STATISTICAL METHODS IMPROVING THE CLINICAL UTILITY OF OMICS DATA

Bryce Rowland

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2022

Approved by:

Yun Li

Michael I. Love

Laura Raffield

Paola Giusti-Rodríguez

Di Wu

ABSTRACT

Bryce Rowland: Statistical Methods Improving the Clinical Utility of Omics Data
(Under the direction of Yun Li)

Variants identified via genome-wide association studies (GWAS) have ushered in an era of deep interest in omics data. Early adopters have used GWAS discoveries to inform drug targets and establish causal relationships using genetic instruments, yet more research must be done to bring the initial boons of GWAS to clinical practice. My dissertation presents three novel statistical methods which could bridge this gap by correcting biases when analyzing omics data and addressing methodological disparities affecting non-European populations. In my first project, I present THUNDER, a novel deconvolution method tailored to the unique challenges of chromatin conformation capture. Prior to our research, differential analysis of chromatin organization was confounded by underlying cell type proportions. Therefore, analyzing across individuals for differential chromatin activity has been of limited utility. THUNDER accurately estimates cell type proportions, allowing for their inclusion as a confounder in future association studies of Hi-C phenotypes. In my second project, I present GAUDI, a fused lasso approach to estimate polygenic risk scores (PRS) in admixed individuals. Our method addresses the decreases in performance of PRS methods in non-European populations, in part due to previously unaccounted for patterns of genetic admixture. Finally, in my third project, I extend polygenic risk score estimation techniques to the variable copy number setting to identify carriers for Spinal Muscular Atrophy (SMA) for which no standard test to identify these carriers exists.

ACKNOWLEDGEMENTS

Thank you God, for your grace and your easy yoke. Your loyal love is everlasting.

There are many wonderful people that deserve my utmost gratitude for their support. I could fill another dissertation with the stories of care summarized here.

To Monica, for your unwavering commitment to love me and support me. I am forever grateful for your willingness to listen when things were hard and celebrate our successes, usually at the Wooden Nickel, when things were easy. I love you.

To my mom, Rhonda, for your sacrifices to make this life possible. Thank you for instilling in me the value of an education and doing the hard things first.

To Joel, for cultivating my curiosity and for mentorship by stove tops and stoplights.

To my advisor, Yun, for cultivating my love for research, for teaching me when I knew so little, and always being willing to entertain my desire for big picture conversations.

To my friends in the Li lab, for your support.

To Chal (aka MDD), for being my day one at UNC, for keeping the bigger picture in front of me, and for the many chats lingering in the halls of Gillings.

To Amanda, for many crucial conversations when the finish line felt far away.

To Mason, Lola, and Harrison, I'm continually thankful for the journey of learning and friendship we're on together.

To my mentors at Centre College, Beau Weston, Rick Axtell, Jeffrey Heath, Joel Kilty, and John Kinkade. I hope that I've built well on the foundation that you laid for me in Danville.

To Summit Chapel Hill, for loving community and picking me up when I stumbled.

Regarding conflicts of interest, I disclose that I was employed as an employee of Illumina while conducting the research in Chapter 4.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACV	across-cell-type variation
AFR	African
AMR	Admixed American
CN	copy number
CTS	cell type specificity
CV	cross validated
DNA	deoxyribonucleic acid
EAS	East Asian
EUR	European
EWAS	epigenome-wide association study
FIRE	frequently interacting region
GWAS	genome-wide association study
LD	linkage disequilibrium
MAD	mean absolute deviation
MAF	minor allele frequency
NMF	nonnegative matrix factorizatin
PC	principal component
PCA	principal component analysis
PCR	Polymerase Chain Reaction
PPV	positive predictive value
PRS	polygenic risk score
RNA	ribonucleic acid
SAS	South Asian
scHi-C	single-cell Hi-C
SMA	spinal muscular atrophy
SNP	single nucleotide polymorphism

TAD	Topologically associating domain
THUNDER	Two Step Hi-C UNsupervised DEconvolution appRoach
WHI	Women's Health Initiative
WGS	whole genome sequencing
1kGP	1000 Genomes Project
3C	chromatin conformation capture

CHAPTER 1: LITERATURE REVIEW

1.1 THUNDER Literature Review

1.1.1 Introduction

Hi-C is an experimental technique to study chromatin organization. Chromatin organization regulates gene expression and facilitates DNA folding. Patterns of chromatin organization vary between different cell types (such as lymphocytes compared to monocytes), much like patterns of gene expression and DNA methylation. As a result, differential analysis of Hi-C phenotypes are biased by underlying cell type proportions. This bias can be addressed by including estimates of cell type proportions in regression analysis as a confounder. Thus, the accurate estimation of cell type proportions underlying the mixture of cell types in Hi-C data is a pressing statistical challenge. Deconvolution is a statistical task where cell type proportions and cell type informative features are inferred from a mixture of distinct cell types. Many deconvolution methods have been developed to analyze RNA-seq and DNA methylation data.[57, 24, 102, 101, 46, 91, 82, 47, 127, 83, 112, 64, 26] However, a relatively small number of methods have been developed to deconvolve Hi-C data. Current deconvolution approaches for omics data, including those previously developed to deconvolve Hi-C data, are of limited utility to estimate cell type proportions or generate cell-type specific features in the multi-sample bulk Hi-C datasets generated today. We present our new statistical method, THUNDER, which addresses this problem by accurately estimating cell type proportions in multi-sample bulk Hi-C datasets.

1.1.2 Hi-C Data Experimental Overview

The chromatin conformation capture (3C) method and the related technologies (broadly known as C-technologies) are experiments that have generated considerable understanding about the spatial organization of chromosomes within a cell.[23, 66, 93] The 3C method was originally developed to detect the contact frequency between two pre-specified genomic loci.[23] However, the proliferation of sequencing technologies combined with experimental techniques from 3C enabled the success of the genome-wide assay of chromatin conformation, Hi-C.[66] The result of the Hi-C experiment is a genome-wide catalog of chromatin fragments that were in close spatial proximity aggregated across all cell types within a tissue sample.[66] Hi-C experiments discovered several levels of chromatin organization in human cells including: A/B compartments[66], topologically associating domains (TADs)[25], frequently interacting regions (FIREs)[98, 22], chromatin loops[93], interchromosomal contacts, and/or intrachromosomal contacts.[122, 121] Structures discovered from Hi-C experiments support the theory that spatial chromatin organization plays a significant role in gene regulation since enhancers and the promoters of a target gene should be in close physical proximity to affect gene regulation.[63, 40] Disruption of chromatin organizational structures, especially TADs, are hypothesized to disrupt gene expression, and this disruption can lead to failure of downstream cell function.[71, 1] Insights from Hi-C data remain of considerable interest to study the spatial organization of chromatin in humans.

1.1.3 scHi-C Data Reveals cell-to-cell variability

Recent experimental advances known as single-cell Hi-C (scHi-C) have revealed that the spatial organization of chromosomes vary across different cell types. scHi-C refers to several different experimental procedures that maintain the chromatin conformation for each cell within a tissue sample.[92, 103, 104, 61] In contrast, bulk Hi-C data summarize the chromatin activity for all cell types in a tissue simultaneously. Initial scHi-C experiments in haploid cell lines detected previously known translocations specific to HAP1 cells that distinguished HAP1 and HeLa cell lines in PCA-based clustering.[92, 29] Further experiments in haploid cells revealed the cell-to-cell variability of

TADs and chromatin loops, but A/B compartments were conserved across cells.[103] scHi-C in diploid cells demonstrated that A/B compartments, TADs, and CCCTC-binding factor (CTCF) loop domains can be identified in single cells and are highly heterogeneous between cells.[104] Thus, scHi-C technologies revealed that cell type variability is a defining feature of spatial chromatin organization.

1.1.4 Defining the Hi-C deconvolution problem.

Cell-to-cell variability in spatial chromatin organization will confound differential analyses of Hi-C phenotypes, and further methodological development is needed to provide a way forward. Despite the interest in single-cell features, scHi-C experiments are expensive and have low yield, and thus Hi-C methods are still quite popular. As researchers continue to conduct Hi-C experiments, there will soon be sufficient individual-level data to conduct studies testing the association between genetic variation and spatial chromatin organization.[40] These future 3D-chromatin-interactome wide association studies (3WAS) and chromatin interactome QTL (iQTL) studies for Hi-C phenotypes will be confounded by underlying cell type proportions because of the evidence of cell-to-cell variability of spatial chromatin organization.[52, 39] Cell type proportion confounding in differential analyses is a well-defined problem in the analysis of RNA-seq and DNA methylation data. Within these fields, the standard approach to account for cell type proportion confounding is to estimate cell type proportions from the mixture data and to include the inferred proportions as a confounding variable in downstream association analyses.[52, 39, 102] To the best of our knowledge, there is no statistical method to infer cell type proportions across multiple bulk Hi-C samples simultaneously, which is capable of leveraging both intrachromosomal and interchromosomal contacts. Thus, the accurate estimation of cell type proportions from bulk Hi-C data is a pressing statistical challenge.

The statistical task of estimating underlying cell type proportions from omics data is known as deconvolution. Many methods to perform deconvolution have been developed to estimate cell type proportions in RNA-seq and DNA methylation mixture data.[57, 24, 102, 101, 46, 91, 82, 47, 127, 83, 112, 64, 26] We define mixture data as omics data aggregated at the sample level that is

composed of several distinct cell types or cell states (hereafter, we refer to both as cell types for brevity). In addition to the accurate estimation of cell type proportions, deconvolution methods can estimate activity at the cell-type-specific level from mixture data. Since mixture data are aggregated at the sample level, cell-type specific features are obscured at the initial data resolution. However, these cell-type specific features are of great biological interest, promoting fields of study to design cell-type specific experimental protocols. These cell-type specific or single-cell datasets are not available in large sample sizes for many cell types. Thus, deconvolution methods provide a valuable tool to estimate both cell type proportions and cell-type specific features when experimental data are rare.

