a block diagonal matrix. Specifically, each block corresponds to an individual i with t_i contexts measured, where t_i is less than or equal to t depending on the number of contexts observed for individual i. Each block is equal to $[\mathbf{1}_{t_i} + \delta I_{t_i}] \in R^{t_i \times t_i}$, where $\mathbf{1}_{t_i}$ is a t_i by t_i matrix composed entirely of 1. These properties of H make its eigendecomposition and inverse directly known.

The eigenvalues of a block diagonal matrix are equal to the union of the eigenvalues of each block. Moreover, the eigenvalues of $[\mathbf{1}_{t_i} + \delta I_{t_i}]$ are $t_i + \delta$ with multiplicity 1 and δ with multiplicity $t_i - 1$. Therefore, H has eigenvalues δ with multiplicity $N - n$ and $t_i + \delta$ for each t_i. This provides our first refactoring

$$\log(|H|) = (N - n)\log(\delta) + \sum_{i=1}^{n} \log(t_i + \delta) \tag{3.4}$$

The inverse of a block diagonal matrix can also be computed by inverting each block individually. Moreover, using the Sherman-Morrison formula [66], the inverse of $[\mathbf{1}_{t_i} + \delta I_{t_i}]$ is known

$$(\mathbf{1}_{t_i} + \delta I_{t_i})^{-1} = -\frac{1}{t + \delta}\mathbf{1}_{t_i} + \frac{1}{\delta} I_{t_i} \tag{3.5}$$

Given each entry of H^{-1}, we can show algebraically that

$$X^T H^{-1} X = \frac{1}{\delta}(E - D) \tag{3.6}$$

$$E_{i,j} = \begin{cases} \sum_{\text{ind} \in f(i)} x_{\text{ind},g(i)} x_{\text{ind},g(j)} & \text{if } f(i) = f(j) \\ 0 & \text{if } f(i) \neq f(j) \end{cases} \tag{3.7}$$

$$D_{i,j} = \sum_{g \in \text{groups}} \frac{1}{t_g + \delta} \sum_{\text{ind} \in f(i), f(j), g} x_{\text{ind},g(i)} x_{\text{ind},g(j)} \tag{3.8}$$

where $f(i) = i\%t$ (modulo operator) provides the context of a given 0-indexed column of X, $g(i) = i//t$ (integer division) provides the covariate of a given index. A group g defines the set of individuals that share the same number of measured contexts t_g. The expression "ind $\in f(i), f(j), g$" indicates the set of all individuals that have t_g measured contexts that include context i and j.

Note that with all values independent of δ pre-computed, specifically the sum of covariate interactions within the sets of individuals indicated above, E is constant with respect to δ and each entry of the symmetric matrix D can be calculated in linear time with respect to the number of groups, which is less than or equal to the number of contexts t. For a given δ, we can compute $X^T H^{-1} X$ in $O(t(tc)^2)$ time complexity. Both the restricted and full log-likelihoods require the calculation of $(X^T H^{-1} X)^{-1}$. The restricted log-likelihood requires the additional calculation of $\log(|X^T H^{-1} X|)$. To calculate both of these terms, we perform a Cholesky decomposition of $X^T H^{-1} X = LL^*$, where * indicates the conjugate transpose. Given this decomposition, we can compute

$$\log(|X^T H^{-1} X|) = \sum_{i=1}^{tc} 2\log(L_{i,i}) \tag{3.9}$$

$$(X^T H^{-1} X)^{-1} = (L^*)^{-1} L^{-1} \tag{3.10}$$

These operations can be done in $O((tc)^3)$ time complexity.

Let $P(X)$ denote a projection matrix and $M(X) = (I - P(X))$. Note that both $P(X)$

and $M(X)$ are idempotent. The term remaining term in the likelihood functions, R, can be reformulated as follows

$$\begin{aligned} \mathbf{y} - X\hat{\beta} &= \mathbf{y} - X(X^T H^{-1} X)^{-1} X^T H^{-1} \mathbf{y} \\ &= (I - X(X^T H^{-1} X)^{-1} X^T H^{-1})\mathbf{y} \\ &= (I - P(X))\mathbf{y} \\ &= M(X)\mathbf{y} \end{aligned} \tag{3.11}$$

$$\begin{aligned} M(X)^T H^{-1} &= (I - X(X^T H^{-1} X)^{-1} X^T H^{-1})^T H^{-1} \\ &= (I - H^{-1} X(X^T H^{-1} X)^{-1} X^T) H^{-1} \\ &= H^{-1} - H^{-1} X(X^T H^{-1} X)^{-1} X^T H^{-1} \\ &= H^{-1}(I - X(X^T H^{-1} X)^{-1} X^T H^{-1}) \\ &= H^{-1} M(X) \end{aligned} \tag{3.12}$$

$$\begin{aligned} R &= (\mathbf{y} - X\hat{\beta})^T H^{-1} (\mathbf{y} - X\hat{\beta}) \\ &= \mathbf{y}^T M(X)^T H^{-1} M(X)\mathbf{y} \\ &= \mathbf{y}^T H^{-1} M(X) M(X)\mathbf{y} \\ &= \mathbf{y}^T H^{-1} M(X)\mathbf{y} \\ &= (\mathbf{y}^T H^{-1}\mathbf{y}) - (\mathbf{y}^T H^{-1} X(X^T H^{-1} X)^{-1} X^T H^{-1}\mathbf{y}) \\ &= a - \mathbf{b}^T (X^T H^{-1} X)^{-1}\mathbf{b} \\ &= a - \mathbf{b}^T (L^*)^{-1} L^{-1}\mathbf{b} \end{aligned} \tag{3.13}$$

The scalar a and vector $\mathbf{b}$ are a function of δ and can be algebraically formulated as

$$a = \frac{1}{\delta}\left(\left(\sum_{i=1}^{N} \mathbf{y}_i^2\right) - \left(\sum_{g \in \text{groups}} \frac{1}{t_g + \delta} \sum_{\text{ind} \in g} (\sum \mathbf{y}_{\text{ind}})^2\right)\right) \tag{3.14}$$

$$\mathbf{b}_i = \frac{1}{\delta}\left(\left(\sum_{\text{ind}\in\text{context(i)}} x_{\text{ind},g(i)}\mathbf{y}_{\text{ind},f(i)}\right) - \left(\sum_{g\in\text{groups}} \frac{1}{t_g+\delta} \sum_{\text{ind}\in f(i),g} x_{\text{ind},g(i)}(\sum \mathbf{y}_{\text{ind}})\right)\right) \tag{3.15}$$

where $\sum \mathbf{y}_{\text{ind}}$ indicates the sum of responses across all contexts for an individual. With values independent of δ pre-calculated, a and $\mathbf{b}$ can be calculated in linear time with respect to the number of groups.

Note that Equations 3.16 and 3.17 remove terms that are independent of δ since they are not required for finding its optimal value, indicated by the $\approx$ symbol. We can reformulate the entire likelihood functions as follows

$$\begin{aligned} l_F(\mathbf{y};\beta,\sigma_g,\delta) &= \frac{1}{2}\left[-N\log\left(2\pi\sigma_g^2\right) - \log(|H|) - \frac{1}{\sigma_g^2}(\mathbf{y}-X\beta)^T H^{-1}(\mathbf{y}-X\beta)\right] \\ &= \frac{1}{2}\left[-N\log\left(2\pi\frac{R}{N}\right) - \log(|H|) - N\right] \\ &= \frac{1}{2}\left[-N\log\left(2\pi\frac{R}{N}\right) - \left((N-n)\log\left(\delta\right) + \sum_{i=1}^{n}\log\left(t_i+\delta\right)\right) - N\right] \\ &\approx -N\log\left(a - \mathbf{b}^T(L^*)^{-1}L^{-1}\mathbf{b}\right) - \left((N-n)\log\left(\delta\right) + \sum_{i=1}^{n}\log\left(t_i+\delta\right)\right) \end{aligned} \tag{3.16}$$

$$\begin{aligned} l_R(\mathbf{y};\beta,\sigma_g,\delta) &= l_F(\mathbf{y};\beta,\sigma_g,\sigma_e) + \frac{1}{2}\left[tc\log(2\pi\sigma_g^2) + \log\left(|X^TX|\right) - \log\left(|X^TH^{-1}X|\right)\right] \\ &\approx (tc-N)\log\left(a - \mathbf{b}^T(L^*)^{-1}L^{-1}\mathbf{b}\right) \\ &- \left((N-n)\log\left(\delta\right) + \sum_{i=1}^{n}\log\left(t_i+\delta\right)\right) - \sum_{i=1}^{tc} 2\log(L_{i,i}) \end{aligned} \tag{3.17}$$

These likelihoods are maximized using the optimize function in R. Each likelihood evaluation has a time complexity of $O((tc)^3+n)$.

3.2.3 Likelihood refactoring with no missing data

When there is no missing data, the likelihood functions can be further simplified. Note that in this case, $N = nt$ and all $t_i = t$. Hence,

$$\begin{aligned} \log(|H|) &= (N-n)\log(\delta) + \sum_{i=1}^{n} \log(t_i + \delta) \\ &= (nt-n)log(\delta) + n\log(t+\delta) \end{aligned} \tag{3.18}$$

If the input terms $\mathbf{y}$, X, and K are permuted resulting in samples being sorted in order of context, and the covariates in X are sorted in order of context, we can decompose H and X into

$$H = (\mathbf{1}_t + \delta I_t) \otimes I_n \tag{3.19}$$

$$X = I_t \otimes X_{\text{dense}} \tag{3.20}$$

where $\otimes$ indicates the Kronecker product and $X_{\text{dense}} \in R^{n \times c}$ is a typical representation of the covariates for each individual without multiple contexts (i.e. samples as rows and covariates as columns). Utilizing the properties of Kronecker products, we can perform the following reformulation

$$\begin{aligned} (X^T H^{-1} X)^{-1} &= ((I_t \otimes X_{\text{dense}}^T)((\mathbf{1}_t + \delta I_t) \otimes I_n)^{-1}(I_t \otimes X_{\text{dense}}))^{-1} \\ &= ((\mathbf{1}_t + \delta I_t)^{-1} \otimes X_{\text{dense}}^T X_{\text{dense}})^{-1} \\ &= (\mathbf{1}_t + \delta I_t) \otimes (X_{\text{dense}}^T X_{\text{dense}})^{-1} \end{aligned} \tag{3.21}$$

$$
\begin{aligned}
\log\left(|(X^T H^{-1} X)^{-1}|\right) &= \log\left(|(\mathbf{1}_t + \delta I_t) \otimes (X_{\text{dense}}^T X_{\text{dense}})^{-1}|\right) \\
&= \log\left(|(\mathbf{1}_t + \delta I_t)|^c |(X_{\text{dense}}^T X_{\text{dense}})^{-1}|^t\right) \\
&= c \log\left(|(\mathbf{1}_t + \delta I_t)|\right) + t \log\left(|(X_{\text{dense}}^T X_{\text{dense}})^{-1}|\right) \\
&= c \log\left(\frac{1}{(t+\delta)\delta^{t-1}}\right) + t \log\left(|(X_{\text{dense}}^T X_{\text{dense}})^{-1}|\right) \\
&= c\left(-\log(t+\delta) - (t-1)\log(\delta)\right) + t \log\left(|(X_{\text{dense}}^T X_{\text{dense}})^{-1}|\right)
\end{aligned}
\tag{3.22}
$$

Note that the remaining determinant in Equation 3.22 will not need to be calculated since it is independent of δ. Next, we show that $\hat{\beta}$ is independent of δ.

$$
\begin{aligned}
\hat{\beta} &= (X^T H^{-1} X)^{-1} X^T H^{-1} \mathbf{y} \\
&= \left((\mathbf{1}_t + \delta I_t) \otimes (X_{\text{dense}}^T X_{\text{dense}})^{-1}\right) X^T H^{-1} \mathbf{y} \\
&= \left((\mathbf{1}_t + \delta I_t) \otimes (X_{\text{dense}}^T X_{\text{dense}})^{-1}\right) (I_t \otimes X_{\text{dense}}^T)((\mathbf{1}_t + \delta I_t)^{-1} \otimes I_n)\mathbf{y} \\
&= \left((\mathbf{1}_t + \delta I_t) \otimes (X_{\text{dense}}^T X_{\text{dense}})^{-1} X_{\text{dense}}^T\right) ((\mathbf{1}_t + \delta I_t)^{-1} \otimes I_n)\mathbf{y} \\
&= \left((\mathbf{1}_t + \delta I_t)(\mathbf{1}_t + \delta I_t)^{-1} \otimes (X_{\text{dense}}^T X_{\text{dense}})^{-1} X_{\text{dense}}^T\right) \mathbf{y} \\
&= \left(I_t \otimes (X_{\text{dense}}^T X_{\text{dense}})^{-1} X_{\text{dense}}^T\right) \mathbf{y}
\end{aligned}
\tag{3.23}
$$

This form of $\hat{\beta}$ shows that the optimal coefficients are equivalent to fitting separate ordinary least squares (OLS) models for each context. We compute $\hat{\beta}$ by concatenating these OLS estimates. Given this term, we can also compute the residuals of this model $\mathbf{s} = (\mathbf{y} - X\hat{\beta})$ and reformulate R as follows.

$$
\begin{aligned}
R &= (\mathbf{y} - X\hat{\beta})^T H^{-1} (\mathbf{y} - X\hat{\beta}) \\
&= \mathbf{s}^T H^{-1} \mathbf{s} \\
&= \sum_{i=1}^{nt} \mathbf{s}_i \sum_{j=1}^{nt} \mathbf{s}_j H^{-1}_{j,i} \\
&= \frac{1}{\delta} \left(\sum_{i=1}^{nt} \mathbf{s}_i^2 \right) + \frac{1}{\delta(t+\delta)} \left(- \sum_{i=1}^{n} \left(\sum \mathbf{s}_{\text{ind}(i)} \right)^2 \right)
\end{aligned}
\tag{3.24}
$$

The term $\sum \mathbf{s}_{\text{ind}(i)}$ denotes the sum of residuals for an individual across all contexts. Let $u = \sum_{i=1}^{nt} \mathbf{s}_i^2$ and $v = -\sum_{i=1}^{n} \left(\sum \mathbf{s}_{\text{ind}(i)} \right)^2$.

$$
R = \frac{1}{\delta} u + \frac{1}{\delta(t+\delta)} v \tag{3.25}
$$

Now we can reformulate the log-likelihoods, omitting terms that do not depend on δ.

$$
\begin{aligned}
l_F(\delta) &= -nt \log(R) - \log(|H|) \\
&= -nt \log \left(\frac{1}{\delta} u + \frac{1}{\delta(t+\delta)} v \right) - (nt - n) \log(\delta) - n \log(t+\delta) \\
&= -nt \log \left(u + \frac{1}{t+\delta} v \right) + n \log \left(\frac{\delta}{t+\delta} \right)
\end{aligned}
\tag{3.26}
$$

$$
\begin{aligned}
l_R(\delta) &= (tc - nt) \log(R) - \log(|H|) - \log(|(X^T H^{-1} X)^{-1}|) \\
&= (tc - nt) \log \left(u + \frac{1}{t+\delta} v \right) + (c - n) \log \left(\frac{t+\delta}{\delta} \right)
\end{aligned}
\tag{3.27}
$$

Both functions are differentiable with respect to δ. Moreover, both derivatives have the same root

$$\hat{\delta} = \frac{-tu - v}{u + v} \tag{3.28}$$

The scalar values u and v can be calculated by performing a separate OLS regression for each context, which can be completed in $O(t(nc^2+c^3))$ time for a naive OLS implementation. Compared to the methods described above, this approach requires no iterative optimization and the estimate is optimal. Our implementation has a time complexity of $O(c^3+nc^2+tcn)$.

3.2.4 Resource requirement simulation comparison

We installed EMMA v1.1.2 and manually built GEMMA from its GitHub source (genetics-statistics/GEMMA.git, commit 9c5dfbc). We edited the source code of GEMMA to prevent the automatic addition of intercept term in the design matrix (commented out lines 1946 to 1954 of src/param.cpp).

Data were simulated using the mcLMM package. Sample sizes of 100, 200, 300, 400, and 500 were simulated with 50 contexts. Context sizes of 4, 8, 16, 32, and 64 were simulated with 500 samples. Data were simulated with $\sigma_e^2 = 0.2$ and $\sigma_g^2 = 0.4$ and a sampling rate of 0.5. Memory usage of each method was measured using the peakRAM R package (v 1.0.2).

3.2.5 False positive rate simulation

We simulated gene expression levels in multiple tissues for individuals where there were no eQTLs. In other words, gene expression levels were not affected by any SNPs. We considered 10,000 genes and 100 SNPs resulting in one million gene-SNP pairs. We simulated 1,000 individuals. We also examined false positive rates with 500 and 800 individuals. We generated 49 such datasets where the number of tissues varied from 2 to 50. To simulate the genotypes for each subject, we randomly generated two haplotypes (vectors consisting of 0 and 1) assuming a minor allele frequency (MAF) of 30%. To simulate gene expression levels from multiple tissues among the same individuals, we sampled gene expression from the following multivariate normal distribution:

$$\mathbf{y} \sim N(0, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}) \tag{3.29}$$

where $\mathbf{y}$ is an $N \times T$ vector representing the gene expression levels of N individuals in T tissues and $\mathbf{K}$ is an $NT \times NT$ matrix corresponding the correlation between the subjects across the tissues. $K_{i,j} = 1$ when i and j are from two tissues of the same individuals, $K_{i,j} = 0$ otherwise. Here, we let $\sigma_g = \sigma_e = 0.5$. We used a custom R function (included with the mcLMM package) to simulate data from this distribution, which is based on sampling with a smaller covariance matrix for each block of measurements from an individual.

After generating the simulation datasets, we first ran mcLMM to obtain the estimated effect sizes and their standard errors, as well as the correlation matrices. The results from mcLMM were used as the input of METASOFT for meta-analysis to evaluate the significance. False positive rate was calculated as the proportion of gene-SNP pairs with p-values smaller than the significance level ($\alpha = 0.05$).

3.2.6 True positive simulations

We developed the true positive simulation framework based on a previous study describing `mash` [63]. We simulated effects for 20,000 gene-SNP pairs in 44 tissues, 400 of which have non-null effects (true positives) and 19,600 of which have null effects. Let β_{jr} denote the effects of the gene-SNP pair j in context/tissue r and β_j is a vector of effects across various tissues, including null effects and non-null effects. We simulated the gene expression levels for 1,000 individuals as:

$$\mathbf{y} = \beta_{jr}^T X + \mathbf{e} \tag{3.30}$$

where X denotes the genotypes of the individuals that were simulated as described in the false positive rate simulation. $\mathbf{e} \sim N(0, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I})$, which is similar to the simulation in the false positive rate simulation. For β_j, we defined two types of non-null effects and simulated them in different ways:

- Shared, structured effects: non-null effects are shared in all tissues and the sharing is

structured. The non-null effects are similar in effect sizes and directions (up-regulation or down-regulation) across all tissues, and this similarity would be stronger among some subsets of tissues. For 19,600 null effects, we set $\beta_j = 0$. For 400 non-null effects, we assumed that each β_j independently followed a multivariate normal distribution with mean 0 and variance ωU_k, where k is an index number randomly sample from $1, \ldots, 8$. $\omega = |\omega'|, \omega' \sim N(0, 1)$ represents a scaling factor to help capture the full range of effects. U_k are 44×44 data-driven covariance matrices learned from the GTEx dataset, which are provided in [63].

- Shared, unstructured effects: non-null effects are shared in all tissues but the sharing is unstructured or independent across different tissues. For 19,600 null effects, we set $\beta_j = 0$. For 400 non-null effects, we sampled β_j from a multivariate normal distribution with mean of 0 and variance of $0.01I$, where I is a 44×44 identity matrix.

After simulating the gene expression levels $\mathbf{y}$, we first ran mcLMM on the simulated datasets to acquire the estimated effect sizes and their standard errors, as well as the correlation matrices. We then applied METASOFT for meta-analysis with mcLMM outputs to evaluate the significance. For `mash`, we first performed simple linear regression to get the estimates of the effects and their standard errors in each tissue separately. These estimates and standard errors were used as the inputs for `mash`, which returned the measure of significance for each effect, the local false sign rate (lfsr). Finally, we employed the "pROC" R package [67] to calculate the receiver operating characteristic (ROC) curve and area under the ROC curve with the significance measures (p-values for mcLMM and METASOFT, lfsr for `mash`) and the correct labels of null effects and non-null effects.

3.2.7 Analysis of the GTEx dataset

The Genotype-Tissue Expression (GTEx) v8 dataset [2] was used in this study. We downloaded the gene expression data, the summary statistics of single-tissue cis-eQTL data using a 1 MB window around each gene, and the covariates in the eQTL analysis from GTEx

portal (https://gtexportal.org/home/datasets). The subject-level genotypes were acquired from dbGaP accession number phs000424.v8.p2. The GTEx v8 dataset includes 49 tissues from 838 donors. We selected 15,627 genes that were expressed in all 49 tissues. We only included SNPs with minor allele frequency (MAF) greater than 1% and missing rate lower than 5%. We applied `mash` and mcLMM plus METASOFT to the GTEx v8 dataset in our analysis.

Since `mash` requires observation of the correlation structure among non-significant tests and data-driven covariance matrices before fitting its model, we prepared its input by selecting the top SNP with the smallest p-value and 49 random SNPs (or all other SNPs if there were fewer than 49 SNPs left in a gene) in every gene from the eQTL analysis in the GTEx v8 dataset. There were 560,475 gene-SNP pairs in total. `mash` uses the estimated effect sizes and standard errors of these gene-SNP pairs to learn the correlation structure of different conditions/tissues. We used the top significant SNPs to set up the data-driven covariances. We then fit `mash` to the random set of gene-SNP pairs with the canonical and data-driven covariances. With the fitted `mash` model, we computed the posterior summaries including local false sign rate (lfsr) [68] for the selected gene-SNP pairs to estimate the significance. We defined significant gene-SNP pairs as those with lfsr < 0.05 in any tissues.

We applied mcLMM to the same set of gene-SNP pairs. We regressed out unwanted confounding factors in gene expression levels for each tissue with a linear model using covariates provided by GTEx. Covariates of each sample included top 5 genotyping principal components, PEER factors [69] (15 factors for tissues with fewer than 150 samples, 30 factors for those with 150-250 samples, 45 factors for those with 250-350 samples, and 60 factors for those with more than 350 samples), sequencing platform, and sex. We ran mcLMM with the genotypes and processed gene expression levels of all 838 individuals across 49 GTEx tissues for each gene-SNP pair. Missing values in gene expression were included in the mcLMM input. The effect sizes, standard errors, and correlation matrices estimated by mcLMM were meta-analyzed with METASOFT to evaluate the significance under both the fixed effects (FE) and random effects (RE2) models. The resulting p-values were converted to q-values

[70] to control false discovery rates. A gene-SNP pair was considered significant if its false discovery rate (FDR) was smaller than 5%.

3.2.8 Analysis of the UK Biobank dataset

This work was conducted using the UK Biobank Resource under application 33127. Samples were filtered for Caucasian individuals (Data-Field 22006)). Hard imputed genotype data from the UK Biobank were LD pruned using a window size of 50, step size of 1, and correlation threshold of 0.2. SNPs were further filtered for minor allele frequency of at least 0.01 and a Hardy-Weinberg equilibrium p-value greater than 1e-7 using Plink 2 [71]. Samples were filtered for unrelated individuals with KING using a cutoff value of 0.125 [72]. Genotype data were split by chromosome and converted to bigsnpr format (v 1.4.4) for memory efficiency [73].

The following data fields were retrieved: age at recruitment (Data-Field 31), sex (Data-Field 21022), BMI (Data-Field 23104), body fat percentage (Data-Field 23099), 10 genetic principal components (Data-Field 22009), HDL Cholesterol (Data-Field 30760), LDL Direct (Data-Field 30780), Apolipoprotein A (Data-Field 30630), Apolipoprotein B (Data-Field 30640), and Triglycerides (Data-Field 30870). Continuous phenotypes were visually inspected and triglycerides were log-transformed due to skewness. Data were filtered for complete observations. All fields were scaled to unit variance and centered at 0.

HDL cholesterol, LDL cholesterol, Apolipoprotein A, Apolipoprotein B, and triglycerides were combined as response variables in the LMM and age, sex, BMI, body fat percentage, and the top 10 genetic principal components were used as additional covariates in the model. Each SNP was marginally fit with mcLMM. The coefficients output by this model for each phenotype were meta-analyzed to calculate FE p-values using METASOFT as packaged with Meta Tissue v 0.5. The top GWAS hits for five different chromosomes (one per chromosome) were validated using the NHGRI-EBI GWAS catalog [74] and compared to studies for LDL and HDL cholesterol (GCST008035 and GCST008037).

3.3 Results

3.3.1 Multi-context linear mixed models

We implement the statistical model described in Meta-Tissue [59], where we model the multi-context data as follows:

$$\mathbf{y} = X\beta + \mathbf{u} + \mathbf{e} \tag{3.31}$$

where $\mathbf{u} \sim N(0, \sigma_g^2 K)$ and $\mathbf{e} \sim N(0, \sigma_e^2 I)$. For n individuals and t contexts, $\mathbf{y}$ is a vector of nt responses, K is an nt by nt binary matrix where a value of 1 indicates that the observations were sampled from the same individual. Compared to a standard regression model, the variance component $\mathbf{u}$ accounts for within-individual variation that may occur with repeated sampling. The design matrix X fits coefficients β for each feature within each context independently. These coefficients, which describe the effect of the feature on the response within each context, can be used in a meta-analytic framework to combine the results. In our pipeline, we utilize the random effects model (RE2) from METASOFT, which assumes that effect sizes may be different across contexts and was shown to outperform existing meta-analysis methods [62].

Fitting this LMM requires estimation of the parameters σ_g^2 and σ_e^2, which can be estimated with traditional likelihood or restricted-likelihood approaches or through various optimized methods that have been developed, such as EMMA and GEMMA [58, 61]. These approaches require an eigendeomposition of the matrix K with is traditionally considered to be an $O((nt)^3)$ operation. mcLMM utilizes the block structure of the matrices in this model to perform matrix operations within contexts and avoids any eigendecomposition operations. This approach provides massive speedups with runtime complexities that are linear with respect to sample size n rather than cubic. As a note, mcLMM is not an approximation and fits identical models to these existing approaches.

3.3.2 mcLMM is computationally efficient

To demonstrate the efficiency of mcLMM compared to existing approaches, we applied our method to simulated data of varying sample sizes and number of contexts. For these simulations, we simulated a sampling rate of 0.5, which indicates that only half of all possible individual-context pairs of observations are expected to be sampled.

We first applied our method to simulations with a fixed number of 50 contexts and varied the sample size from 100 to 500. From these experiments, we observed that mcLMM requires computational time orders of magnitude less than EMMA and GEMMA. Similarly, when we fixed the number of samples at 500 and varied the context sizes from 4 to 64, we observed dramatically reduced runtimes for mcLMM.

In these experiments, mcLMM also significantly reduces the memory footprint compared to EMMA and GEMMA, since we avoid creating any nt by nt matrices. In these simulations, existing approaches quickly grow memory requirements, with usages that grow to dozens of gigabytes for modestly sized datasets in the thousands of samples. mcLMM allows large-scale studies to be performed on relatively little computational resources (Figure 3.1).

In cases where there is no missing data, mcLMM allows for further speedups. We ran similar simulations to compare mcLMM with no missing data (optimal model) and mcLMM with missing data (iterative model). We observed a dramatic speedup, with sample sizes of 500,000 individuals across 10 contexts completed in under 10 seconds for the optimal model compared to around 15 minutes for the iterative model (Figure 3.2).