1.1.5 Statistical Deconvolution Methods for Omics Data

Statistical methods to deconvolve omics data can be divided into two categories based on their required input data: reference-free and reference-based methods. As a generality, the input data common to both categories is a matrix of n subjects and p features. Reference-based deconvolution methods additionally require a set of cell-type specific profiles as input. These cell-type specific profiles prime reference-dependent methods to identify distinguishing features across cell types in order to more effectively deconvolve the mixture samples. Alternatively, reference-free methods do not require additional data sources to perform deconvolution, only the nxp matrix of aggregated data. Reference-based methods have outperformed reference-free methods in comparative studies, but reference-free methods are more robust to the practical challenges of deconvolving omics data.

The following reference-based methods demonstrate the diversity of statistical methods to deconvolve omics data, with particular attention to methods applied in Chapter 2. Shen-Orr et al. developed csSAM, a standard least-squares based method to deconvolve microarray gene expression data.[102] Newman et al. developed the popular method CIBERSORT in 2015 and extended the method to CIBERSORTx in 2019.[82, 83] Both CIBERSORT and CIBERSORTx are reference-based approaches leveraging singular-value decomposition to estimate cell type proportions, and are tailored to unique characteristics of gene expression data, especially sample contamination

for tumor data. MuSiC is another reference-based deconvolution method which uses weighted non-negative least squares regression to estimate cell type proportions from bulk RNAseq data based on multi-subject single cell RNAseq data.[112] MuSiC leverages features which demonstrate cross-cell and cross-sample consistency to apply cell-type-specific feature information in estimating cell type proportions. MuSiC additionally applies a tree-based procedure to address collinearity in closely related cell types within a bulk tissue. The profiled reference-based methods all estimate cell type proportions with high accuracy when all cell types in the mixture data are contained in the reference.[82, 126] However, small perturbations in the reference panel can result in biased estimates.[112, 64] Thus, reference-based methods are preferred to reference-free methods when high quality reference panels are available.

When high quality reference panels are unavailable for all cell types in RNA-seq and DNA methylation mixture data, reference-free deconvolution methods are used to estimate cell type proportions. Reference-free methods infer latent variables which summarize the variability across the mixture data samples, assuming the number of cell types in the sample are known. The inferred cluster specific features can correspond to features from either cell types. Non-negative Matrix Factorization (NMF) and its extensions are commonly used for reference-free deconvolution, particularly in deconvolution of RNA-seq data.[59, 60, 12, 57, 32, 64] Due to the centrality of NMF to our work, we profile NMF extensively in the following section. In the DNA-methylation literature, there are methods that either explicitly infer cell type proportions or adjust for the underlying differences when performing EWAS. TOAST is a recently proposed unsupervised deconvolution and feature selection algorithm which iteratively searches for cell type-specific features and estimates composition.[64] TOAST was developed for the flexible application to both gene expression and DNA methylation data. BayesCCM is another reference-free deconvolution method that uses prior knowledge about the distribution of cell counts within a tissue to infer cell counts from DNA methylation data.[91] Houseman et al. 2014 does not explicitly apply a deconvolution method, but rather adjusts the estimates of an EWAS using an SVD model to correct the unadjusted effect sizes in confounded EWAS analysis.[46] These methods all estimate cell type

proportions accurately in their designed context, however little consensus remains as to the best approach for reference-free deconvolution across omics data.

For our method to deconvolve Hi-C data, we adopted a reference-free approach due to the lack of cell-type specific reference panels for Hi-C data and a lack of understanding of the differences between the spatial interactome between healthy and diseased individuals, which may bias reference-based methods. As stated above, scHi-C or cell-type specific are rare for cells in understudied tissues and even for some cells in well-studied tissues. Additionally, cell-type specific Hi-C datasets are usually collected in a small number of healthy individuals. Samples from healthy individuals may not be applicable to deconvolve samples from patients of different age, sex, or other relevant phenotypes. For example, reference-based methods performed poorly to deconvolve DNA methylation data when reference samples are from adults and the mixture samples are from newborns.[124] Assuming this principle holds for Hi-C data, reference-free methods are more relevant to the problem at hand of deconvolving Hi-C data, where such reference panels are still rare.

1.1.6 Non-Negative Matrix Factorization (NMF)

Non-negative matrix factorization is the statistical foundation behind our proposed method to deconvolve Hi-C data, THUNDER. Following Lee and Seung, 1999, we discuss the properties of Non-negative Matrix Factorization (NMF) as a matrix decomposition method well-suited to learning about a data object as the sum of distinct parts.[59] The problem of deconvolving complex statistical data is analogous to decomposing a matrix into a product of two matrices. We will show that deconvolving complex statistical data with NMF leads to a natural interpretation of the basis matrix as a matrix of cell-type specific features and of the coefficient matrix as estimates of cell type proportions.

Consider the problem of deconvolving data represented in a $p \times n$ matrix, V, with p features and n mixture samples. We let $k > 0$ be an integer specified for the number of distinct cell types in the mixture sample and is chosen a priori. A matrix decomposition approach to the deconvolution

problem seeks to find an approximation for V such that $V \approx WH$. W , a $p \times k$ (or feature by cell type) matrix, is often referred to as a basis matrix, and H, a $k \times n$ (or cell type by sample) matrix, is known as the coefficient matrix.

NMF applies a non-negativity constraint to the elements of a basis matrix, which makes it a factorization method well suited to deconvolution. Factorization methods can be characterized by mathematical constraints on the elements of the basis and coefficient matrices. Some factorization methods, such as PCA, allow elements of the basis matrix or coefficient matrix to be negative. However, negative matrix elements are not biologically interpretable. In the NMF decomposition, only additive combinations of basis vectors can be used to reconstitute the original matrix V. Therefore, the NMF non-negativity constraints correspond to the intuition that the sum of the parts make a whole. This non-negativity constraint makes NMF estimates particularly useful for deconvolving genetic data. The elements of the coefficient matrix can be scaled to estimate cell type proportions and the elements of the basis matrix may correspond to cell-type specific features. NMF is a useful tool for reference-free deconvolution, and it provides the mathematical foundation for THUNDER.

1.1.7 Review of Hi-C Deconvolution Methods

There exist two particular challenges of applying existing methods developed in other omics data to deconvolve Hi-C data: the lack of cell-type-specific Hi-C reference profiles and the lack of an ubiquitous aggregating unit for summarizing Hi-C data. First, existing reference-based methods, such as CIBERSORT or MuSiC, can not be directly applied to Hi-C data. For most cell types, there is a paucity of single-cell or cell-type specific Hi-C datasets to serve as reference panels for reference-based methods. While several cell types have been profiled with scHi-C technology as described above, these resources are not widely available for all cell types, or even all major cell types in a given tissue. Second, Hi-C data can be summarized at several different structural levels, as discussed above, and it is unclear which level(s) of measurement are most scientifically relevant or effective for deconvolution purposes. In contrast, when deconvolving gene expression

data, it is clear that the aggregating unit of interest is the gene. As a result, the feature space for Hi-C deconvolution is theoretically unbounded, or at least several orders of magnitude larger than applications to deconvolve gene expression data. Without a feature selection strategy tuned to the context of Hi-C data, deconvolution methods will not perform well in Hi-C data. These two challenges make previously developed deconvolution methods unsuitable to deconvolve Hi-C data in practice.

Several methods have attempted to infer cell-type proportions from bulk Hi-C data as either a primary or secondary aim of the method. 3CDE is a matrix-based deconvolution approach for bulk Hi-C data, which infers non-overlapping domains of chromatin activity in each cell type and uses a linear combination of binary interaction information at these domains to deconvolve the contact frequency matrix.[99] Junier et al. put forth a method to infer overlapping domains of chromatin activity as well as their mixture proportions.[53] To the best of our knowledge, no software accompanies the work by Junier et al.[53] Carstens et al. infer chromatin structure ensembles from bulk Hi-C contact information using a Bayesian approach but does not infer cell type proportions directly.[15] None of these methods analyze both interchromosomal and intrachromosomal contacts simultaneously, which we demonstrate in Chapter 2 are sometimes essential and always helpful to deconvolve Hi-C data. Additionally, 3CDE does not analyze multiple samples of Hi-C data simultaneously, limiting practical utility to match latent clusters across multiple samples (see Figure A.1). Due to these limitations, the problem of Hi-C deconvolution is an open question.

1.1.8 Conclusion

In Chapter Two, we present the details of our method, THUNDER, a reference-free deconvolution method to estimate cell type proportions from bulk Hi-C data. THUNDER extends NMF to estimate cell type proportions in multiple samples simultaneously and incorporate data from interchromosomal contacts. Our method also generates biologically informative estimates of cell-type specific chromatin interaction. We hope that THUNDER will become a useful tool to estimate cell type proportions as a confounding factor underlying future differential organization studies.

1.2 GAUDI Literature Review

1.2.1 Introduction

Polygenic risk scores (PRS) have been successfully incorporated into clinical risk models to perform therapeutic interventions and disease screening. [78, 81, 56, 48, 105] However, the benefits of PRS in personalized medicine are disproportionately concentrated in European populations.[76] European ancestry populations are over-represented in genetic studies used to create PRSs.[43] But, PRS trained in European populations are not transferable to non-European populations, and they transfer particularly poorly to African ancestry populations.[75, 76] Additionally, genomes from recently admixed populations, including African Americans and Hispanic/Latino individuals, often have non-negligible and varying levels of genetic admixture. Genetic admixture further complicates PRS transferability, even if more data was available in non-European populations. Several recent methods have been proposed to estimate PRS in admixed populations, but they do not adequately account for differential allele frequencies across populations or utilize ancestry information in admixed genomes in modeling.[73, 10] Our method, GAUDI, addresses methodological disparities in estimating PRSs in admixed individuals by modeling local ancestry and borrowing information across ancestral segments to estimate ancestry-shared effects. GAUDI is a penalized regression based PRS method which combines fusion and sparsity parameters to estimate population-shared and population-specific effects. We anticipate that GAUDI will help increase equity in the benefits of PRS research.

1.2.2 Polygenic Risk Score Overview

In its simplest form, a PRS is the sum of the number of risk alleles carried by an individual weighted by their effect sizes from GWASs. PRSs are an attempt to summarize the proportion of

variability of a trait explained by an individuals genetics. Mathematically, for an individual i, a PRS built using p SNPs is defined as,

$$PRS_i = \sum_{j=1}^{p} x_{ij}\hat{\beta}_j \tag{1.1}$$

where x_{ij} is the number of risk alleles carried by individual i at SNP j, and β_j is the GWAS effect size for SNP j on the phenotype of interest. The problem of estimating an individuals PRS consists of two primary tasks: selecting relevant genetic variants for inclusion and weighting these variants appropriately. The power and predictive accuracy of PRSs is a function of the sample sizes in training and testing samples, explained genetic variance, and method for selecting and weighting variants for the score.[27]

Initially, a proposed solution to the variable selection problem was to include only genome-wide significant variants from a large GWAS, but the constructed PRSs were found to be sub-optimal.[51] The pruning and thresholding method, commonly referred to as the P+T method, was proposed to identify the optimal p-value cutoff using cross-validation and allowed for sub-genome-wide significant variants to be included in the PRS.[51] Sub-genome-wide significant variants may be useful for prediction if they have small but non-zero true effect sizes on the phenotype.[51] Further, variants included in the P+T PRS are commonly subject to LD pruning, such that variants in high-LD are considered to duplicate information contained in a more significant variant.[30] Another PRS method, LDPred, was the first PRS method to utilize external LD reference panels to re-estimate variant effect sizes rather than lift them directly from large GWASs.[111] By not excluding variants via LD pruning but re-estimating their effect sizes, LDPred often outperforms the P+T method.[111, 68] Subsequently, the use of external LD reference panels in PRS estimation has become quite common.[68, 123] One such method relying on LD reference panels, Deterministic Bayesian Sparse Linear Mixed Model (DBSLMM), utilizes a Bayesian framework to estimate a PRS under a range of possible genetic architectures. DBSLMM demonstrates advantages over traditional PRS methods by assuming a proportion of SNPs have large effect sizes, and all other SNPs have infinitesimal effect sizes. Their model innovates on previous methods by incorporating a range

of possible genetic architectures a priori and allows the data to drive assumptions about genetic architecture.[123] However, despite the increasing methodological complexity in PRS-research, including the recent attempts to incorporate deep-learning methods[120] it is still most common to estimate PRS using the P+T method.

1.2.3 Clinical Utility of Polygenic Risk Scores

Increased access to genotype data for large numbers of individuals via biobank studies have led to a renewed interest in PRSs. PRSs have improved clinical risk prediction models in several heritable phenotypes. Risk prediction models are clinically useful when the population can be stratified into risk groups that substantially affect the risk-benefit balance of a public health intervention.[16] Generally, the higher the risk of a public health intervention, the higher the perceived clinical benefit must be to recommend the intervention. The clinical utility of PRSs have been demonstrated primarily in PRS-informed therapeutic interventions[78, 81, 56]and PRS-informed disease screening.[48, 109, 105] For example, several studies have demonstrated that a PRS for coronary artery disease, along with other environmental predictors, stratified individuals into clinically meaningful risk categories. Two studies have demonstrated that the use of statins to treat coronary heart disease in individuals with the highest PRS resulted in a 45% relative risk reduction of the 10 year risk of a heart attack or coronary artery disease related death.[56, 78] PRS have also demonstrated success when applied to PRS-informed disease screening for prostate cancer. Prostate-specific antigen (PSA) screening is not recommended by the US Preventive Services Task force since benefits are outweighed by false positives which result in overtreatment.[8] A prostate cancer PRS helped identify men at significantly elevated risk of disease, and the score may be used to counsel actions following a positive PSA test.[85, 100] With biobank study resources, future applications or PRS are well on their way to clinical applications.

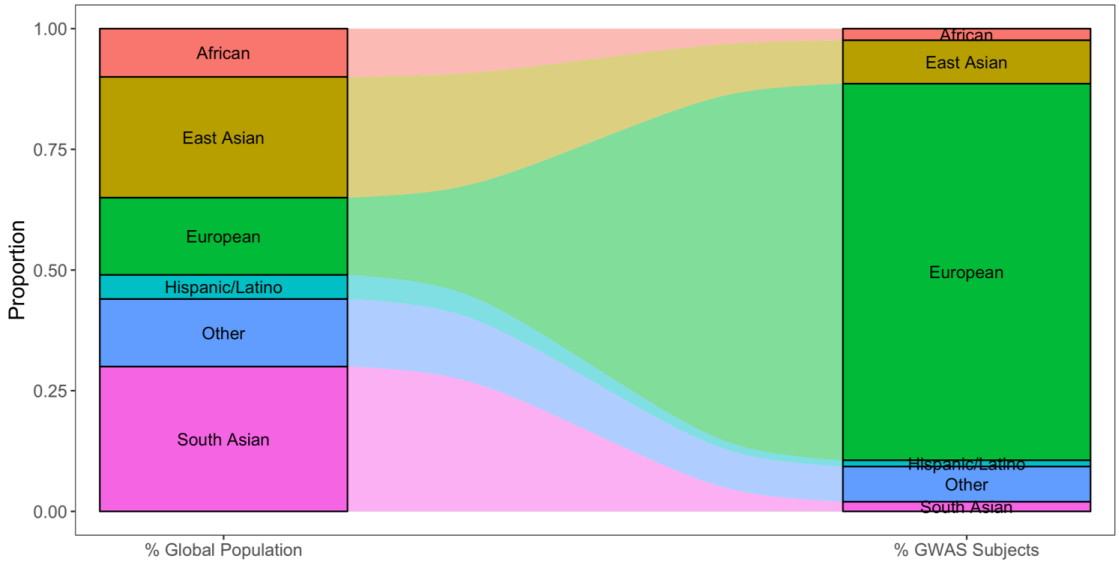

Figure 1.1: Disparities in GWAS Research. I modified Figure 1 from Gurdasani et al. 2019 and Figure 1 from Martin *et al.* 2019 because I found both figures powerful individually, but more powerful when synchronized with one another. Europeans are over-represented in genetic studies compared to their proportion of the global population.

1.2.4 PRS have limited utility in global populations

However, the benefits of PRS research have not been equitably distributed across populations.[76] In the majority of studies to date, PRSs predict individual risk more accurately in European populations than in non-European populations.[75, 76, 73, 10] This unfortunate reality is not surprising, since there has been a global underinvestment in genetic association studies in non-European populations (Figure 1.1).[43, 76] Additionally, several theories in population genetics suggest that prediction accuracy will decrease between populations as genetic divergence increases. This decrease in prediction accuracy is due in part to differences in the site frequency spectrum, linkage disequilibrium, and environmental factors across populations.[76, 110, 10, 113]. These challenges have resulted in calls to action to increase genetic diversity in GWASs, as well as the need for methodological innovation to increase the transferability of PRSs across populations.[76, 20, 43] Despite their methodological shortcomings, PRS are proposed for inclusion in clinical risk prediction as described above, which may exacerbate health disparities.[76] Statistical methods to increase

the transferability of PRS across populations can be categorized into two strategies: identifying causal variants with epigenetic annotations and leveraging results from multi-population association studies. The first strategy consists of methods which identify variants more likely to be causal across global populations using external epigenetic datasets.[3, 65, 115] These methods hypothesize that PRS containing the true causal variants will be more transferable across populations. This further assumes that the causal variants are the same across populations, and their effects are identical. Examples of variants demonstrating ancestry-specific effects due to being nearly fixed in Europeans include rs2814778, the Duffy null variant residing in Atypical Chemokine Receptor 1 (*ACKR1*) explaining 7% variation in white blood cell counts among African Americans[95, 94], and the trypanolytic *APOL1* G1/G2 alleles conferring an estimated 20% lifetime risk of developing chronic kidney disease.[67, 36]. Several recent examples of differential effect sizes across populations have been reported and are potentially caused by gene-gene or gene-environment interactions which current GWAS are under-powered to study in depth.[11, 86] The second strategy is to utilize multi-population studies to better estimate GWAS weights or account for the lack of transferability across populations.[17, 14] However, European populations are often still the largest populations included in multi-population meta-analyses, so the merits of these methods can not fully be assessed until GWAS studies are more representative of global populations. Until the disparity in PRS prediction applicability is addressed, a disparity in health outcomes between European and non-European populations is inevitable.

1.2.5 PRS in Admixed Individuals

The problem of PRS transferability is further complicated when applying PRSs to individuals with recent genetic admixture due to the unique mosaic structure of individuals of admixed ancestry.[75, 73, 10] A useful theoretical model to understand the genomes of admixed individuals is that of a mosaic of genomic regions from ancestral populations. Due to crossover events, recently admixed genomes are composed of large chunks inherited from distinct ancestral populations. In part due to these alternating ancestral mosaics, PRSs trained in ancestral populations are not directly

applicable to individuals of recent admixture.[73, 10] However, a robust assessment of this question is challenging because of small sample sizes in non-European populations.[75, 73] Furthermore, the definition of an LD-reference panel in admixed individuals is an open question, and the utility of an in-sample LD calculation for admixed individuals has shortcomings when allele frequencies are highly differentiated across populations.[87] As a result, the above-mentioned PRS methods relying on external LD reference panels for PRS estimation will have limited usability in admixed individuals. Due to the unique structure of admixed genomes as well as methodological limitations, the transferability of PRS trained in ancestral populations to admixed populations remains an open question.

Recently, several methods have been proposed to estimate PRSs in admixed individuals by accounting for the mosaic structure of admixed genomes. Accounting for this structure in modeling assumes the accurate estimation of either global or local ancestry. Global ancestry is the total proportion of each ancestral population within an individual genome. Local ancestry is a finer measurement where ancestry information is available for each base-pair in the human genome. Since genotyping technology is agnostic to ancestry, the task of estimating local and global ancestry from genotype data is known as local ancestry inference. Ancestry inference methods always require genotype reference panels from the ancestral populations.[119] Ancestry inference, therefore, is the first step in applying these innovative PRS methods for admixed individuals.