3.3.3 mcLMM enables powerful meta analyses to detect eQTLs

We utilized mcLMM to reduce the computational resource requirements of the Meta-Tissue pipeline, which fits a multiple-context LMM and combines the resulting effect sizes using METASOFT [59]. While powerful, the existing approach utilizes EMMA to fit the LMM. For a recent release from the GTEx consortium [2], each pair of genes and single nucleotide polymorphisms (SNPs) required over two hours to run. Across hundreds of thousands of

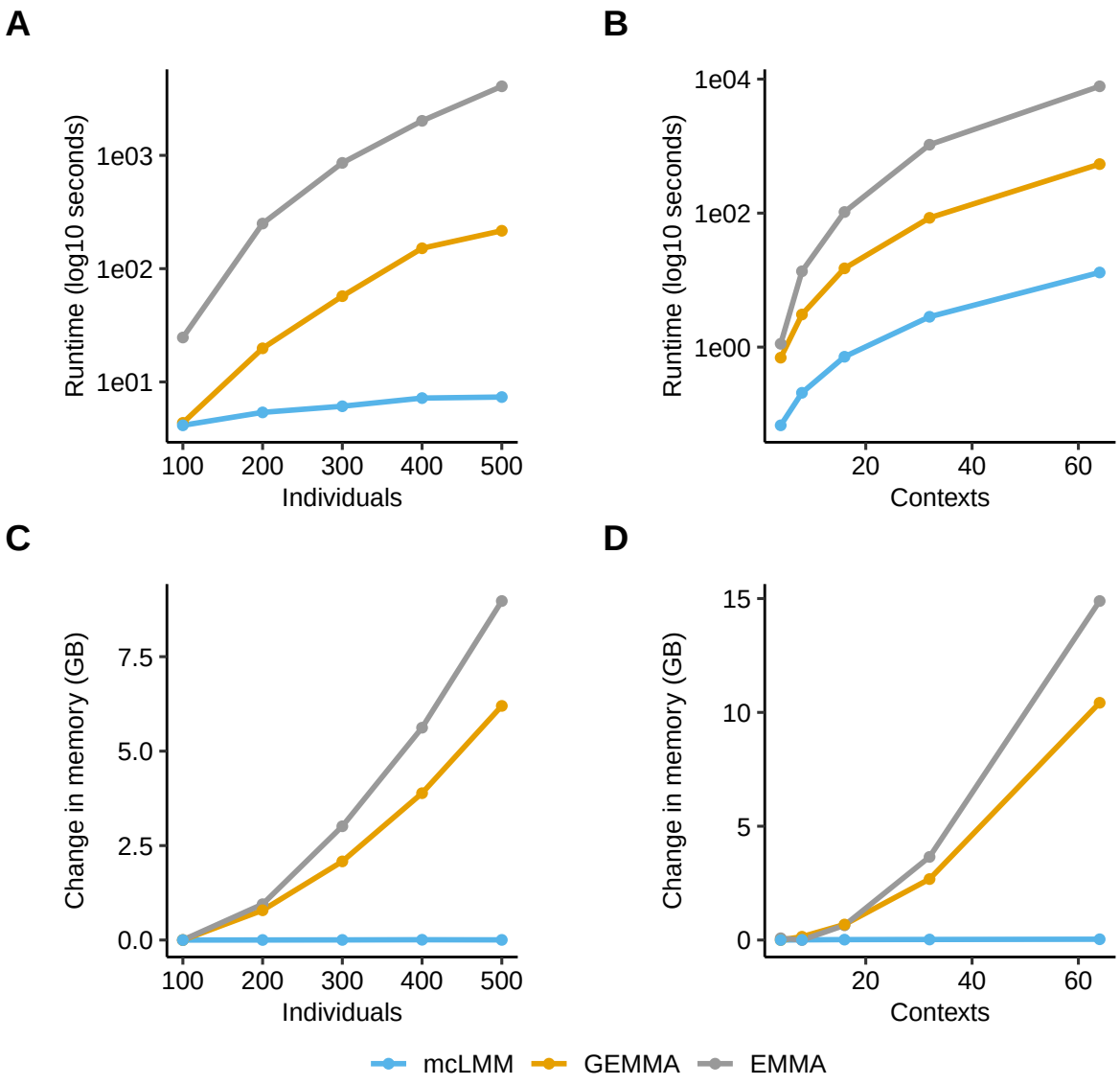

Figure 3.1: Resource requirements of mcLMM, GEMMA, and EMMA across various simulated individual and context sizes with missing values (sampling rate of 0.5). For varying individuals, contexts were fixed at 50. For varying contexts, individuals were fixed at 500. (A-B) Runtime with log10(seconds) on the y-axis and number of individuals or contexts simulated on the x-axis. (C-D) Memory usage (GB) on the y-axis and number of individuals or contexts simulated on the x-axis.

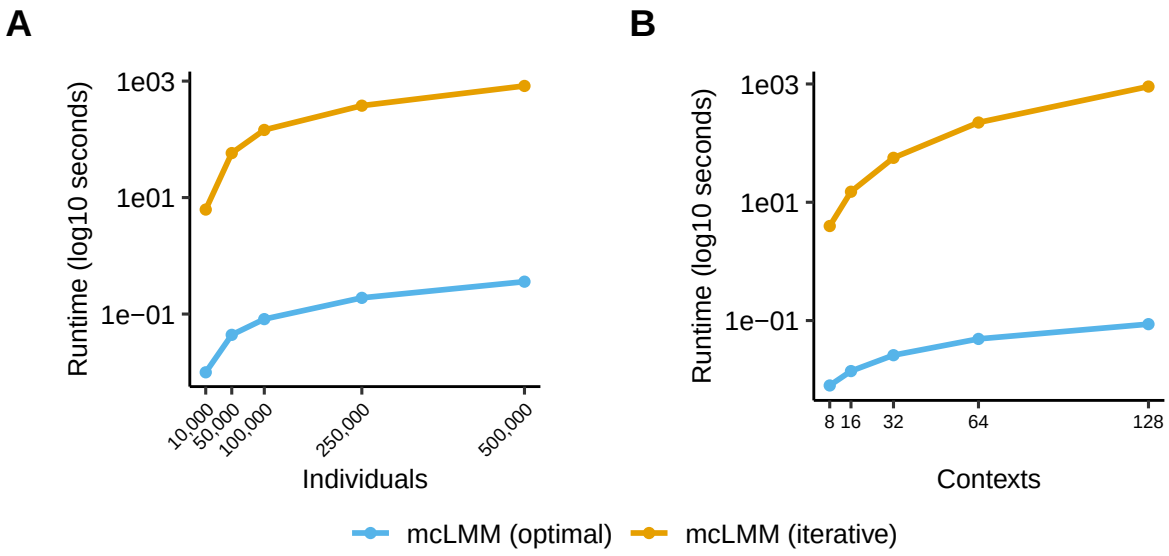

Figure 3.2: Runtime comparison of iterative and optimal mcLMM algorithms for data with no missing values. For varying individuals, contexts were fixed at 10. For varying contexts, individuals were fixed at 10,000. (A) Runtime across varying individuals. (B) Runtime across varying contexts.

gene-SNP pairs, this method would require years of computational runtime to complete. Utilizing mcLMM, we were able to complete this analysis in 3 days parallelized over each chromosome.

We compared our approach to a method known as `mash` [63]. This approach utilizes effect sizes estimated within each context independently and employs a Bayesian approach to combine their results for meta-analysis. In order to estimate the power of these methods, we performed simulations as described in the methods. In null simulations, we observed well-controlled false positive rates at $\alpha = 0.05$ for mcLMM coupled with METASOFT (Figure 3.3). In our simulation with true positives, we observed an increased area under the receiver operating characteristic (AUROC) for mcLMM coupled with the random effects (RE2) METASOFT model compared to `mash` (Figure 3.4).

Next, we compared the number of significant associations identified in the GTEx dataset. The `mash` approach utilized gene-SNP effect sizes estimated by the GTEx consortium within each tissue independently. Concordant with our simulations, we observed that the Meta-Tissue approach, utilizing mcLMM for vast speedup, identified more significant eQTLs than

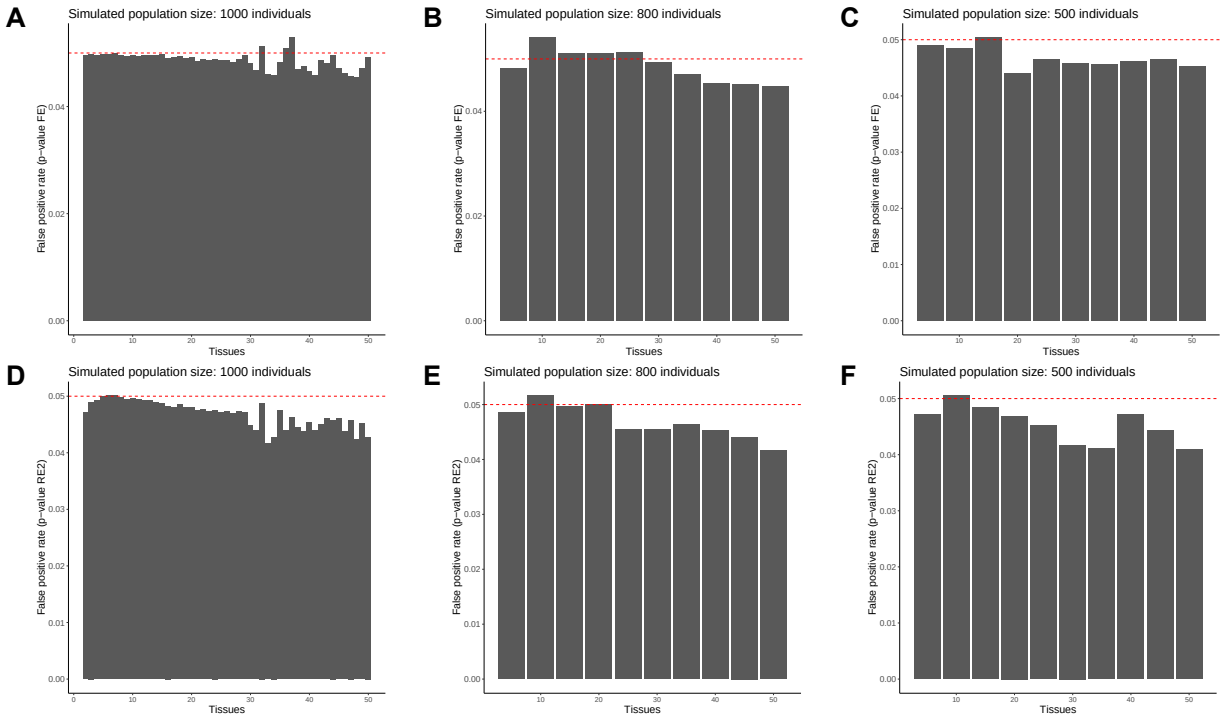

Figure 3.3: False positive rates of mcLMM + METASOFT in simulated data with 2-50 tissues. We estimated false positive rates with the p-values from METASOFT fixed effects (FE) model on the simulated data with (A) 1000 individuals, (B) 800 individuals, and (C) 500 individuals. Also, we estimated false positive rates with the p-values from METASOFT random effects (RE2) model on the simulated data with (D) 1000 individuals, (E) 800 individuals, and (F) 500 individuals.

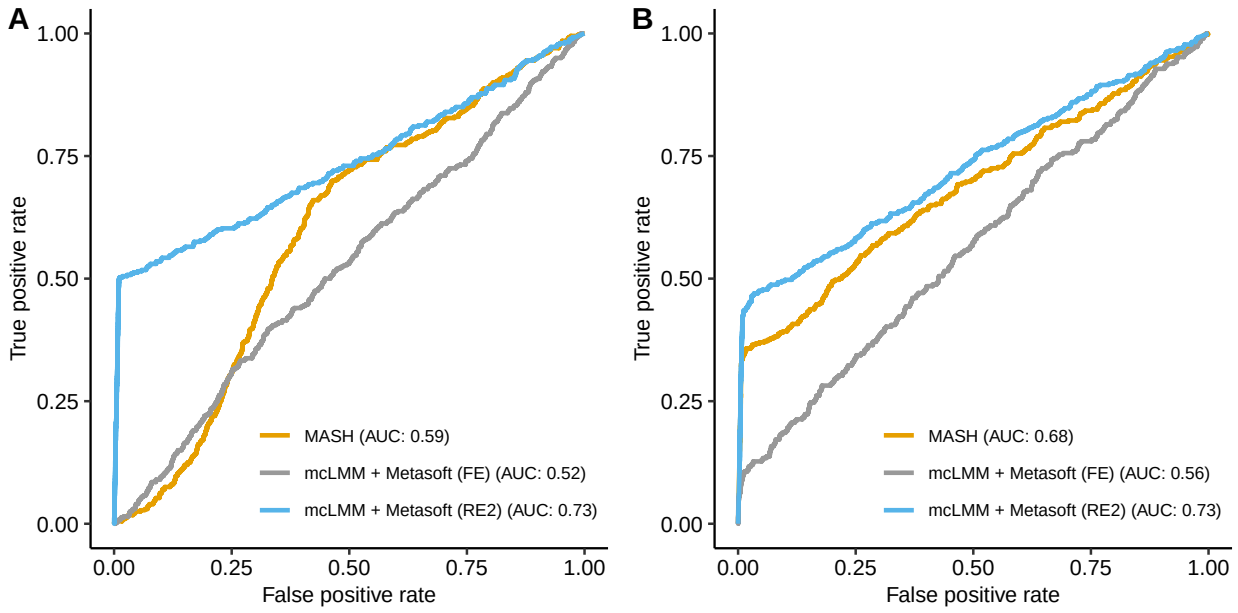

Figure 3.4: AUROC curves of mcLMM+METASOFT and `mash` in simulated data, assuming the effects of gene-SNP pairs are (A) shared and unstructured, and (B) shared and structured.

`mash` (Figure 3.5). These associations allow researchers to better understand the link between genetic variation and complex phenotypes through possible mediation of gene expression.

3.3.4 mcLMM scales to millions of samples across related phenotypes

As a practical application of the efficiency of mcLMM, we performed a multiple phenotype GWAS in the UK Biobank. A multiple phenotype GWAS associates SNPs with several related phenotypes in order to increase the effective sample size for greater power, under the assumption that the phenotypes are significantly correlated. For our analysis, we combined HDL and LDL cholesterol, Apolipoprotein A and B, and triglyceride levels across 323,266 unrelated caucasian individuals in the UK Biobank. In total, 1,616,330 observations of these related phenotypes were fit as responses in the LMM.

The mcLMM approach completed this analysis over 211,642 SNPs with an additional 14 covariates, parallelized over each chromosome, within a day. Each chromosome was analyzed on a single core machine with 32 GB of memory, with each test taking around 2 seconds to complete. We identified several significant loci, a subset of which replicate previous

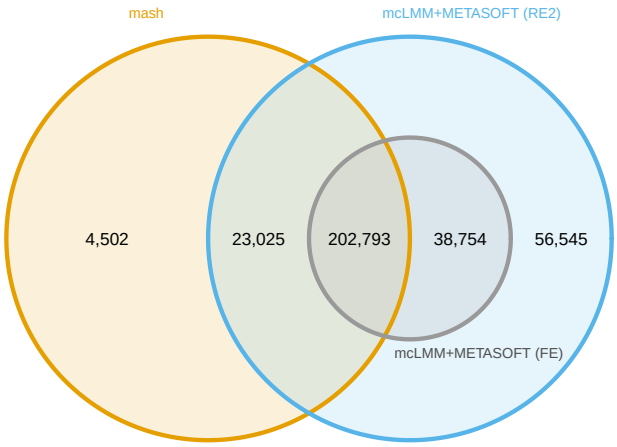

Figure 3.5: Venn diagram of significant eQTLs identified by meta-analysis methods in the GTEx dataset. We compared mcLMM using the random effects and fixed effects models in METASOFT (RE2 and FE, respectively) to `mash`. Note that areas are not proportional to the number of eQTLs in each region. mcLMM+METASOFT (RE2) identified a total of 321,117 significant associations that contained 225,818 eQTLs identified by `mash`.

findings for specific phenotypes included in the model, such as HDL cholesterol [75] (Figure 3.6). Existing approaches, namely EMMA and GEMMA, require orders of magnitude more memory to begin this analyses and could not be run on the available computational resources.