Modern statistical and machine learning tools perform highly accurate estimation of local ancestry in admixed samples, even when reference panels in ancestral populations have small sample size or are unphased. Early methods to infer local ancestry relied on large reference panels of phased genotypes from ancestral populations. However, methods such as RFMix and ELAI have relaxed those assumptions, allowing local ancestry inference in understudied populations where such reference panels may not exist.[72, 42] Local ancestry inference methods primarily perform estimation via hidden Markov models (HMMs).[37, 119] These methods can be broadly classified into two groups: those which account for background LD and admixture explicitly (e.g. ELAI), and those which remove variants in LD (e.g. RFMix).[37] Both ELAI and RFMix have demonstrated

accurate estimation of local ancestry in simulations and real-data analysis. To summarize, given the right reference panels, ancestry inference methods have been highly successful at recovering global and local ancestry from genotype data.

Assuming that local ancestry inference can be performed with high accuracy in admixed individuals, we will now profile several recent attempts to incorporate estimates of genetic ancestry in PRSs. The approach taken by Marnetto et al. 2020 infers local ancestry for all individuals in the PRS training data, and it applies effect sizes from ancestry specific GWASs piecewise across the genome, an approach they name the combined ancestry-specific polygenic risk score (casPRS).[73] Consider the problem of fitting a PRS for a recently admixed individual with genomic segments from populations A and B. Marnetto et al. defines a partial PRS for some subset of variants across the genome such that $k < p$,

$$pPRS_i = \sum_{j=1}^{k} x_{ij}\beta_j$$

An ancestry specific partial PRS (aspPRS) is defined as a proxy for the total PRS that uses only the genomic portion pertaining to ancestry A, which in practice is inferred using local ancestry inference methods (specifically ELAI in their work[42]). Applying the same aspPRS methodology for regions assigned to population B via local ancestry inference, one can construct the casPRS by adding the aspPRS weighted by the global ancestry proportions for each individual in the training sample. Using real-data based simulations in several admixed populations, they demonstrate that total PRS without local ancestry information is the weighted average of aspPRS. Additionally, in an analysis of UK Biobank and Biobank Japan individuals, aspPRS never outperform the total PRS, and casPRS has equivalent performance to the total PRS in most settings.[73]

Bitarello and Mathieson provide three alternatives to the approach taken by Marnetto et al.[10] In their simplest approach, they weight two population specific PRSs for populations A and B by a constant, α, ranging between 0-1. For an individual i,

$$PRS^1_{C,i} = \alpha PRS_{A,i} + (1-\alpha)PRS_{B,i} \tag{1.2}$$

This approach was also first proposed in a separate study by Marquez-Luna et al.[74] Second, in addition to weighting by α, they weight by the global ancestry proportion for each individual for population B.

$$PRS^2_{C,i} = \alpha(1 - p_{B,i})PRS_{A,i} + (1 - \alpha - \alpha p_{B,i})PRS_{B,i} \tag{1.3}$$

Finally, utilizing local ancestry inferred regions, they construct a PRS where weighted averages of GWAS effect sizes from populations$A$ and B are utilized for regions corresponding to population A, and effect sizes from population B are used for SNPs in population B inferred regions,

$$PRS^3_{C,i} = \alpha[\sum_{j \in A} \beta_{i,A} x_{ij}] + (1 - \alpha)[\sum_{j \in B} \beta_{i,B} x_{ij}] + \sum_{j \in B} \beta_{i,B} x_{ij}$$

Conceptually, this is nearly identical to the model proposed by Marnetto et al. but using a weighted average of GWAS effects for one population rather than the aspPRS.

The proposed PRS methods for admixed individuals have several limitations. Each of the three methods here rely on utilizing summary statistics from previously conducted GWAS in the ancestral populations for the admixed individual. Largely, use of GWAS results from ancestral populations assumes independent genetic architectures between population groups, despite evidence that the majority of variants have concordant effect sizes across populations.[110] Additionally, as mentioned above, GWAS studies are biased toward discovering variants that are common in the population of the studied cohort.[76] Therefore, even after incorporating local ancestry estimates into PRS estimation, the above methods are biased toward SNPs which are common in the ancestral populations. Additionally, if rare variants tend to have larger GWAS effect sizes, and these effect sizes are correlated across populations, these scores will underestimate the effect size of variants which are common in one population but are rare in another population. Thus, a PRS method which jointly models ancestry shared and ancestry specific effects using individual admixed genomes may improve PRS estimation in admixed individuals.

1.2.6 Brief Overview of Penalized Regression

Our approach to estimate PRS in admixed individuals models both shared and specific genetic effects across ancestry groups via a novel application of penalized regression. Penalized regression methods are a popular set of statistical approaches where optimizing a regression likelihood is modified to include some penalty to influence the estimation of the model parameters. These techniques are often presented as data-driven methods, which replace performing challenging statistical tasks, such as variable selection, with a slightly less challenging task of estimating tuning parameters. Ubiquitous methods categorized as penalized regression methods include lasso regression[106], ridge regression[45], and elastic net[49]. In this work, I will detail two relevant penalized regression methods to our PRS application: the fused lasso[107] and the generalized lasso[108] (where the fused lasso is a special but illustrative case of the generalized lasso).

In the original paper proposing the fused lasso, Tibsherani presents the method with both sparsity and fusion penalties as fused lasso, but for the purposes of our exposition, we will separate the two penalties to build clarity within the generalized lasso framework. Consider the prediction problem with N observations with outcome $y_1, ..., y_n$ and features x_{ij}, $i = 1, ..., N$, $j = 1, .., p$. We assume that the x_{ij} are realizations of features X_j and the features can be ordered as $X_1, ..., X_p$ in some meaningful way. We are then interested in predicting Y using $X_1, ..., X_p$, and especially in the case that $p >> N$.

First, consider the regression model,

$$y_i = \sum_j x_{ij}\beta_j + \epsilon_i$$

with the errors having mean zero and constant variance. The fused lasso solution is defined as,

$$\hat{\beta} = argmin\{\sum_i (y_i - x_{ij}\beta_j)^2\} \quad \text{subject to} \quad \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leqslant s$$

In order to account for the ordering in the features, the fused lasso penalty encourages coefficients which are ordered adjacent to one another to be similar. The Lagrangian formulation of this problem is,

$$\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$$

The fused lasso is a special case of the generalized lasso, which formulates the penalized regression problem as,

$$\text{minimize}_{\beta \in R^p} \quad \frac{1}{2}||y - X\beta||_2^2 + \lambda||D\beta||_1$$

Where D is chosen by the statistician to the sparsity of $D\beta$ corresponds to some desired behavior of β. The fused lasso is recovered by specifying D to be,

$$D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

And sparsity can be introduced to any penalty by appending the rows of D with the identity matrix as such,

$$D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

1.2.7 Conclusion

In Chapter Three, I present GAUDI, a PRS estimation method for admixed individuals that models local ancestry and estimates ancestry-shared effects by borrowing information across ancestral segments in admixed genomes using penalized regression. Unlike previous methods, GAUDI does not rely on the use of external GWAS results from ancestral populations. By doing so, we bring our PRS modeling in line with the belief that genetic architectures are not independent across ancestry groups. Additionally, GWAS results in ancestral populations could be quite rare depending on the populations comprising an admixed population. GAUDI can also model the PRS with high accuracy in the presence of ancestry-specific effects by balancing fusion and sparsity penalties. Additionally, by estimating a PRS from individual admixed genomes, GAUDI more efficiently utilizes information across ancestral mosaic segments to accurately estimate effect sizes of rare variants.

1.3 SMA Silent Carrier Screening Literature Review

Spinal muscular atrophy is an autosomal recessive neuromuscular disorder characterized by loss of alpha motor neurons and causes muscle atrophy shortly after birth.[69, 79] After cystic fibrosis, it is the leading genetic cause of infant death.[88] The disease causing gene, *SMN1*, and its

paralog, *SMN2*, reside in a 2Mb region on chromosome 5 characterized by a complex pattern of gene duplication and inversions, which make characterizing the region through traditional whole-genome sequencing techniques (WGS) challenging. Additionally, the *SMN1/2* region is characterized by variable copy number, with differential copy-number frequencies across global populations.[44]. Finally, *SMN1* and *SMN2* share >99.9 sequence similarity.[19] One of the base-pair differences is NM_000344.3:c.840C>T (c.840T), the biallelic absence of which causes 95% of SMA cases. Population-wide carrier screening is recommended by the American College of Medical Genetics and Genomics.[89] Previous screening approaches determine the copy number of *SMN1* based on the differences between *SMN1* and *SMN2* at c.840C>T. Recently, a WGS-based approach to diagnose SMA and identify carriers demonstrated comparable precision and recall to previous PCR-based screening methods.[19] However, one current limitation n of this approach is that individuals with a *SMN1* copy number (CN) genotype of 2+0 are currently not correctly identified as carriers. Due to the differential copy number frequencies across global populations, any screening tool must demonstrate accurate performance across diverse samples. For example, 0.4% of African Americans are SMA silent carriers (2+0 genotype) compared to 0.1% of Europeans and 0.07% of Hispanic individuals. Using novel bioinformatics and biostatistical tools, we propose extending existing PRS-like prediction methods in this variable copy number setting. We will show that using multi-population analyses result in better population-screening results than population-specific methods, addressing the concerns of variable *SMN1* allele frequency across populations.