3.4 Discussion

We presented mcLMM, an efficient method for fitting LMMs used for multiple-context association studies. Our method provides exact results and scales linearly in time and memory with respect to sample size, while existing methods are cubic. This efficiency allows mcLMM to process hundreds of thousands of samples over several contexts within a day on minimal computational resources, as we showed in simulation and in the UK Biobank. The association parameters learned by mcLMM can further be utilized with the METASOFT framework to provide powerful meta-analysis of the associations, as we showed in the GTEx dataset.

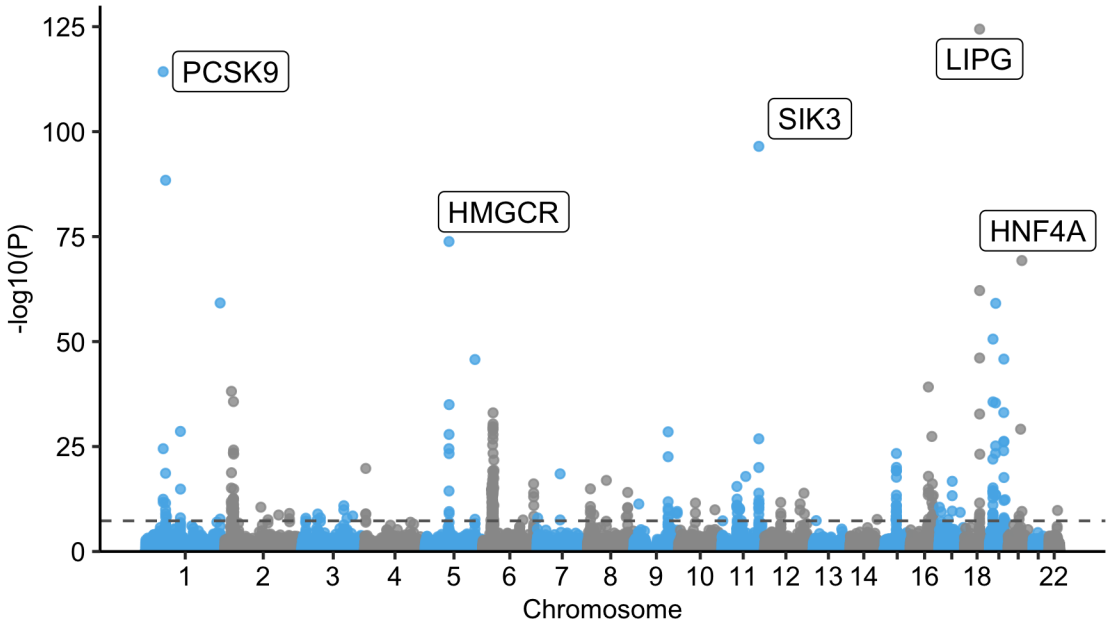

Figure 3.6: Multiple phenotype GWAS results from UK Biobank. Five phenotypes (LDL cholesterol, HDL cholesterol, Apolipoprotein A, Apolipoprotein B, and triglyceride levels) were used as responses in the mcLMM framework. The model was fit with 1,616,330 observations from 323,266 unrelated Caucasian individuals. In total, 211,642 SNPs were tested with an additional 14 covariates. Each test required around 2 seconds to run on a 32GB machine and was parallelized over each chromosome. The -log10 of the p-values are plot on the y-axis and genomic positions on the x-axis. The horizontal dashed line indicates the genome wide significance level at $p = 0.05/1e6$. The top hit for 5 different chromosomes is annotated with the gene containing the SNP. These genes have been previously identified as associated with a subset of these phenotypes.

Previous approaches have derived related speedups for LMMs when the matrix K is low rank, such as in the case when multiple samples are genetically identical or clustered in genome wide association studies as described in FaST-LMM [65]. In this approach, the authors show that the likelihood function can be evaluated in linear time with respect to the number of individuals after singular value decomposition of a matrix that is also linear with respect to the number of individuals. Other work has similarly used block structures and Kronecker refactorizations in studies with structured designs, such as multi-trait GWAS, to significantly speed up these approaches as well [76, 77].

Our approach builds upon these findings and we optimize the method specifically for the low rank matrix with known eigenvalues described in the model, thus avoiding any spectral or singular value decompositions. Furthermore, when there is no missing data, our method can compute the optimal model parameters with a closed form solution requiring no iterative optimization of likelihood functions. We also note that mcLMM models covariance across contexts within an individual while the FaST-LMM approach, described above, models covariance across individuals within each context. This specific model fit by mcLMM arises in multiple-context association studies, such as the approach employed by Meta Tissue [59] for identifying eQTLs across tissues utilizing the cubic EMMA algorithm. Applied within this framework for eQTL and multi-trait genome wide association studies, our method provides exact results and scales to hundreds of thousands of samples with minimal computational resources. mcLMM is available as an R package at `https://github.com/brandonjew/mcLMM`.

CHAPTER 4

Selection contributes to skewed X chromosome inactivation across human tissues

4.1 Background

X chromosome inactivation (XCI) occurs in blastocysts during embryonic development. Under no external pressures, the process is assumed to be random and is established in about a dozen cells and all daughter cells will inherit the XCI status of these cells [78]. Skew in XCI has been widely observed in female mammals and is accepted to be common across individuals [79]. Selective pressures have been implicated in contributing to XCI skew, especially in the context of diseases such as cancer [80, 17, 81]. In addition, XCI skew was found to be heritable and correlated with age based on a twin study [82]. Recent work argued that observed XCI skew in the general female population is the result of expected randomness in early development [79].

Here, we show that estimated XCI skew in a general female population in the GTEx dataset [2] is associated with genetic scores related to deleteriousness and proliferative potential. We hypothesize that differences in the fitness of maternal and paternal X chromosomes may contribute to skewed XCI across tissues into adulthood (Figure 4.1a). Furthermore, we perform a scan of common variants across the X chromosome to identify specific loci associated with overall XCI skew in these individuals.

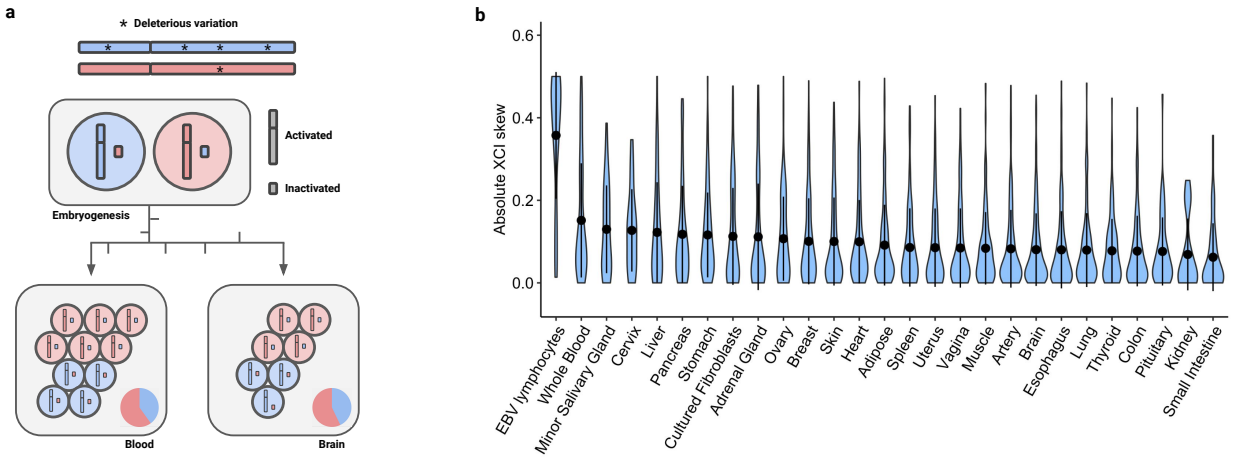

Figure 4.1: Measuring XCI skew across tissues **a**, Schematic overview of selection hypothesis. An individual inherits one haplotype (red) that is more fit than the other (blue) due to genetic variation. Given equal population sizes in embryonic development, we hypothesize that fitness differences will produce skewed populations in fully developed tissues. **b**, Violin plot of absolute XCI skew on the y-axis. Tissues are sorted by mean on the x-axis. Dots indicate mean of the absolute skew with vertical lines indicating one standard deviation.

4.2 Methods

4.2.1 Transcriptomic and genetic data

The following data were processed and made available by the GTEx consortium [2]. GENCODE v26 and the GRCh38 human reference genome was used to process both WGS and RNA-seq data. RNA-seq data was aligned using STAR with WASP filtering. Allele-specific read counts at heterozygous sites were quantified with GATK ASEReadCounter. Genetic variants were called from whole genome sequencing data and phased with read-aware SHAPEIT2. All analyses were restricted to caucasian samples.

4.2.2 Quantifying XCI skew

Allele-specific read counts were matched to haplotypes from the phased WGS data. Only heterozygous sites within fully inactivated genes reported by Carrel and Willard [83] and Cotton et al [84] were considered. We further selected genes where no females were observed to escape inactivation in Cotton et al. This filtering yielded heterozygous sites within 117 genes con-

sidered to be fully inactivated. LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) was used to convert these reported gene windows from GRCh37 to GRCh38 positions. Moreover, we only considered variants that had at least 10 overlapping reads. XCI skew was calculated at each heterozygous site in the gene windows as the number of reads coming from one haplotype divided by the total number of reads at the site. The final XCI skew measurement for each sample was the median of these per-site observations as done in previous work [79]. The haplotypes were arbitrarily labeled as H1 and H2, and skew was calculated as the number of reads mapping to H1 divided by the total reads at each site. Since haplotypes are not readily distinguishable as maternal or paternal, we also calculated the absolute XCI skew as the absolute deviation of this value from 0.5. Correlation between skewing in tissues was calculated as the Pearson correlation of pairwise-complete observations between tissue pairs where 25 or more individuals had both tissues sampled.

We also calculated skew in the same manner in autosomes. Specifically, if m heterozygous sites were used to calculate XCI in a sample, we randomly selected m heterozygous sites with at least 10 overlapping RNA-seq reads on the q-arm of each non-acrocentric autosome. We used these sites to calculate skew for each autosome as the median value. Since extreme XCI skewing was observed in EBV-transformed lymphocytes, we removed these samples from all downstream analyses for both X and autosomal chromosomes.

4.2.3 Genetic score association

Four scores were defined to measure genetic burden of X chromosome haplotypes in females. Each score was calculated using SNPs and indels in the SHAPEIT2 phased VCF files provided by GTEx [2]. Each score was calculated using all heterozygous variants with physical positions that did not fall within the 117 genomic windows corresponding to fully inactivated genes used to calculate skew. First, we considered the difference in average SNP CADD score [85] between the two haplotypes. A higher CADD score corresponds to an increased likelihood of deleteriousness. CADD Phred scores for SNPs were retrieved using VEP [86]. The CADD score associated with all alternate SNP alleles occuring on a haplotype was averaged.

CADD burden scores were defined as the average CADD score of haplotype H1 subtracted from the average CADD score of haplotype H2.

The remaining scores were proportions of different types of mutations carried by the H1 haplotype. We considered the proportion of alternate alleles (both SNPs and indels), missense SNPs, and synonymous SNPs carried by a haplotype. This proportion is calculated as the count of the variant-type on haplotype H1 divided by the total number of these variants found across both haplotypes. Annotations indicating the coding consequence of variants were obtained using VEP.

We associated these scores with XCI skew using a linear mixed model. Specifically, we used the lmer function from the lmerTest package [87] to fit the following model across all tissues, where the (1|Grouping) notation indicates a random intercept for the specified grouping:

$$\text{H1 XCI Skew} \sim \text{H1 Burden} + \text{Age} + (1|\text{Individual}) + (1|\text{Tissue}) \tag{4.1}$$

In this model, we include random intercepts for the individual and tissue of origin for each sample. The significance of these associations were determined using a one-sided t-test, under the assumption that increased burden will decrease skewing towards a haplotype. The p-values were calculated as the value of the cumulative density function of the t-distribution given the t-values and degrees of freedom returned by the lmer function.

Blood-related polygenic scores were retrieved from PGS Catalog [88]. We used scores for the following cell counts: platelet (PGS000186), red blood cell (PGS000187), basophil (PGS000163), neutrophil (PGS000182), eosinophil (PGS000165), monocyte (PGS000177), and lymphocyte (PGS000172) [89]. GRCh38 positions of the variants used for these scores were retrieved from SNP Nexus [90]. We calculated each polygenic score for both haplotypes as the linear combination of the score weights and the corresponding effect alleles if they occur on the haplotype. We modeled the difference in polygenic scores as follows:

$$\text{H1 XCI Skew} \sim (\text{H1 PGS - H2 PGS}) + \text{Age} + (1|\text{Individual}) + (1|\text{Tissue}) \tag{4.2}$$

The significance of these associations were determined by using a one-sided t-test, under the assumption that increased proliferative potential will increase skewing towards a haplotype. Specifically, p-values were calculated by subtracting the cumulative density function of the t-distribution given the t-values and degrees of freedom from one.

Each of the scores described above were calculated for autosomal haplotypes using variants that occur on the p-arm of each non-acrocentric chromosome. As controls, we repeated the association tests for these autosomes with the following model:

$$\begin{aligned}\text{H1 XCI Skew} \sim \text{(H1 PGS - H2 PGS)} + \text{Age} + \\ \text{(1|Individual)} + \text{(1|Tissue)} + \text{(1|Chromosome)}\end{aligned} \tag{4.3}$$

where "H1 score" refers to the burden and proliferation polygenic scores defined above and "Chromosome" refers to the autosome used to generate the sample. A Bonferroni-corrected significance threshold of 0.05/8 was used for the 4 burden scores and 0.05/14 was used for the 7 blood-related polygenic scores.

4.2.4 Identifying significant XCI loci

The phased X chromosome variants were filtered with PLINK 2 [71]. SNPs were extracted with a minor allele frequency threshold of 0.01 and Hardy-Weinberg equilibrium p-value threshold of 1e-6. LD pruning was performed with a window size of 100 kilobases, step size of 5 variants, and r^2 threshold of 0.5.

These filtered variants were marginally tested in a linear mixed model to measure association with absolute XCI skew. We fit the following linear mixed model:

$$\text{Absolute XCI Skew} \sim \text{Heterozygous} + \text{Age} + \text{(1|Individual)} + \text{(1|Tissue)} \tag{4.4}$$

where Heterozygous is 1 if the sample is heterozygous for the variant and 0 if the sample is homozygous. We modeled absolute skewing under the assumption that heterozygosity will lead to differences in fitness and consequently increased overall skewing. Given this

assumption, we determined the significance of these associations using a one-sided t-test under this hypothesis of a positive effect size if one exists. The resulting p-values were converted to q-values [91] and considered significant using a threshold corresponding to a 0.05 local false discovery rate.

At the two significant loci, this linear mixed model was also used to determine the significance of the difference in skewing between homozygous samples and groups defined by the phasing of the variant in heterozygous individuals. For these associations, we fit the following model separately for individuals with the variant on H1 and on H2:

$$\text{H1 XCI Skew} \sim \text{Heterozygous} + \text{Age} + (1|\text{Individual}) + (1|\text{Tissue}) \tag{4.5}$$

Since we do not assume the direction of individual variant effects on XCI skewing, we performed a two-sided t-test to determine significance of these associations.

4.2.5 Associating XCI-linked genetics in males

To address the possibility that the burden and proliferation scores are capturing regulatory effects, such as downregulation of a haplotype rather than increased inactivation, we calculated these scores within males. Like in the female samples, we utilized variants with physical positions that did not fall within the windows of the 117 fully inactivated genes used to calculate skew. For each individual, only variants reported as homozygous in the VCF files were considered, since heterozygous calls should not occur in the males with one X chromosome. Given that males have one X chromosome, we associated these scores with the expression of each gene on the X chromosome. We fit a model similar to those used for identifying expression quantitative trait loci. Specifically, we performed a linear regression within each tissue independently, correcting for the top 5 genetic principal components, PCR, and platform. We also corrected for PEER factors [92], using 15 factors for sample sizes less than 150 and adding an additional 15 factors for each 100 samples. The resulting p-values from the linear regression across all tissues were converted to q-values and considered significant

at a local false discovery rate threshold of 0.05.

4.3 Results

4.3.1 Estimating XCI skew from RNA-seq data

We analyzed data generated by the GTEx consortium [2], which included RNA-seq and phased whole genome sequencing data measured across 243 caucasian female samples and 472 caucasian male samples. We measured XCI skew in females as the difference in expression measured from each X chromosome in an RNA-seq experiment. Since XCI does not completely silence the entire X chromosome, we restricted our analysis to genes that have been previously observed to be fully inactivated [83, 84, 93]. Read counts at heterozygous sites in these regions were assigned to haplotypes determined with phased whole genome sequencing data. The skew in RNA-seq reads was calculated at each site as the number of reads coming from one haplotype divided by the total number of reads. XCI skew for a sample was calculated as the median of these values as done by previous work [79]. A median of 45 heterozygous sites were used for this calculation across the samples (Figure 4.2). We observed variability in XCI skew across tissues, with EBV-transformed lymphocytes exhibiting nearly complete skew in many samples (Figure 4.1b). Given this extreme skewing likely due to culture conditions, we removed these samples from downstream analyses. Furthermore, we observed correlation of XCI skew across the different tissues (Figure 4.3).

Inferring XCI skew from expression data can be confounded by both technological and biological factors that may lead to similar associations with our genetic burden scores. Reference bias, the tendency for RNA-seq reads with reference alleles to better map to a reference genome than non-reference alleles [94], may be a significant confounding variable. This bias is especially relevant to the genetic burden defined by the proportion of alternate alleles carried by a haplotype. For example, an alternate allele may have lower observed expression than the reference allele simply due to reference bias rather than XCI skew. To account for this source of bias, we utilized RNA-seq reads that were aligned with a filter utilizing

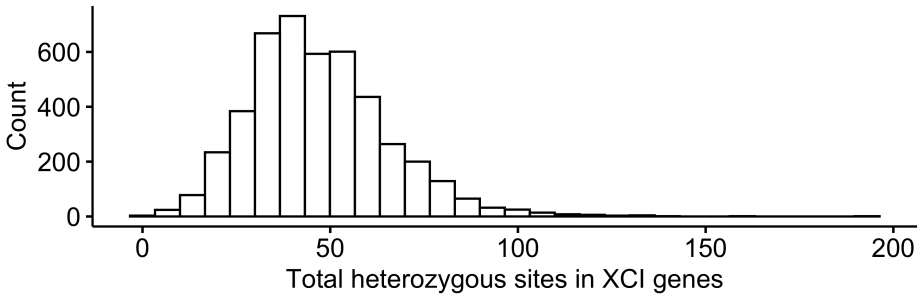

Figure 4.2: Histogram of number of heterozygous sites with coverage of 10 or more RNA-seq reads and within fully inactivated genes used to calculate XCI skew for each sample. We observed a mean of 46.805 and median of 45 variants used to calculate skew (the median of per-site skew).

WASP, a method that accounts for reference bias by mapping reads with swapped genotypes [95]. Regulatory effects of genetic variation may also lead to imbalances in expression [96] measured from each X chromosome that may be erroneously attributed to skew in X chromosome inactivation. To address this source of confounding, we performed similar experiments using the non-acrocentric autosomes (chromosomes with arms of roughly equal size) of the female samples. In short, we calculated skewing at heterozygous sites that were captured by RNA-seq reads on the q or long arm of the autosome. We observed no significant level of average 'skewing' in these chromosomes (Figure 4.4).

We also performed an analysis of the male GTEx samples to test the possibility of regulatory effects rather than skewing in inactivation. Since males only have one X chromosome, we associated genetic features with expression of each gene on the X chromosome. If our genetic scores are associated with downregulation of expression rather than XCI skew, we anticipated this experiment to show a significant negative association with the expression of X chromosome genes across males.

4.3.2 Genetic burden is associated with XCI skew

We defined the following genetic burden scores for this analysis: the proportion of missense and synonymous mutations carried by each haplotype, and the difference in average CADD score [85] between the two haplotypes. CADD scores provide a quantitative measure of

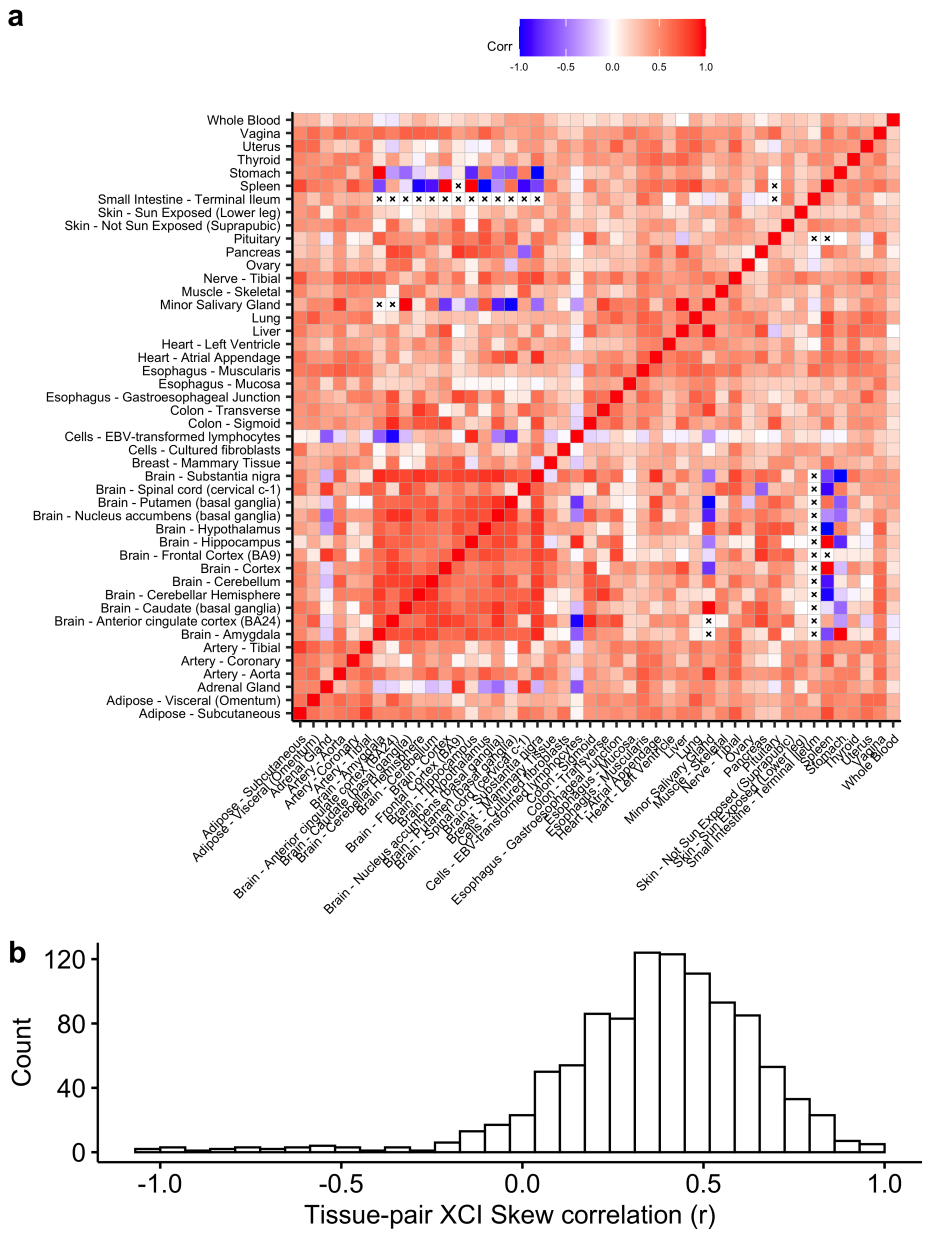

Figure 4.3: Pearson correlation of XCI skew between tissues. **a**, Correlation matrix of tissue-specific XCI skew. White boxes with 'X' symbol indicate that less than 25 observations were available for the tissue pair. **b**, Histogram of correlation between 1,017 non-identical tissue pairs. We observed a mean correlation of 0.3663 and median of 0.3992.

the predicted deleteriousness of a variant. Only genetic variation with genomic positions outside of the gene windows that were used to calculate XCI skew were considered in these scores to avoid capturing cis regulatory effects. To associate these burden measures with XCI skew, we fit a linear mixed model with random intercepts accounting for the individual

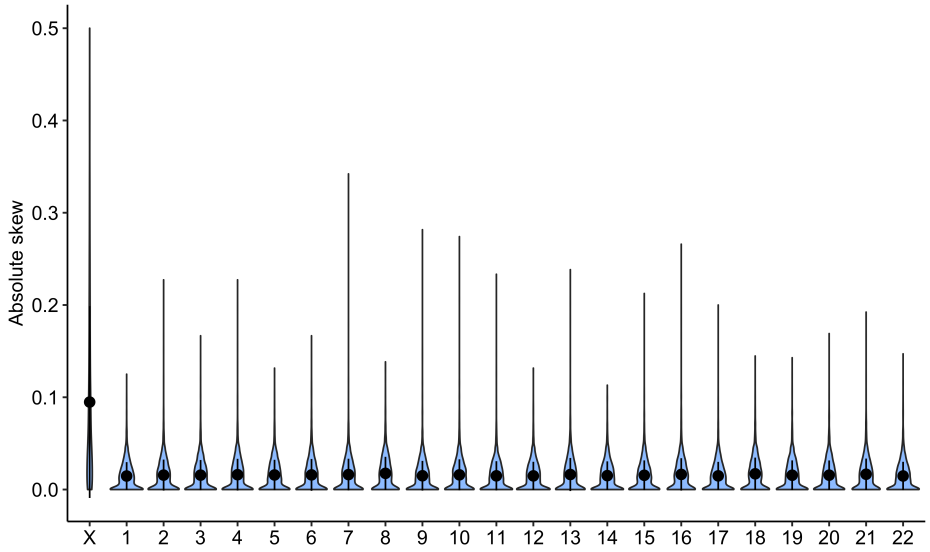

Figure 4.4: Estimated absolute skew in expression at heterozygous sites on the X chromosome and non-acrocentric autosomes. On the X chromosome, this value is used as the estimate of inactivation skewing since heterozygous sites are within fully inactivated genes. On the autosomes, an equal number of heterozygous sites used on the X chromosome for each sample were randomly selected from the q-arm to estimate the median skew in expression. Dots indicate mean of absolute skew with vertical lines indicating one standard deviation.

of origin for each tissue sample and tissue type. In this model, the difference in CADD score between haplotypes had a significant negative effect on XCI skew (coefficient = -0.1289 $\pm$ 0.0451, p = 0.0023) (Figure 4.5a). This result suggests that a higher genetic burden in terms of deleterious variation is associated with higher rates of inactivation of the haplotype. We found that this association was not significant across the non-acrocentric autosomes (Table 4.1). In addition, these burden scores were not associated with the regulation of gene expression on the X chromosome in male samples.

4.3.3 Variation in proliferation-related polygenic scores is associated with XCI Skew

We utilized publicly available polygenic risk score weights to calculate the proliferative potential of each haplotype in the female samples. Specifically, we used various blood cell counts as phenotypes under the assumption that these scores are concordant with hematopoiesis

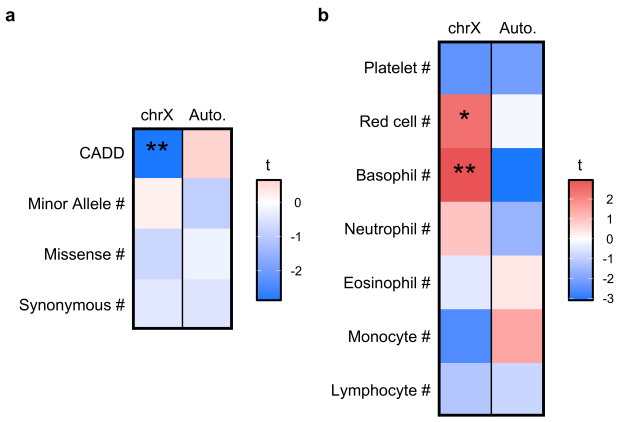

Figure 4.5: Association between genetic scores and skew estimated from expression at heterozygous sites on X chromosome (chrX) and non-acrocentric autosomes (Auto.). Colors indicate the relative t-score from a linear mixed model association test with an asterisks (*) indicating significance at $\alpha = 0.05$ and double asterisks (**) indicating significance after Bonferroni correction. **a**, Association results of genetic burden scores with estimated skew. A one-sided t-test was performed under the assumption that increased genetic burden decreases skew towards the haplotype. CADD indicates difference in mean CADD score of each haplotype. The remaining scores compare the number of the indicated mutations on each haplotype. **b**, Association results of proliferative polygenic scores with estimated skew. Counts of different blood cell types were used as a proxy for proliferative potential. A one-sided t-test was performed under the assumption that increased proliferation genetic scores will increase skew towards the haplotype.

and proliferative cell activity [97]. We found that the haplotype with the higher red blood cell polygenic score (coefficient = 0.1116 ± 0.0462, p = 0.0082) and basophil polygenic score (coefficient = 0.1355 ± 0.0460, p = 0.0018) tends to be overrepresented in terms of XCI skewing (Figure 4.5b). These results imply that higher proliferative-related polygenic scores are associated with an enrichment for cells with this haplotype active across tissues. We also ran similar autosomal and male analyses to test the alternative hypothesis that these polygenic scores may be capturing gene regulatory effects. We found no significant associations between these scores in the female non-acrocentric autosomes (Table 4.2). In the analysis of male samples, the basophil polygenic score was significantly associated with increased expression of AFF2 in 4 tissues (Table 4.3). However, this gene is considered a variable XCI escape gene and therefore was not used to calculate XCI skew in the female samples.