CHAPTER 2: THUNDER: A REFERENCE-FREE DECONVOLUTION METHOD TO INFER CELL TYPE PROPORTIONS FROM BULK HI-C DATA

2.1 Methods

2.1.1 THUNDER Overview

In order to estimate the underlying cell type proportions found in bulk Hi-C datasets, we propose a Two Step Hi-C UNsupervised DEconvolution appRoach (THUNDER).[96] THUNDER consists of a feature selection step and a deconvolution step, both of which rely on non-negative matrix factorization, or NMF (Figure 2.1). For Hi-C data, V denotes the $p \times n$ mixture matrix of bulk Hi-C samples with p bin-pairs and n columns of mixture samples. We let $k > 0$ be an integer specified for the number of distinct cell types in the mixture sample and is chosen a priori. NMF seeks to find an approximation $V \approx WH$, where $W$and H are $p \times k$ and $k \times n$ non-negative matrices. We refer to $W$and H as the cell type profile and proportion matrices, respectively. The NMF problem can be solved by finding a local minimum for the Euclidean norm between V and WH, $||V - WH||^2$, under the constraint that W and H are non-negative. We use the NMF R package[35] with the updates provided by Lee and Seung [60] with random initialization of the W and H matrices.

In step one of THUNDER, we perform an initial NMF deconvolution estimate on the $p \times n$ matrix V to obtain the deconvolution estimate $V \approx W_1 H_1$ where W_1 is a $p \times k$ matrix and H_1 is a $k \times n$ matrix. We then perform feature selection using the decomposition to identify informative bin-pairs across cell types. THUNDER performs feature selection on intrachromosomal and interchromosomal contacts separately. Let $W_1(i, j)$ denote the element in the i^{th} row and j^{th} column of the cell-type specific profile matrix W_1. Let S_{intra} and S_{inter} denote the set of intrachromosomal and interchromosomal bin-pairs respectively.

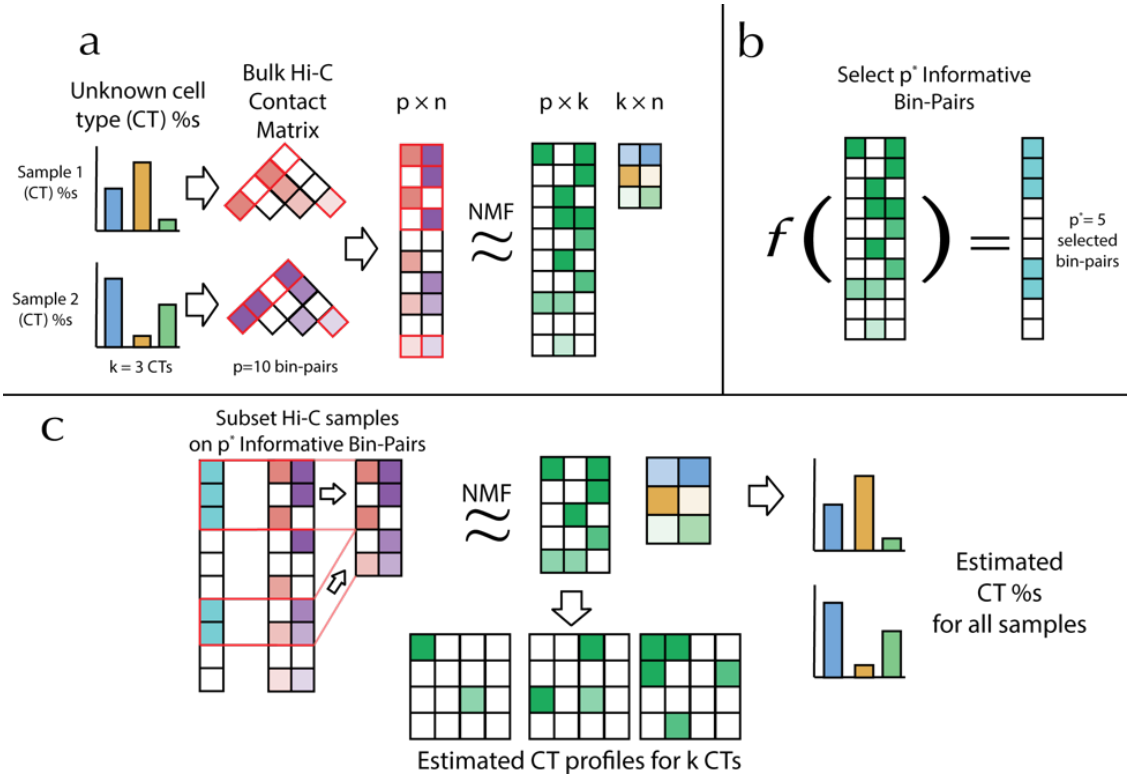

Figure 2.1: Overview of THUNDER Procedure. (a) Overview of nonnegative matrix factorization (NMF) in the context of bulk Hi-C data. Three underlying cell types each contribute to the observed contact frequencies in two bulk Hi-C samples. The first step of the THUNDER algorithm is to deconvolve the input bulk Hi-C data into two estimated matrices: the cell type profile matrix and the proportion matrix. (b) In order to select informative bin-pairs for deconvolution, THUNDER utilizes a feature selection algorithm specifically tailored to Hi-C data to analyze the contact frequency distribution of the bin-pairs in the cell type profile matrix. (c) In the final step of THUNDER, we subset the bin-pairs contained in the input bulk Hi-C samples to only informative bin-pairs and perform NMF a second time. The proportion matrix is scaled to be estimates of the underlying cell type proportions in the bulk Hi-C samples. The cell type profile matrix estimates cell-type specific contact distributions.

Standard deviation across cell types for bin-pair i is defined as,

$$SD_i = \frac{1}{k-1}\sum_{j=1}^{k}(W_1(i,j) - \frac{1}{k}W_1(i,\cdot))^2$$

Feature score across cell types for bin pair i is defined as follows,

$$FS_i = 1 + \frac{1}{log_2(k)}\sum_{j=1}^{k} p(i,j)log_2(p(i,j))$$

where $p(i, \Omega)$ is the probability that the i^{th} pairwise bin contributes to cell type Ω, i.e.,

$$p(i, \Omega) = \frac{W_1(i, \Omega)}{\sum_{j=1}^{k} W_1(i, j)}$$

Feature scores range from $[0, 1]$ with higher scores representing bin-pairs with higher cell-type specificity. We further define,

$$\hat{\mu}_{SD,inter} = \frac{1}{|S_{inter}|} \sum_{i:i \in S_{inter}} SD_i$$

$$\hat{\sigma}_{SD,inter} = \frac{1}{|S_{inter}| - 1} \sum_{i:i \in S_{inter}} (SD_i - \hat{\mu}_{SD,inter})^2$$

$$\hat{m}_{FS,intra} = \text{median}_{\{i \in S_{intra}\}}(FS_i)$$

$$\hat{s}_{FS,intra} = \text{median}_{\{i \in S_{intra}\}}(\hat{m}_{FS,intra} - FS_i)$$

THUNDER's feature selection algorithm is as follows. Intrachromosomal bin-pair i is defined to be an informative bin-pair if $FS_i > \hat{m}_{FS,intra} + 3\hat{s}_{FS,intra}$, and interchromosomal bin pair j is defined to be an informative bin pair if $SD_j > \hat{\mu}_{SD,inter} + 3\hat{\sigma}_{SD,inter}$

Let p^* be the number of informative bin-pairs identified via feature selection. We subset Von all informative bin-pairs to form the reduced $p^* \times n$ mixture matrix V^*. We then perform NMF on V^* to arrive at our final estimates, W^* (of dimension $p^* \times k$) and H^* (of dimension $k \times n$). Finally, we adjust the columns of H^* to sum to one to represent cell type proportions. The scaled elements of H^* are cell type proportion estimates in the p mixture samples. The columns of W^* are parsimonious cell-type specific contact profiles. These parsimonious contact profiles estimate Hi-C contact frequencies at the bin-pairs which most differentiate the inferred cell types in the Hi-C samples.

2.1.2 Simulating Bulk Hi-C Data

2.1.2.1 Ramani *et al.* Data

Cellular indices were downloaded from GSE84920 which included 6 libraries: ML1, ML2, ML3, ML4, PL1 and PL2.[92] For our simulations, we used data from all libraries except ML4. These libraries are composed of scHi-C data from five distinct human and mouse cell lines. Within each cell, we followed the same preprocessing procedure as outlined in Ramani *et al.* Specifically, cellular indices with fewer than 1000 unique reads, a cis:trans ratio less than 1, and cells with less than 95% of reads aligning uniquely to either the mouse or human genomes were filtered out before analysis. Additionally, we removed reads whose genomic distance was $<$15Kb due to self-ligation, and only considered unique reads. For the four libraries containing HAP1 and HeLa cells (ML1, ML2, PL1 and PL2), we discarded cellular indices where the proportion of sites where the non-reference allele was found was between 57% and 99%.

To account for varying levels of single-cell sequencing depth across libraries, we considered only cells with filtered reads greater than the 20^{th} quantile and less than the 90^{th} quantile of reads and across all libraries and cell types considered in the simulated mixture sample. We then downsampled each cell via multinomial sampling to the number of contacts in the cell with the fewest number of contacts across all cell types considered in the sample. We constructed contact matrices on the filtered and downsampled scHi-C data at three levels of data representation at 10Mb bin-pair resolution: interchromosomal contacts only, intrachromosomal contacts only, and both interchromosomal contacts and intrachromosomal contacts together. The total number of cells in each mixture sample is equal to the smallest number of cells present in a cell line after the filtering step across cells in the mixture sample. We normalized the mixture samples by dividing the observed contacts by the total number of contacts to generate normalized contact frequencies for each sample.

To test proposed feature selection methods for THUNDER, we generated three cell type mixtures of GM12878, HAP1, and HeLa cells. We generated 5 replications of 12 bulk samples (3

pure samples and 9 mixture samples) which are mixtures of the three cell lines at the proportions given in Table A.2. These proportions are a subset of those used by Shen-Orr and Tibsherani in their simulated mixture data.[102]

2.1.2.2 Lee *et al.* Dataset.

4,238 scHi-C profiles from the prefrontal cortex region of two postmortem adult human brains were downloaded from GSE130711. Non-neuronal cell types were previously identified via clustering based on CG methylation signature, followed by fine clustering of neuronal subtypes using non-CG methylation. For each cell, we removed reads with genomic distance $<$15kb and only considered unique reads.