Table 4.1: Burden associations with skewing in XCI and autosomal q-arm expression

Burden Score	Chromosome(s)	β (s.e.)	df	t	P
CADD	chrX	-0.1289 (0.0451)	240.6684	-2.8579	0.0023
	Auto.	0.0024 (0.0037)	60762.2663	0.6516	0.7427
Minor Allele #	chrX	0.0097 (0.0453)	240.5593	0.2141	0.5847
	Auto.	-0.0034 (0.0037)	58076.8103	-0.9263	0.1772
Missense #	chrX	-0.0355 (0.0476)	238.7277	-0.7452	0.2285
	Auto.	-0.0009 (0.0037)	59330.7798	-0.2314	0.4085
Synonymous #	chrX	-0.0202 (0.0463)	239.3741	-0.4356	0.3318
	Auto.	-0.0018 (0.0037)	39637.5931	-0.4748	0.3175

Table 4.2: Proliferation-related polygenic score associations with skewing in XCI and autosomal q-arm expression

Polygenic Score	Chromosome(s)	β (s.e.)	df	t	P
Platelet #	chrX	-0.1093 (0.0456)	240.8663	-2.3997	0.9914
	Auto.	-0.0079 (0.0037)	58466.4918	-2.1498	0.9842
Red cell #	chrX	0.1116 (0.0462)	238.7682	2.4172	0.0082
	Auto.	-0.0006 (0.0037)	54155.3786	-0.1623	0.5645
Basophil #	chrX	0.1355 (0.0460)	237.3432	2.9487	0.0018
	Auto.	-0.0113 (0.0037)	60017.3402	-3.0875	0.9990
Neutrophil #	chrX	0.0451 (0.0469)	238.1159	0.9606	0.1689
	Auto.	-0.0059 (0.0037)	57022.3082	-1.6151	0.9469
Eosinophil #	chrX	-0.0224 (0.0475)	238.1788	-0.4710	0.6810
	Auto.	0.0014 (0.0037)	60407.8660	0.3883	0.3489
Monocyte #	chrX	-0.1209 (0.0463)	237.1836	-2.6126	0.9952
	Auto.	0.0055 (0.0037)	50177.1222	1.4848	0.0688
Lymphocyte #	chrX	-0.0560 (0.0473)	238.1101	-1.1834	0.8811
	Auto.	-0.0031 (0.0037)	57465.9629	-0.8419	0.8001

4.3.4 Variation in specific loci is significantly associated with XCI skew

While the previous analyses suggest that genetic burden and proliferative potential is associated with skewed XCI, they do not provide specific variants or loci that contribute to this association. We performed an association study on the X chromosome to address this question. Specifically, we associated variants on the X chromosome with the absolute XCI skew measurements (Figure 4.6a). We found 2 variants that were significant at a FDR of 0.05 (Table 4.4). The first variant (rs141680486) is an intronic SNP found within the DMD gene, also known as dystrophin, and was significantly associated with increased absolute

Table 4.3: Covariates associated with chromosome X gene regulation in males.

Covariate	Gene Name	Gene ID	XCI status	Tissue	β (s.e.)	P	Q
Basophil #	AFF2	ENSG00000155966.13	Variable	Adipose (Subc.)	0.1522 (0.0329)	5.701e-06	0.0490
				Lung	0.1825 (0.0374)	1.841e-06	0.0317
				Skin (Sup.)	0.1616 (0.0344)	4.353e-06	0.0490
				Thyroid	0.1499 (0.0296)	7.902e-07	0.0272
rs141680486	TMEM187	ENSG00000177854.7	Inactive	EBV lymphocytes	-0.3917 (0.0696)	8.734e-07	0.0302

XCI skew (coefficient = 0.0203 ± 0.0041, p=7.978e-07). We observed large absolute skew in samples that were heterozygous for this variant compared to those who were homozygous (Figure 4.6b). Moreover, we found that the haplotype with the minor allele tends to be more inactivated than the other haplotype in heterozygous samples (Figure 4.6c). The second variant (rs73227260) is an intronic SNP in LOC101928359, a non-coding transcript, and was also associated with increased absolute XCI skew (coefficient = 0.0199 ± 0.0041, p = 1.265e-06). We observed a high level of skewing in heterozygous individuals with no clear difference in directional skewing depending on the haplotype carrying the minor allele (Figure 4.7).

Since eQTLs may also cause skewing in expression within cells, we tested the association between these variants and gene expression on the X chromosome in males. We found that the DMD variant (rs141680486) is a significant eQTL for TMEM187 expression in EBV-transformed lymphocytes (coefficient = -0.3917 ± 0.0696, p=8.734e-07). However, we note that these lymphocyte samples were excluded from our female analyses due to extreme XCI skewing.

4.4 Discussion

These analyses support the hypothesis that differences in genetic variation on the maternal and paternal X chromosomes influences skewed inactivation through selection [80, 17]. Specifically, these results demonstrate a significant association between higher CADD scores of a haplotype and estimated skewing away from this haplotype. Furthermore, we showed that higher polygenic scores related to proliferation, specifically red blood cell and basophil

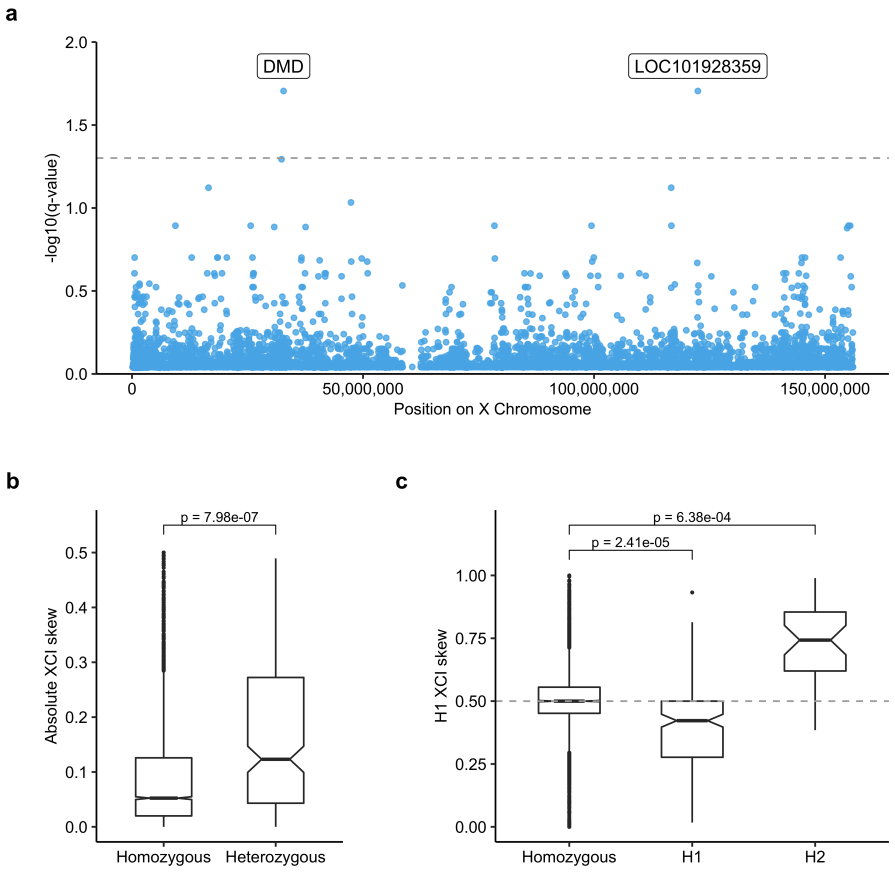

Figure 4.6: Associating specific variation on the X chromosome with inactivation skew. **a**, Manhattan plot of -log10(q-values) generated from a linear mixed model associating absolute XCI skew with heterozygous status, accounting for age as well as individual and tissue groupings of samples. A one-sided test was performed under the hypothesis that heterozygous status increases absolute skew. Dotted horizontal line indicates local false discovery rate of 0.05. **b**, Boxplot of absolute XCI skew in samples that are homozygous (n = 4,232) or heterozygous (n = 230) for the DMD variant (rs141680486). Indicated p-value is from the model described above. **c**, Boxplot of skewing toward haplotype 1 (H1), where the grouping on the x-axis describes individuals without the DMD variant (Homozygous, n = 4,232), with the variant on haplotype 1 (H1, n = 190), and with the variant on haplotype 2 (H2, n = 40). Indicated p-values are from the model described above but with a two-sided test, since we do not assume the direction of skewing associated with a specific variant.

counts, tend to be associated with skewing towards the haplotype. These results imply that skewing in X inactivation may arise from higher proliferation or reduced deleteriousness of cells with one haplotype activated compared to the other subpopulation. We also identified common variation within specific loci on the X chromosome, such as the dystrophin gene,

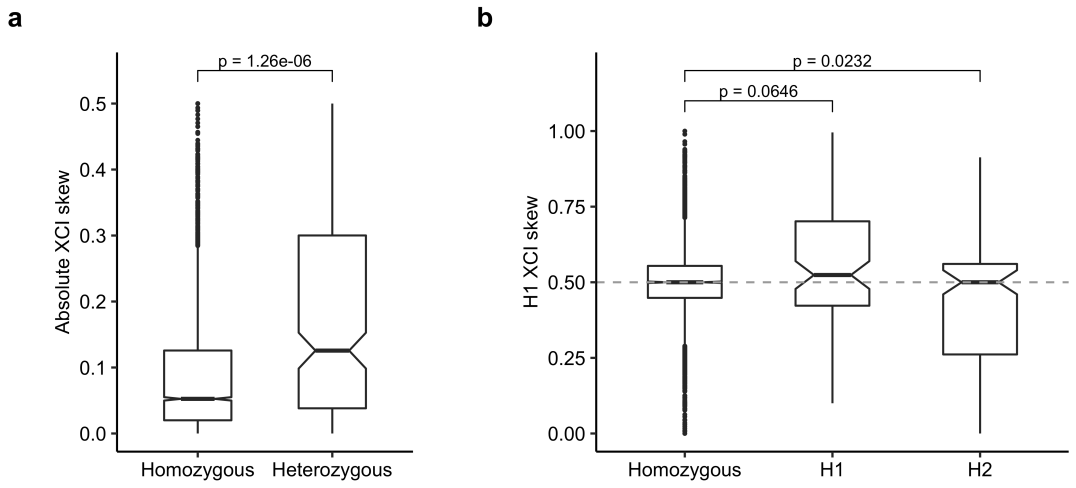

Figure 4.7: Association of LOC101928359 variant (rs73227260) with XCI skew. Indicated p-values are from a linear mixed model accounting for individual and tissue of origin. **a**, Boxplot of absolute XCI skew in homozygous samples (n = 4,230) and heterozygous samples (n = 232). Indicated p-value is from a one-sided t-test. **b**, Boxplot of skewing toward haplotype 1 (H1), where the grouping on the x-axis describes individuals without the LOC variant (Homozygous, n = 4,230), with the variant on haplotype 1 (H1, n = 92), and with the variant on haplotype 2 (H2, n = 140). Indicated p-values are from a two-sided t-test.

associated with XCI skew that may also contribute to these differences in proliferation or deleteriousness.

Several confounding factors could influence the results of this study. First, haplotype estimation by phasing algorithms may be inaccurate. Phasing errors would reduce the power of these analyses, since genetic scores may include variants from both haplotypes. In our analyses, we observed phase switches in EBV-transformed lymphocyte samples with extreme XCI skewing that suggests some phasing errors had occurred. Second, mapping biases may cause artificially higher observed expression of reference alleles compared to alternate alleles. This issue is addressed by the WASP filtering [95] used in the alignment of RNA-seq reads. Third, differences in eQTL effects across haplotypes may lead to skewed expression in RNA-seq experiments that may be incorrectly attributed to skewed XCI. We consider this source of confounding by taking the median of expression skew measured across heterozygous sites in 117 fully inactivated genes and by testing whether our features of interest are significant

Table 4.4: Top 10 variants associated with absolute skewing in XCI

Locus	SNP	Position	Alleles	AF	β (s.e.)	$P_{\text{one-sided}}$
DMD	rs141680486	32794582	C/G	0.0179	0.0203 (0.0041)	7.978e-07
LOC101928359	rs73227260	122471389	C/T	0.0203	0.0199 (0.0041)	1.265e-06
DMD	rs144615018	32355895	C/G	0.0328	0.0175 (0.0039)	4.891e-06
None	rs140456300	116700067	A/T	0.0316	0.0157 (0.0036)	9.868e-06
None	rs113265091	16560720	T/C	0.0185	0.0177 (0.0041)	1.21e-05
ZNF157	rs140428205	47388509	G/A	0.0215	0.0174 (0.0041)	1.782e-05
None	rs60102760	9394922	T/G	0.0263	0.0168 (0.0041)	2.917e-05
FUNDC2	rs782070621	155057031	T/G	0.0369	0.0169 (0.0042)	3.742e-05
None	rs139564137	155434572	C/T	0.0358	0.0169 (0.0042)	3.742e-05
None	rs6623805	78439230	A/T	0.9475	0.0159 (0.004)	4.263e-05

in the non-acrocentric autosomes or as eQTLs in the male samples. These issues could be circumvented with sufficiently large single-cell RNA-seq datasets. Haplotype phasing, as well as inactivation status, could be estimated within each cell individually and XCI skew could be calculated by simply counting cells. Furthermore, these data would allow cell-type-specific analysis of the relationship between genetics and skewed XCI.

The influence of genetics on XCI skewing across the general female population highlights interesting mechanisms that should be further explored. For example, the effect of a specific variant on a complex phenotype may be dampened or amplified depending on the haplotype that carries it in females. For example, if an individual homozygous for a disease-related variant carries it on a haplotype with reduced proliferation, its effect may be dampened since a majority of cells will be expressing the wild-type variant. Epistatic effects and reduced penetrance due to allele-specific expression has been previously reported in the context of allele-specific gene regulation [96, 98]. Here, these effects would arise from skewed X inactivation across a tissue rather than skewed expression within individual cells. This effect on penetrance has been described in the context of X-linked disorders, where skewed inactivation blurs the distinction of dominant and recessive traits across males and females [99]. The results shown here imply that selective pressures on genetic differences across the female population can influence skewing and consequently influence the penetrance of variation across the X chromosome.

REFERENCES

[1] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 10 2018.

[2] GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, September 2020.

[3] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W. J. H. Penninx, Rick Jansen, Eco J. C. de Geus, Dorret I. Boomsma, Fred A. Wright, Patrick F. Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J. Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T. Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L. Price, Päivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, Mar 2016.

[4] Peter M. Visscher, Matthew A. Brown, Mark I. McCarthy, and Jian Yang. Five years of gwas discovery. *American journal of human genetics*, 90(1):7–24, 01 2012. 22243964[pmid].

[5] Alexandra C. Nica and Emmanouil T. Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1620):20120362–20120362, May 2013. 23650636[pmid].

[6] Alison A. Motsinger-Reif, Eric Jorgenson, Mary V. Relling, Deanna L. Kroetz, Richard Weinshilboum, Nancy J. Cox, and Dan M. Roden. Genome-wide association studies in pharmacogenomics: successes and lessons. *Pharmacogenetics and genomics*, 23(8):383–394, Aug 2013. PMC3003940[pmcid].

[7] Michael G Walker. Pharmaceutical target identification by gene expression analysis. *Mini reviews in medicinal chemistry*, 1(2):197–205, 2001.

[8] Marjan Farahbod and Paul Pavlidis. Untangling the effects of cellular composition on coexpression analysis. *Genome research*, 30(6):849–859, 2020.

[9] Harm-Jan Westra, Danny Arends, Tõnu Esko, Marjolein J Peters, Claudia Schurmann, Katharina Schramm, Johannes Kettunen, Hanieh Yaghootkar, Benjamin P Fairfax, Anand Kumar Andiappan, Yang Li, Jingyuan Fu, Juha Karjalainen, Mathieu Platteel, Marijn Visschedijk, Rinse K Weersma, Silva Kasela, Lili Milani, Liina Tserel, Pärt Peterson, Eva Reinmaa, Albert Hofman, André G Uitterlinden, Fernando Rivadeneira, Georg Homuth, Astrid Petersmann, Roberto Lorbeer, Holger Prokisch, Thomas Meitinger, Christian Herder, Michael Roden, Harald Grallert, Samuli Ripatti, Markus Perola, Andrew R Wood, David Melzer, Luigi Ferrucci, Andrew B Singleton, Dena G Hernandez,

Julian C Knight, Rossella Melchiotti, Bernett Lee, Michael Poidinger, Francesca Zolezzi, Anis Larbi, De Yun Wang, Leonard H van den Berg, Jan H Veldink, Olaf Rotzschke, Seiko Makino, Veikko Salomaa, Konstantin Strauch, Uwe Völker, Joyce B J van Meurs, Andres Metspalu, Cisca Wijmenga, Ritsert C Jansen, and Lude Franke. Cell specific eQTL analysis without sorting cells. *PLoS Genet.*, 11(5):e1005223, May 2015.

[10] 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, 2010.

[11] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[12] Maria-Ioanna Christodoulou, Margaritis Avgeris, Ioanna Kokkinopoulou, Eirini Maratou, Panayota Mitrou, Christos K. Kontos, Efthimios Pappas, Eleni Boutati, Andreas Scorilas, and Emmanuel G. Fragoulis. Blood-based analysis of type-2 diabetes mellitus susceptibility genes identifies specific transcript variants with deregulated expression and association with disease risk. *Scientific Reports*, 9(1):1512, Feb 2019.

[13] Zong Miao, Marcus Alvarez, Arthur Ko, Yash Bhagat, Elior Rahmani, Brandon Jew, Sini Heinonen, Linda Liliana Muñoz-Hernandez, Miguel Herrera-Hernandez, Carlos Aguilar-Salinas, Teresa Tusie-Luna, Karen L. Mohlke, Markku Laakso, Kirsi H. Pietiläinen, Eran Halperin, and Päivi Pajukanta. The causal effect of obesity on prediabetes and insulin resistance reveals the important role of adipose tissue in insulin resistance. *PLOS Genetics*, 16(9):1–23, 09 2020.

[14] Zhe Zhou, Yizhang Zhu, Yang Liu, and Yuxin Yin. Comprehensive transcriptomic analysis indicates brain regional specific alterations in type 2 diabetes. *Aging*, 11(16):6398–6421, Aug 2019. 31449493[pmid].

[15] Francis Robert and Jerry Pelletier. Exploring the impact of single-nucleotide polymorphisms on translation. *Frontiers in genetics*, 9:507, 2018.

[16] Marco Garieri, Olivier Delaneau, Federico Santoni, Richard J. Fish, David Mull, Piero Carninci, Emmanouil T. Dermitzakis, Stylianos E. Antonarakis, and Alexandre Fort. The effect of genetic variation on promoter usage and enhancer activity. *Nature Communications*, 8(1):1358, Nov 2017.

[17] B R Migeon. Non-random X chromosome inactivation in mammalian cells, 1998.

[18] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.*, 19(1A):A68–77, 2015.

[19] GTEx Consortium. Human genomics. the Genotype-Tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015.

[20] Oskar Bruning, Wendy Rodenburg, Paul F K Wackers, Conny van Oostrom, Martijs J Jonker, Rob J Dekker, Han Rauwerda, Wim A Ensink, Annemieke de Vries, and Timo M Breit. Confounding factors in the transcriptome analysis of an In-Vivo exposure experiment. *PLoS One*, 11(1):e0145252, January 2016.

[21] Wolf Herman Fridman, Franck Pagès, Catherine Sautès-Fridman, and Jérôme Galon. The immune contexture in human tumours: impact on clinical outcome. *Nat. Rev. Cancer*, 12(4):298–306, March 2012.

[22] J Rahier, R M Goebbels, and J C Henquin. Cellular composition of the human diabetic pancreas, 1983.

[23] Ping Hu, Wenhua Zhang, Hongbo Xin, and Glenn Deng. Single cell isolation and analysis, 2016.

[24] Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, January 2017.

[25] Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, Osnat Penn, Trygve Bakken, Vilas Menon, Jeremy Miller, Olivia Fong, Karla E Hirokawa, Kanan Lathia, Christine Rimorin, Michael Tieu, Rachael Larsen, Tamara Casper, Eliza Barkan, Matthew Kroll, Sheana Parry, Nadiya V Shapovalova, Daniel Hirschstein, Julie Pendergraft, Heather A Sullivan, Tae Kyung Kim, Aaron Szafer, Nick Dee, Peter Groblewski, Ian Wickersham, Ali Cetin, Julie A Harris, Boaz P Levi, Susan M Sunkin, Linda Madisen, Tanya L Daigle, Loren Looger, Amy Bernard, John Phillips, Ed Lein, Michael Hawrylycz, Karel Svoboda, Allan R Jones, Christof Koch, and Hongkui Zeng. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, November 2018.

[26] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, May 2015.

[27] Jingshu Wang, Mo Huang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, John Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 115(28):E6437–E6446, July 2018.

[28] Shahin Mohammadi, Neta Zuckerman, Andrea Goldsmith, and Ananth Grama. A critical survey of deconvolution methods for separating cell types in complex tissues, 2017.

[29] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, 12(5):453–457, May 2015.

[30] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, Douglas A Melton, and Itai Yanai. A Single-Cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst*, 3(4):346–360.e4, October 2016.

[31] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.*, 10(1):380, January 2019.

[32] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, 65(4):631–643.e4, February 2017.

[33] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, August 2018.

[34] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36(5):411–420, June 2018.

[35] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression.

[36] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection, 2018.

[37] Oscar Franzén, Li-Ming Gan, and Johan L M Björkegren. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, 2019, January 2019.

[38] Montserrat Esteve Ràfols. Adipose tissue: cell heterogeneity and functional diversity. *Endocrinol. Nutr.*, 61(2):100–112, February 2014.

[39] Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, Maximilian Diehn, and Ash A Alizadeh. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.*, May 2019.

[40] Evan D Rosen and Bruce M Spiegelman. What we talk about when we talk about fat, 2014.

[41] Craig A Glastonbury, Alexessander Couto Alves, Julia El-Sayed Moustafa, and Kerrin S Small. Cell-type heterogeneity in adipose tissue is associated with complex traits and reveals disease-relevant cell-specific eQTLs.

[42] Kirsty L Spalding, Erik Arner, Pål O Westermark, Samuel Bernard, Bruce A Buchholz, Olaf Bergmann, Lennart Blomqvist, Johan Hoffstedt, Erik Näslund, Tom Britton, Hernan Concha, Moustapha Hassan, Mikael Rydén, Jonas Frisén, and Peter Arner. Dynamics of fat cell turnover in humans, 2008.

[43] Stuart P Weisberg, Daniel McCann, Manisha Desai, Michael Rosenbaum, Rudolph L Leibel, and Anthony W Ferrante. Obesity is associated with macrophage accumulation in adipose tissue, 2003.

[44] Tracey McLaughlin, Li-Fen Liu, Cindy Lamendola, Lei Shen, John Morton, Homero Rivas, Daniel Winer, Lorna Tolentino, Okmi Choi, Hong Zhang, Melissa Ch'ng, and Edgar Engleman. T-Cell profile in adipose tissue is associated with insulin resistance and systemic inflammation in humans. *Arterioscler. Thromb. Vasc. Biol.*, 34(12):2637, December 2014.

[45] Manish Gutch, Sukriti Kumar, Syedmohd Razi, Kumarkeshav Gupta, and Abhinav Gupta. Assessment of insulin sensitivity/resistance, 2015.

[46] Sara Mostafavi, Chris Gaiteri, Sarah E Sullivan, Charles C White, Shinya Tasaki, Jishu Xu, Mariko Taga, Hans-Ulrich Klein, Ellis Patrick, Vitalina Komashko, Cristin McCabe, Robert Smith, Elizabeth M Bradshaw, David E Root, Aviv Regev, Lei Yu, Lori B Chibnik, Julie A Schneider, Tracy L Young-Pearse, David A Bennett, and Philip L De Jager. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of alzheimer's disease. *Nat. Neurosci.*, 21(6):811–819, June 2018.

[47] Ellis Patrick, Mariko Taga, Ayla Ergun, Bernard Ng, William Casazza, Maria Cimpean, Christina Yung, Julie A Schneider, David A Bennett, Chris Gaiteri, Philip L De Jager, Elizabeth M Bradshaw, and Sara Mostafavi. Deconvolving the contributions of cell-type heterogeneity on cortical gene expression.

[48] Bruce A Yankner. Mechanisms of neuronal degeneration in alzheimer's disease, 1996.

[49] David V Hansen, Jesse E Hanson, and Morgan Sheng. Microglia in alzheimer's disease. *J. Cell Biol.*, 217(2):459–472, February 2018.

[50] Victoria Navarro, Elisabeth Sanchez-Mejias, Sebastian Jimenez, Clara Muñoz-Castro, Raquel Sanchez-Varo, Jose C Davila, Marisa Vizuete, Antonia Gutierrez, and Javier Vitorica. Microglia in alzheimer's disease: Activated, dysfunctional or degenerative, 2018.

[51] Max Schelker, Sonia Feau, Jinyan Du, Nav Ranu, Edda Klipp, Gavin MacBeath, Birgit Schoeberl, and Andreas Raue. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.*, 8(1):2032, December 2017.

[52] Quy H Nguyen, Nicholas Pervolarakis, Kevin Nee, and Kai Kessenbrock. Experimental considerations for Single-Cell RNA sequencing approaches. *Front Cell Dev Biol*, 6:108, September 2018.

[53] Haojia Wu, Yuhei Kirita, Erinn L Donnelly, and Benjamin D Humphreys. Advantages of Single-Nucleus over Single-Cell RNA sequencing of adult kidney: Rare cell types and novel cell states revealed in fibrosis. *J. Am. Soc. Nephrol.*, 30(1):23–32, January 2019.

[54] Trygve E Bakken, Rebecca D Hodge, Jeremy A Miller, Zizhen Yao, Thuc Nghi Nguyen, Brian Aevermann, Eliza Barkan, Darren Bertagnolli, Tamara Casper, Nick Dee, Emma Garren, Jeff Goldy, Lucas T Graybuck, Matthew Kroll, Roger S Lasken, Kanan Lathia, Sheana Parry, Christine Rimorin, Richard H Scheuermann, Nicholas J Schork, Soraya I Shehata, Michael Tieu, John W Phillips, Amy Bernard, Kimberly A Smith, Hongkui Zeng, Ed S Lein, and Bosiljka Tasic. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One*, 13(12):e0209648, December 2018.

[55] The All of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019. PMID: 31412182.

[56] François Aguet, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, YoSon Park, Princy Parsana, Ayellet V. Segrè, Benjamin J. Strober, Zachary Zappala, Beryl B. Cummings, Ellen T. Gelfand, Kane Hadley, Katherine H. Huang, Monkol Lek, Xiao Li, Jared L. Nedzel, Duyen Y. Nguyen, Michael S. Noble, Timothy J. Sullivan, Taru Tukiainen, Daniel G. MacArthur, Gad Getz, Anjene Addington, Ping Guan, Susan Koester, A. Roger Little, Nicole C. Lockhart, Helen M. Moore, Abhi Rao, Jeffery P. Struewing, Simona Volpi, Lori E. Brigham, Richard Hasz, Marcus Hunter, Christopher Johns, Mark Johnson, Gene Kopen, William F. Leinweber, John T. Lonsdale, Alisa McDonald, Bernadette Mestichelli, Kevin Myer, Bryan Roe, Michael Salvatore, Saboor Shad, Jeffrey A. Thomas, Gary Walters, Michael Washington, Joseph Wheeler, Jason Bridge, Barbara A. Foster, Bryan M. Gillard, Ellen Karasik, Rachna Kumar, Mark Miklos, Michael T. Moser, Scott D. Jewell, Robert G. Montroy, Daniel C. Rohrer, Dana Valley, Deborah C. Mash, David A. Davis, Leslie Sobin, Mary E. Barcus, Philip A. Branton, Nathan S. Abell, Brunilda Balliu, Olivier Delaneau, Laure Frésard, Eric R. Gamazon, Diego Garrido-Martín, Ariel D. H. Gewirtz, Genna Gliner, Michael J. Gloudemans, Buhm Han, Amy Z. He, Farhad Hormozdiari, Xin Li, Boxiang Liu, Eun Yong Kang, Ian C. McDowell, Halit Ongen, John J. Palowitch, Christine B. Peterson, Gerald Quon, Stephan Ripke, Ashis Saha, Andrey A. Shabalin, Tyler C.