We generated 5 replications of 18 mixtures of scHi-C data at 10Mb resolution that consisted of 6 cell groups: oligodendrocyte (ODC), oligodendrocyte progenitor cell (OPC), astrocyte (Astro), microglia (MG), endothelial (Endo), and the 8 neuronal subtypes as one group (Neuron). We generated mixtures at the same three resolutions of Hi-C data as the mixtures from Ramani et al. (see Table A.3).

In order to assess the robustness of the reference-based deconvolution method, MuSiC, compared to reference-free deconvolution approaches we estimated cell type proportions under three scenarios.[112] First, we estimated cell type proportions where all cell types in the mixture were present in the reference panel. Second, we randomly removed one or two cells, respectively, from the reference panel and estimated the cell type proportions of the remaining cells.

2.1.2.3 Window Size

In large part, the 10Mb window choice was limited by the library size of current scHi-C datasets and sparsity of contacts from which to generate synthetic bulk Hi-C datasets such that the true cell type proportions are known. Additionally, we reported our computation test on 10Kb resolution Hi-C data that THUNDER scales up to the much larger feature space of finer resolution Hi-C data.

As single-cell technologies improve and with more data accumulating, we will be able to test Hi-C deconvolution methods at finer data resolutions where truth is known.

2.1.2.4 Feature Selection

The eleven feature selection methods either performed feature selection on the bulk Hi-C contact frequencies or on the derived cell-type specific profiles after an initial NMF fit. Strategies in the former group identify bin-pairs with high Fano Factor estimates across all samples. Strategies in the latter group identify informative bin-pairs with high cell-type specificity and/or high variation across inferred cell types. Cell type specificity was measured by feature score within a bin-pair and across estimated cell types. Across-cell-type variation was measured by standard deviation within a bin-pair and across estimated cell types. For both metrics, we used empirical thresholds based on the distribution of these estimates across all bin-pairs for feature selection.

2.1.2.5 Choosing Hi-C Readout for Deconvolution

Using the 12 simulated mixtures of HAP1, HeLa, and GM12878 cell lines from Ramani et al, we summarized the Hi-C contacts into varying readouts: 10Mb intrachromosomal contacts, 10Mb interchromosomal contacts, 1Mb intrachromosomal contacts, 1Mb interchromosomal contacts, 1Mb A/B compartment PC scores, and 100Kb insulation score. We computed normalized insulation score for 100Kb contacts with a sliding window size of 1.2Mb [**?**]. For insulation scores and compartment PCs, we apply the absolute value transformation to ensure that the input mixture matrices are non-negative. For each sample, we applied THUNDER to estimate cell type proportions using k=3. We compared the deconvolution performance at each readout of Hi-C data using MAD and correlation between estimated cell type proportions and true cell type proportions. Additionally, we computed the proportion of explained variance of the mixture matrix by the NMF fit. Specifically, given mixture matrix V and THUNDER estimated matrix, $H^* \times W^* = \hat{V}$,

$$RSS = \sum_{ij} (V_{ij} - \hat{V}_{ij})^2$$

and

$$\text{Proportion of Variance Explained} = 1 - \frac{RSS}{\sum_{ij} \hat{V}^2}$$

Similar measures of performance have been used previously to determine goodness of fit for NMF deconvolution estimates ([50, 33]).

2.1.2.6 Competing Methods

MuSiC and TOAST are each profiled in Chapter 1, but details of their application to our simulated Hi-C data are given below. To run MuSiC, we used the MuSiC R package (version 0.1.1) with default parameters. We constructed a scHi-C reference dataset using cells from Lee et al. which match cells considered in the simulated mixtures. Using multinomial sampling, we selected n cells from each cell type in the mixture where n is 75% of the minimum number of cells available in a given cell type within the Lee el al. dataset. We used the TOAST Bioconductor package version 1.0.0 using the default 1,000 features for deconvolution. Additionally, we used NMF with KL divergence function as the deconvolution engine of TOAST.

2.1.3 Real Data Analysis

2.1.3.1 Giusti-Rodríguez *et al.* eHi-C data

Anterior temporal cortex was dissected from postmortem samples from three adults of European ancestry with no known psychiatric or neurological disorder. Protocol for generating Hi-C data on these samples has been described previously[38]. We applied THUNDER to the three adult samples at 1Mb, 100Kb, and 40Kb resolutions. We ran THUNDER on intrachromosomal contacts only, and performed feature selection on each chromosome separately. To obtain the final estimated cell type proportions, we concatenated selected features across all chromosomes before running step 2 of the THUNDER algorithm. We assumed a range of possible values for the number of cells in the mixture ($k = 3, ..., 7$), and ran THUNDER for 100 iterations for both feature selection and cell

type proportion estimation. For downstream analysis, we chose the Hi-C bin size resolution and k value which maximized the proportion of variance explained in the subset mixture matrix by the final THUNDER deconvolution estimate.

After running THUNDER, we identified bin-pairs that demonstrated specificity to each inferred cell-type-profile. Informative bin-pairs were selected as specific to each inferred cell-type-profile if the row-normalized element of the basis matrix was greater than or equal to 0.3. This threshold was chosen to select a sufficient number of bin-pairs for each feature. We then compared the unique bins in these bin-pairs with cell-type specific epigenomic annotations (described below). We assigned cell types to the THUNDER inferred cluster-specific contact profiles based on the enrichment of epigenetic features within the THUNDER bins based on the results of a chi-squared test. Finally, we compared the THUNDER estimated cell-type proportions for each labeled cluster with the distribution of cell types within cortex tissue.

We tested if THUNDER bin pairs identify biologically relevant bin pairs by examining the gene expression distributions for cell-type-specifically expressed genes in each THUNDER cluster. Specifically, for each THUNDER feature of the final deconvolution estimate, we identified all cell-type-specifically expressed genes for neurons, oligodendrocytes, microglia, and astrocytes. After assigning the THUNDER features as described above, we tested the hypothesis that the gene expression distribution for genes in a THUNDER feature would be higher in the assigned cell type compared to other cell types using pairwise two-sample Wilcoxon rank sum tests.

2.1.3.2 Enhancer Annotations

We obtained cell-type specific enhancer annotations for neurons, microglia, oligodendrocytes, and astrocytes generated from Nott et al. They performed ATAC-seq as well as H3K27ac and H3K4me3 chromatin immunoprecipitation sequencing on cell-type specific nuclei. We did not consider cell-type specific enrichments for promoters due to previous evidence supporting that promoters are mostly conserved across cell types.[84]

2.1.3.3 Cell Type Specifically Expressed Genes.

We used cell-type specific RNA-seq data in neurons, microglia, oligodendrocytes, and astrocytes generated by Zhang et al. to identify cell type specific genes.[125] We defined a cell type specific gene as a gene where the difference between the cell type specific expression and the mean expression level of all other genes was greater than one. To examine overlap with Hi-C bins, we check the region within 2kb of the gene transcription start site.

2.1.3.4 High-confidence regulatory chromatin interactions

High confidence regulatory chromatin interactions (HCRCIs) are genomic regions physically proximal in the nuclear 3D space. HCRCIs were identified for the three adult cortex tissue samples as described above in a previous study.[38] HCRCIs are interacts that demonstrated significant evidence of increased interaction frequency ($p < 2.31 * 10^{-11}$) and overlapped with open chromatin, active histone marks, or transcription start sites of brain-expressed genes. Data were generated with two 10 Kb anchors that are $\geqslant$20 Kb and $\leqslant$ 2Mb apart.

2.1.3.5 Computation Test

In order to assess the computational costs of THUNDER on genome-wide Hi-C data, we applied THUNDER to intrachromosomal Hi-C data at 10Kb resolution in YRI samples.[40] We randomly selected 5 samples to be included in the analyses. First, we performed feature selection for each chromosome through simple parallelization. Then, we concatenated the selected features across all chromosomes for the final deconvolution estimate. We used computing time and memory usage to assess the computational efficiency for both feature selection and estimation of cell type proportions across the three datasets.

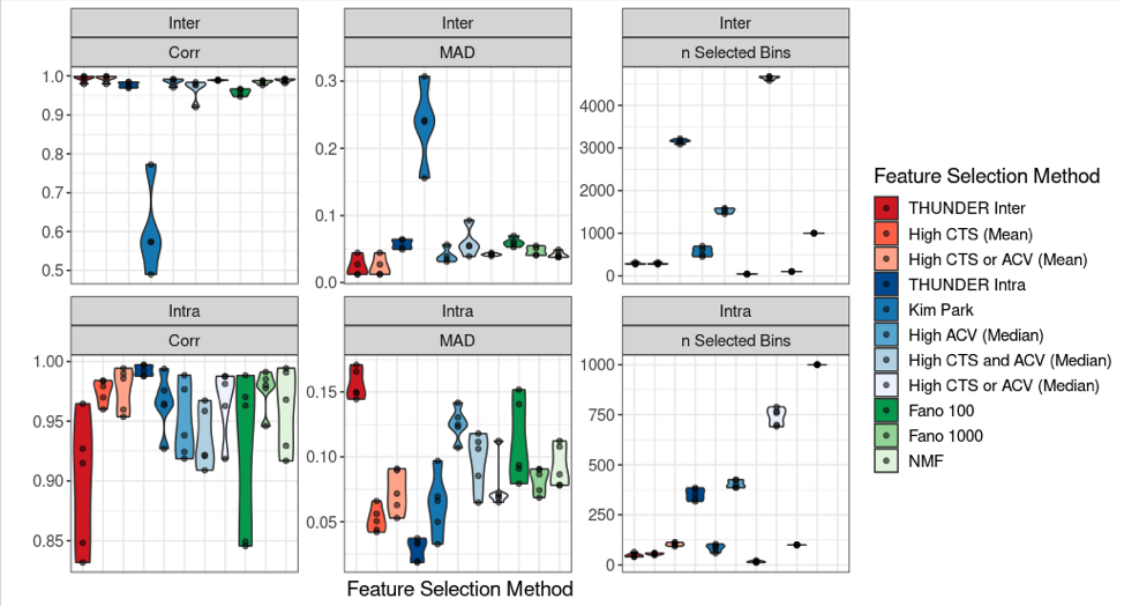

Figure 2.2: Performance of Feature Selection Strategies for Unsupervised Hi-C Deconvolution in HAP1, HeLa, and GM12878 Mixtures. We test 11 feature selection strategies including no feature selection (NMF), Fano 100, Fano 1,000, and 8 feature selection strategies combining bin-pairs with high cell-type specificity (CTS) and high across-cell-type variation (ACV). Colors are grouped such that the reds are strategies analyzing the estimated cell-type specific profiles using the mean across bin-pairs for thresholding, blues are feature score strategies analyzing the estimated cell-type specific profiles using the median across bin-pairs for thresholding, and greens are NMF with no feature selection or a pre-specified number of features based on Fano factor. Distributions are presented across simulation replicates.