Shimko, Jae Hoon Sul, Nicole A. Teran, Emily K. Tsang, Hailei Zhang, Yi-Hui Zhou, Carlos D. Bustamante, Nancy J. Cox, Roderic Guigó, Manolis Kellis, Mark I. McCarthy, Donald F. Conrad, Eleazar Eskin, Gen Li, Andrew B. Nobel, Chiara Sabatti, Barbara E. Stranger, Xiaoquan Wen, Fred A. Wright, Kristin G. Ardlie, Emmanouil T. Dermitzakis, Tuuli Lappalainen, Robert E. Handsaker, Seva Kashin, Konrad J. Karczewski, Duyen T. Nguyen, Casandra A. Trowbridge, Ruth Barshir, Omer Basha, Alexis Battle, Gireesh K. Bogu, Andrew Brown, Christopher D. Brown, Lin S. Chen, Colby Chiang, Farhan N. Damani, Barbara E. Engelhardt, Pedro G. Ferreira, Ariel D.H. Gewirtz, Roderic Guigo, Ira M. Hall, Cedric Howald, Hae Kyung Im, Eun Yong Kang, Yungil Kim, Sarah Kim-Hellmuth, Serghei Mangul, Jean Monlong, Stephen B. Montgomery, Manuel Muñoz-Aguirre, Anne W. Ndungu, Dan L. Nicolae, Meritxell Oliva, Nikolaos Panousis, Panagiotis Papasaikas, Anthony J. Payne, Jie Quan, Ferran Reverter, Michael Sammeth, Alexandra J. Scott, Reza Sodaei, Matthew Stephens, Sarah Urbut, Martijn van de Bunt, Gao Wang, Hualin S. Xi, Esti Yeger-Lotem, Judith B. Zaugg, Joshua M. Akey, Daniel Bates, Joanne Chan, Melina Claussnitzer, Kathryn Demanelis, Morgan Diegel, Jennifer A. Doherty, Andrew P. Feinberg, Marian S. Fernando, Jessica Halow, Kasper D. Hansen, Eric Haugen, Peter F. Hickey, Lei Hou, Farzana Jasmine, Ruiqi Jian, Lihua Jiang, Audra Johnson, Rajinder Kaul, Muhammad G. Kibriya, Kristen Lee, Jin Billy Li, Qin Li, Jessica Lin, Shin Lin, Sandra Linder, Caroline Linke, Yaping Liu, Matthew T. Maurano, Benoit Molinie, Jemma Nelson, Fidencio J. Neri, Yongjin Park, Brandon L. Pierce, Nicola J. Rinaldi, Lindsay F. Rizzardi, Richard Sandstrom, Andrew Skol, Kevin S. Smith, Michael P. Snyder, John Stamatoyannopoulos, Hua Tang, Li Wang, Meng Wang, Nicholas Van Wittenberghe, Fan Wu, Rui Zhang, Concepcion R. Nierras, Latarsha J. Carithers, Jimmie B. Vaught, Sarah E. Gould, Nicole C. Lockart, Casey Martin, Anjene M. Addington, Susan E. Koester, GTEx Consortium, Lead analysts:, Data Analysis & Coordinating Center (LDACC): Laboratory, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, Data Analysis & Coordinating Center (LDACC)-Analysis Working Group Laboratory, Statistical Methods groups-Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, and Biospecimen Collection Source Site-NDRI. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 10 2017.

[57] Vardhman K. Rakyan, Thomas A. Down, David J. Balding, and Stephan Beck. Epigenome-wide association studies for common human diseases. *Nature reviews. Genetics*, 12(8):529–541, 07 2011. 21747404[pmid].

[58] Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[59] Jae Hoon Sul, Buhm Han, Chun Ye, Ted Choi, and Eleazar Eskin. Effectively identifying eqtls from multiple tissues by combining mixed model and meta-analytic approaches. *PLOS Genetics*, 9(6):1–13, 06 2013.

[60] Jong Wha J Joo, Eun Yong Kang, Elin Org, Nick Furlotte, Brian Parks, Farhad Hormozdiari, Aldons J Lusis, and Eleazar Eskin. Efficient and Accurate Multiple-Phenotype Regression Method for High Dimensional Data Considering Population Structure. *Genetics*, 204(4):1379–1390, 12 2016.

[61] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–4, 2012.

[62] Buhm Han and Eleazar Eskin. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics*, 88(5):586–598, 05 2011.

[63] Sarah M. Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, 51(1):187–195, 01 2019.

[64] S. J. Welham and R. Thompson. Likelihood ratio tests for fixed model terms using residual maximum likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):701–714, 1997.

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

[66] Jack Sherman and Winifred J. Morrison. Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. *The Annals of Mathematical Statistics*, 21(1):124 – 127, 1950.

[67] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: an open-source package for R and s+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77, March 2011.

[68] Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

[69] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, 7(3):500–507, February 2012.

[70] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[71] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 02 2015. s13742-015-0047-8.

[72] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 10 2010.

[73] Florian Privé, Hugues Aschard, Andrey Ziyatdinov, and Michael G B Blum. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, 34(16):2781–2787, 03 2018.

[74] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L Whetzel, Ridwan Amode, Jose A Guillen, Harpreet S Riat, Stephen J Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, 11 2018.

[75] Genevieve L. Wojcik, Mariaelisa Graff, Katherine K. Nishimura, Ran Tao, Jeffrey Haessler, Christopher R. Gignoux, Heather M. Highland, Yesha M. Patel, Elena P. Sorokin, Christy L. Avery, Gillian M. Belbin, Stephanie A. Bien, Iona Cheng, Sinead Cullina, Chani J. Hodonsky, Yao Hu, Laura M. Huckins, Janina Jeff, Anne E. Justice, Jonathan M. Kocarnik, Unhee Lim, Bridget M. Lin, Yingchang Lu, Sarah C. Nelson, Sung-Shim L. Park, Hannah Poisner, Michael H. Preuss, Melissa A. Richard, Claudia Schurmann, Veronica W. Setiawan, Alexandra Sockell, Karan Vahi, Marie Verbanck, Abhishek Vishnu, Ryan W. Walker, Kristin L. Young, Niha Zubair, Victor Acuña-Alonso, Jose Luis Ambite, Kathleen C. Barnes, Eric Boerwinkle, Erwin P. Bottinger, Carlos D. Bustamante, Christian Caberto, Samuel Canizales-Quinteros, Matthew P. Conomos, Ewa Deelman, Ron Do, Kimberly Doheny, Lindsay Fernández-Rhodes, Myriam Fornage, Benyam Hailu, Gerardo Heiss, Brenna M. Henn, Lucia A. Hindorff, Rebecca D. Jackson, Cecelia A. Laurie, Cathy C. Laurie, Yuqing Li, Dan-Yu Lin, Andres Moreno-Estrada, Girish Nadkarni, Paul J. Norman, Loreall C. Pooler, Alexander P. Reiner, Jane Romm, Chiara Sabatti, Karla Sandoval, Xin Sheng, Eli A. Stahl, Daniel O. Stram, Timothy A. Thornton, Christina L. Wassel, Lynne R. Wilkens, Cheryl A. Winkler, Sachi Yoneyama, Steven Buyske, Christopher A. Haiman, Charles Kooperberg, Loic Le Marchand, Ruth J. F. Loos, Tara C. Matise, Kari E. North, Ulrike Peters, Eimear E. Kenny, and Christopher S. Carlson. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762):514–518, 06 2019.

[76] Arthur Korte, Bjarni J. Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44(9):1066–1071, Sep 2012.

[77] Barbara Rakitsch, Christoph Lippert, Karsten Borgwardt, and Oliver Stegle. It is all in the noise: Efficient multi-task gaussian process inference with structured residuals. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[78] N Takagi and M Sasaki. Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature*, 256(5519):640–642, August 1975.

[79] Ekaterina Shvetsova, Alina Sofronova, Ramin Monajemi, Kristina Gagalova, Harmen H M Draisma, Stefan J White, Gijs W E Santen, Susana M Chuva de Sousa Lopes, Bastiaan T Heijmans, Joyce van Meurs, Rick Jansen, Lude Franke, Szymon M Kiełbasa, Johan T den Dunnen, Peter A C 't Hoen, BIOS consortium, and GoNL consortium. Skewed x-inactivation is common in the general female population. *Eur. J. Hum. Genet.*, 27(3):455–465, March 2019.

[80] C J Brown. Skewed x-chromosome inactivation: cause or consequence? *J. Natl. Cancer Inst.*, 91(4):304–305, February 1999.

[81] Susan S Brooks, Alissa L Wall, Christelle Golzio, David W Reid, Amalia Kondyles, Jason R Willer, Christina Botti, Christopher V Nicchitta, Nicholas Katsanis, and Erica E Davis. A novel ribosomopathy caused by dysfunction of RPL10 disrupts neurodevelopment and causes x-linked microcephaly in humans. *Genetics*, 198(2):723–733, October 2014.

[82] Antonino Zito, Matthew N Davies, Pei-Chien Tsai, Susanna Roberts, Rosa Andres-Ejarque, Stefano Nardone, Jordana T Bell, Chloe C Y Wong, and Kerrin S Small. Heritability of skewed x-inactivation in female twins is tissue-specific and associated with age. *Nat. Commun.*, 10(1):5339, November 2019.

[83] Laura Carrel and Huntington F Willard. X-inactivation profile reveals extensive variability in x-linked gene expression in females. *Nature*, 434(7031):400–404, March 2005.

[84] Allison M Cotton, Bing Ge, Nicholas Light, Veronique Adoue, Tomi Pastinen, and Carolyn J Brown. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol.*, 14(11):R122, November 2013.

[85] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, 47(D1):D886–D894, January 2019.

[86] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biol.*, 17(1):122, June 2016.

[87] Alexandra Kuznetsova, Per B Brockhoff, and Rune H B Christensen. lmertest package: Tests in linear mixed effects models, 2017.

[88] Samuel A Lambert, Laurent Gil, Simon Jupp, Scott C Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, Gad Abraham, Michael Chapman, Helen Parkinson, John Danesh, Jacqueline A L MacArthur, and Michael Inouye. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.*, 53(4):420–425, April 2021.

[89] Dragana Vuckovic, Erik L Bao, Parsa Akbari, Caleb A Lareau, Abdou Mousas, Tao Jiang, Ming-Huei Chen, Laura M Raffield, Manuel Tardaguila, Jennifer E Huffman, Scott C Ritchie, Karyn Megy, Hannes Ponstingl, Christopher J Penkett, Patrick K Albers, Emilie M Wigdor, Saori Sakaue, Arden Moscati, Regina Manansala, Ken Sin Lo, Huijun Qian, Masato Akiyama, Traci M Bartz, Yoav Ben-Shlomo, Andrew Beswick, Jette Bork-Jensen, Erwin P Bottinger, Jennifer A Brody, Frank J A van Rooij, Kumaraswamy N Chitrala, Peter W F Wilson, Hélène Choquet, John Danesh, Emanuele Di Angelantonio, Niki Dimou, Jingzhong Ding, Paul Elliott, Tõnu Esko, Michele K Evans, Stephan B Felix, James S Floyd, Linda Broer, Niels Grarup, Michael H Guo, Qi Guo, Andreas Greinacher, Jeff Haessler, Torben Hansen, Joanna M M Howson, Wei Huang, Eric Jorgenson, Tim Kacprowski, Mika Kähönen, Yoichiro Kamatani, Masahiro Kanai, Savita Karthikeyan, Fotios Koskeridis, Leslie A Lange, Terho Lehtimäki, Allan Linneberg, Yongmei Liu, Leo-Pekka Lyytikäinen, Ani Manichaikul, Koichi Matsuda, Karen L Mohlke, Nina Mononen, Yoshinori Murakami, Girish N Nadkarni, Kjell Nikus, Nathan Pankratz, Oluf Pedersen, Michael Preuss, Bruce M Psaty, Olli T Raitakari, Stephen S Rich, Benjamin A T Rodriguez, Jonathan D Rosen, Jerome I Rotter, Petra Schubert, Cassandra N Spracklen, Praveen Surendran, Hua Tang, Jean-Claude Tardif, Mohsen Ghanbari, Uwe Völker, Henry Völzke, Nicholas A Watkins, Stefan Weiss, VA Million Veteran Program, Na Cai, Kousik Kundu, Stephen B Watt, Klaudia Walter, Alan B Zonderman, Kelly Cho, Yun Li, Ruth J F Loos, Julian C Knight, Michel Georges, Oliver Stegle, Evangelos Evangelou, Yukinori Okada, David J Roberts, Michael Inouye, Andrew D Johnson, Paul L Auer, William J Astle, Alexander P Reiner, Adam S Butterworth, Willem H Ouwehand, Guillaume Lettre, Vijay G Sankaran, and Nicole Soranzo. The polygenic and monogenic basis of blood traits and diseases. *Cell*, 182(5):1214–1231.e11, September 2020.

[90] Jorge Oscanoa, Lavanya Sivapalan, Emanuela Gadaleta, Abu Z Dayem Ullah, Nicholas R Lemoine, and Claude Chelala. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res.*, 48(W1):W185–W192, July 2020.

[91] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.*, 100(16):9440–9445, August 2003.

[92] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, 7(3):500–507, February 2012.

[93] Taru Tukiainen, Alexandra-Chloé Villani, Angela Yen, Manuel A Rivas, Jamie L Marshall, Rahul Satija, Matt Aguirre, Laura Gauthier, Mark Fleharty, Andrew Kirby, Beryl B Cummings, Stephane E Castel, Konrad J Karczewski, François Aguet, Andrea Byrnes, GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI,

Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Tuuli Lappalainen, Aviv Regev, Kristin G Ardlie, Nir Hacohen, and Daniel G MacArthur. Landscape of X chromosome inactivation across human tissues. *Nature*, 550(7675):244–248, October 2017.

[94] Kraig R Stevenson, Joseph D Coolon, and Patricia J Wittkopp. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome, 2013.

[95] Bryce van de Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*, 12(11):1061–1063, November 2015.

[96] Stephane E Castel, Alejandra Cervera, Pejman Mohammadi, François Aguet, Ferran Reverter, Aaron Wolman, Roderic Guigo, Ivan Iossifov, Ana Vasileva, and Tuuli Lappalainen. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.*, 50(9):1327–1334, September 2018.

[97] Po-Ru Loh, Giulio Genovese, and Steven A McCarroll. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature*, 584(7819):136–141, August 2020.

[98] Tuuli Lappalainen, Stephen B Montgomery, Alexandra C Nica, and Emmanouil T Dermitzakis. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am. J. Hum. Genet.*, 89(3):459–463, September 2011.

[99] William B Dobyns, Allison Filauro, Brett N Tomson, April S Chan, Allen W Ho, Nicholas T Ting, Jan C Oosterwijk, and Carole Ober. Inheritance of most x-linked traits is not dominant or recessive, just x-linked, 2004.