2.2 Results

2.2.1 THUNDER Feature Selection

In order to determine the feature selection method for THUNDER, using scHi-C data generated from Ramani et al. [92], we simulated 12 mixtures of Hi-C data at 10Mb resolution consisting of three cell lines, HAP1, HeLa, and GM12878, where underlying composition proportions were known (details in Methods). We evaluated the performance of 11 published and novel NMF feature selection strategies for intrachromosomal only and interchromosomal only bin-pairs (see Table A.1).

Our simulation results suggest that the optimal feature selection method differs for deconvolving interchromosomal and intrachromosomal contacts (Figure 2.2). For intrachromosomal contacts, the best feature selection method is High CTS (median) which prioritizes features with high cell-type specificity using median-based empirical thresholds and selects an average of 353 informative bin-pairs out of an average of 2,590 input intrachromosomal contact features. The best performing interchromosomal feature selection method is High ACV. High ACV prioritizes features with high across-cell-type variation (ACV) using mean-based empirical thresholds and selects an average of 287 informative bin-pairs out of an average of 42,871 input interchromosomal contact features. We refer to these two methods hereforeward as THUNDER-intra and THUNDER-inter, respectively. Compared to NMF with no feature selection, THUNDER-intra reduced average MAD (mean absolute deviation, smaller indicates better performance) by 42% and increased average Pearson correlation by 0.4%. Similarly, THUNDER-inter reduced average MAD by 69% and increased average Pearson correlation by 3.3%. Feature selection methods that require specifying the number of informative bin-pairs a priori such as Fano-100 and Fano-1000, which selects the top 100 and 1000 features with highest Fano factor respectively, exhibit the most variable performance across simulations, and perform poorly relative to other methods despite specifying a similar number of bins.

Using simulated Hi-C mixtures from Ramani et al, we assessed THUNDERs performance across a variety of Hi-C data readouts including 10Mb intrachromosomal contacts, 10Mb interchromosomal contacts, 1Mb intrachromosomal contacts, 1Mb interchromosomal contacts, 1Mb A/B compartment PCs, and 100Kb insulation scores. We measured performance using proportion of explained variance by the THUNDER fit, MAD with true cell type proportions, and correlation with true cell type proportions. In all three measures, THUNDER deconvolution estimates were most accurate on 10Mb interchromosomal contacts and by 10Mb intrachromosomal contacts (Figure A.2 a-c). Notably, the next best performing inputs were 100Kb-resolution TAD insulation score and 1Mb-resolution A/B compartment PCs. This suggests that deconvolution of Hi-C data may be enhanced by summarizing Hi-C data to biologically relevant features before analysis. Across all simulation

results, MAD was negatively correlated with the proportion of variance explained by the THUNDER fit (Figure A.2 d). Additionally, the proportion of variance explained does not require knowledge of the true underlying cell type proportions to compute the goodness of fit. We therefore propose to use the proportion of variance explained as a practical solution to choose Hi-C data readout and the number of underlying cell types, k.

2.2.2 Simulations based on scHi-C from brain (Lee *et al.*)

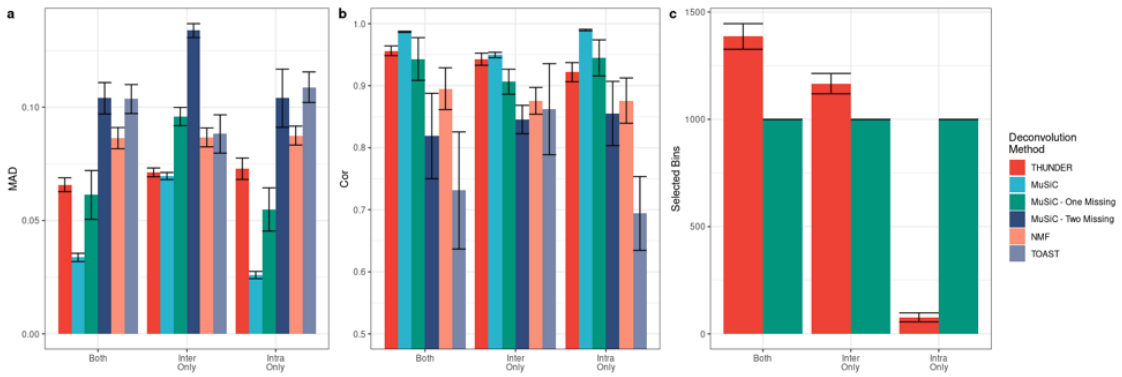

Figure 2.3: Performance of Deconvolution Methods on Mixtures with 6 Human Brain Cell Types. (a,b) The average mean absolute deviation (MAD) and average Pearson correlation comparing the true underlying cell type proportions to the simulated true proportions across simulations across 5 simulation replicates. Lower MAD and higher Pearson correlation indicates better performance. Error bars are equal to the standard deviation across simulation replications. (c) Number of bin-pairs selected by deconvolution methods which perform feature selection.

We tested the accuracy of THUNDER cell type proportion estimates using scHi-C data from Lee et al.[61] to simulate 18 Hi-C mixtures at 10Mb resolution of 6 brain cell types: microglia, astrocytes, oligodendrocytes, oligodendrocyte progenitor cells, endothelial cells, and neuronal cells. THUNDER cell type proportion estimates were most accurate when deconvolving intrachromosomal and interchromosomal contacts together, reducing MAD by 9.7% and 7.6% and increasing Pearson correlation by 3.7% and 1.4% compared to intrachromosomal contacts and interchromosomal contacts respectively. We compared THUNDERs performance to NMF with no feature selection, MuSiC, and TOAST on mixtures with both intrachromosomal and interchromosomal contacts, intrachromosomal contacts only, and interchromosomal contacts only (Figure

2.3). THUNDER outperformed all alternative reference-free deconvolution approaches in each simulation. When deconvolving both intrachromosomal and interchromosomal contacts together, THUNDER decreased average MAD by 23% and 36% and increased Pearson correlation by 6% and 31% relative to NMF and TOAST, respectively. MuSiC, a reference-based deconvolution approach, outperformed THUNDER in all simulation scenarios when all cell types in the mixtures are present in the reference panel. However, due to the current paucity of cell-type specific Hi-C reference panels, we tested the performance of MuSiC with one and two cell types randomly removed from the reference panel (Methods). In all three simulation settings, MuSiCs performance decreased with the number of cell types randomly removed from the reference (MuSiC, MuSiC - One Missing, and Music - Two Missing in Figure 2.3a,b). The performance of MuSiC one-missing was comparable to THUNDER in all simulation settings, and MuSiC - Two Missing was either worst or close to the worst performing methods. From our simulations, THUNDER performed best among reference free methods, and was more robust compared to MuSiC which performed poorly when cell types are missing from the reference panel. We anticipate reference based methods such as MuSiC will become more advantageous as we accumulate resources to build a comprehensive reference panel. Currently, with limited resources to construct a reference dataset, reference free methods are more valuable.

2.2.3 THUNDER estimates cell-type specific features from real brain Hi-C data (Giusti-Rodriguez *et al.*)

We applied THUNDER to bulk Hi-C data generated on cortex tissue from three postmortem adults samples (Methods). In downstream analysis, we proceeded with the deconvolution results when k=6 due to the greatest consistency across samples (see Figure A.3).

In order to assign plausible cell type labels to the 6 THUNDER inferred clusters, we compared the cluster-specific bins to cell-type specific enhancers and genes from four cell types commonly found in cortex tissue. 5 out of 6 THUNDER features (all except THUNDER cluster 4) demonstrated enrichment for neuronal enhancers ($p < 0.05/48 = 1.04 * 10^{-3}$), so we assigned each cluster to a

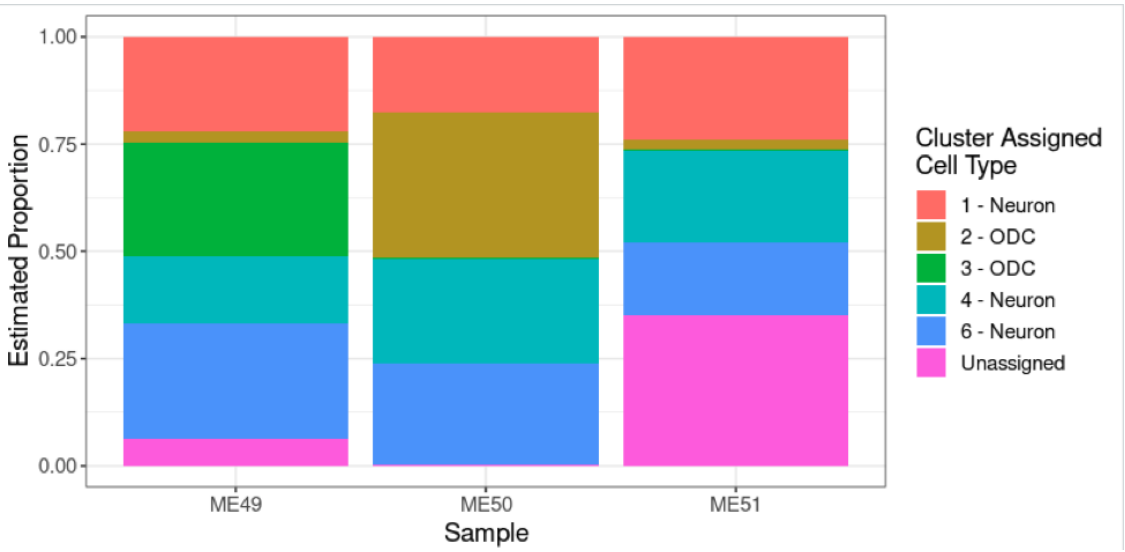

Figure 2.4: THUNDER Estimated Cell Type Proportions in 3 Samples of Human Cortex Tissue. We use THUNDER to estimate cell type proportions for 3 Hi-C samples from cortex tissue and perform enrichment analyses to assign brain cell types to THUNDER clusters. Our results match the expected ratio of neuronal to non-neuronal cells in cortex tissue.

cortical cell type based on other significant enrichments if possible. THUNDER cluster 1 showed evidence of enrichment for neuronal specifically expressed genes (p = 3.2e-3) and was thus assigned as neurons. THUNDER cluster 6 demonstrated enrichment for neuronal enhancers (p = 3.79e-9) and a trend (although not statistically significant) for enrichment of neuron specific genes (p = 0.104). We assigned THUNDER cluster 6 to neurons. THUNDER cluster 4 demonstrated enrichment with neuronal enhancers (p = 1.89e-3), and was thus assigned to neurons as well. Bins distinct to THUNDER clusters 2 and 3 demonstrated consistent evidence of enrichment of oligodendrocytes (ODC) features, in terms of enhancers (p = 3.3e-4 and p = 7.5e-9) and ODC-specifically expressed genes (p = 7.5e-3 and p = 4.78e-3). Therefore, both were assigned as ODC cells. THUNDER cluster 5 was not assigned to a cell type due to a lack of specific enrichments.

With these assigned cell type labels to the clusters, THUNDER estimated 62.7-65.2% neurons, 2.3-34.5% ODCs, and 0.3-35% unassigned for the three samples, largely matching the expected ratio of neuronal to non-neuronal cells in cortex tissue (see Figure 2.4).

To examine the biological relevance of the THUNDER inferred features, we compared the distribution of cell-type-specific gene expression across the four cell types in genes identified in feature-specific bin-pairs. Genes in bin-pairs specific to the THUNDER feature were enriched for cell-type-specifically expressed genes of the assigned cell type compared to the other three possible cell types (Fig A.4).

Additionally, THUNDER informative bin-pairs identified biologically relevant cell-type specific interactions. For example, the bin-pair defined by genomic regions chr5:130Mb-131Mb and chr5:131Mb-132Mb was an informative bin pair for THUNDER cluster 6, which was assigned to neurons via enrichment analysis. This bin-pair contained 14 high-confidence regulatory chromatin interactions (HCRCI) identified in the three adult cortical samples in a previous study with genomic coordinates within chr5:130600000-130970000 and chr5:131100000-131730000, respectively.[38] Further, two neuron-specific genes identified in our analysis of data from Zhang et al. were contained in chr5:131,100,000-131,730,000, *ACSL6* and *P4HA2*. Together, these results suggest that this THUNDER informative bin pair may correspond to a group of neuron-specific chromatin interactions. Another such example is the THUNDER informative bin-pair defined by the genomic regions chr12:121Mb-122Mb and chr12:122Mb-123Mb for THUNDER cluster 3, which enrichment analysis suggested as ODCs. The two regions defining this bin pair contained 64 HCRCIs, and two ODC specifically expressed genes, *P2RX7* and *ANAPC5*. Our results suggest that THUNDER estimated cell-type specific profiles can identify biologically meaningful cell-type specific interactions from bulk Hi-C data.

2.2.4 Computations on 10Kb Hi-C data

THUNDER scales linearly with both the number of samples under inference and the number of input features (see Tables A.4 - A.6). We assessed THUNDERs computing performance on Hi-C data of lymphoblastoid cell lines (LCLs) derived from five YRI (Yoruba in Ibadan, Nigeria) individuals.7 Specifically, we analyzed intrachromosomal contacts at 10Kb resolution, with 38,343,298 unique intrachromosomal bin-pairs ranging from 380,000 to 3.5 million bin-pairs per chromosome. To obtain cell type proportion estimates genome-wide using THUNDER, we first perform feature

selection by chromosome, then concatenate the selected features across chromosomes as input for the final deconvolution estimate. THUNDERs average computing time is 3.4 hours (range 0.6-7.2 hours) with an average of 57GB memory (range 18GB - 103GB) per chromosome using a single core on a 2.50 GHz Intel processor with 256GB of RAM. The final genome-wide estimation step to obtain cell type proportions, with 693,771 (2%) bin-pairs selected as informative, took 2.5 hours and 18GB of memory (see Table A.5). Similar summaries are presented for analyzing 3 and 10 YRI samples respectively (see Tables A.4 and A.6). One advantage of THUNDERs feature selection method when analyzing genome-wide Hi-C data is the ease with which it can be parallelized by subsetting the original input matrix in smaller regions than by chromosome, then concatenating Hi-C data for the final cell type proportion estimation step. This run time and memory usage serves as an upper limit on the computational costs of running THUNDER, as 10Kb is one of the finest resolutions of Hi-C data currently analyzed in practice.

2.3 Discussion

THUNDER is the first unsupervised deconvolution method for Hi-C data that integrates both intrachromosomal and interchromosomal contact information to estimate cell type proportions in multiple bulk Hi-C samples. Across all simulations, THUNDERs accuracy in estimating cell type proportions exceeded all reference-free alternative approaches tested. Importantly, THUNDERs feature selection strategy for identifying informative bin-pairs before deconvolution improves performance relative to NMF with no feature selection. We found THUNDER to be a robust alternative to reference-dependent methods which may not estimate cell type proportions accurately when cells are missing from the reference panel, a realistic scenario in practice with Hi-C data deconvolution. Further, we found that even in non-cancerous cell lines, the inclusion of sparse interchromosomal contact information (in addition to intrachromosomal contacts) improves deconvolution performance. This, however, comes at the cost of increased computational cost. THUNDER also provides an approach to infer cell-type-specific contact frequency from bulk Hi-C data.

We demonstrated that THUNDER successfully integrates interchromosomal contacts to improve deconvolution estimates for Hi-C data. In most cell types, we have more reliable Hi-C data at a much larger number of intrachromosomal bin-pairs compared to interchromosomal bin-pairs. For this reason, previous methods to deconvolve Hi-C data restricted their estimation to these intrachromosomal contacts. However, even in simulations with no strong interchromosomal signatures (for example, in the Lee et al. human brain data), THUNDERs performance improves when integrating interchromosomal and intrachromosomal data for deconvolution relative to only using intrachromosomal contacts. Our results suggest some value in including interchromosomal contacts bulk Hi-C deconvolution, though at the trade-off of computational efficiency. Since we analyze Hi-C data by grouping contacts into bin-pairs, the feature space increases rapidly with increasing bins. As demonstrated in our computation test, THUNDERs computation costs increase linearly as the number of features increases. Despite this trade-off, our results suggest that interchromosomal bin-pairs contain useful information that warrant consideration before excluding these bin-pairs in Hi-C deconvolution. Additionally, we demonstrate that THUNDER estimated cell-type-specific profiles are enriched for relevant cell-type-specific enhancers and specifically expressed genes through our analysis of 3 adult human cortex samples. We demonstrate how existing cell-type specific annotations can be used to label THUNDER inferred clusters, and thus provide cell type proportion estimates in real Hi-C data. Thus, the estimated cell type profile matrix serves a dual purpose: identifying informative bin-pairs from the large input feature space (dimension reduction) and accurately estimating relative cell-type-specific contact frequency at informative bin-pairs.

An additional application of these cell-type-specific contact profiles could be in fine mapping of GWAS variants in non-coding regions of the genome. Genome-wide association studies (GWAS) have identified over 300,000 unique associations between single-nucleotide polymorphisms (SNPs) and common diseases or traits of interest.[13] However, the majority of these SNPs reside in non-coding regions where little is understood about their underlying functional mechanisms, which has limited the adoption of variant-trait associations into revealing molecular mechanisms and further into transforming clinical practice. Functional annotation of GWAS results are often most

relevant in a cell-type-specific fashion due to important variability across cell types[63]. By further understanding the cell-type-specific interactome via THUNDERs estimated profiles, we anticipate more informative linking putatively causal variants identified by GWAS to the target genes on which they act.

We provide a statistical approach for selecting Hi-C bin size and k for the THUNDER deconvolution estimates that was correlated with accurately estimated cell type proportions in our real data-based simulations. Selecting these parameters are essential to an effective deconvolution approach for Hi-C data. We demonstrated the practical utility of our approach through our real Hi-C data analysis on data from Giusti et al. In addition to our goodness of fit metric proposed here, we recommend that analysts consider relevant information from histological experiments regarding the number of major cell types present in a tissue sample and the expected range of cell type proportions when evaluating the estimates provided by THUNDER. Additionally, analysts must consider the read depth of the Hi-C data when selecting the optimal resolution for deconvolution.

While we have presented results for Hi-C data here, the THUNDER algorithm could easily be modified to other variations of Hi-C data such as HiChIP/PLAC-seq data (HP data), which couple standard Hi-C with chromatin immunoprecipitation to profile chromatin interactions anchored at genomic regions bound by specific proteins or histone modifications, with reduced cost and enhanced resolution.[80, 31] Used in concert with methods to identify long-range chromatin interactions from HP data[54], our method is anticipated to efficiently leverage interchromosomal contacts jointly with high quality intrachromosomal contacts to estimate underlying cell type proportions. The robustness of our feature selection strategy and subsequent deconvolution performance warrant future interrogation in the setting of HP data.

There are two primary limitations of our study. First, due to the number of cells present in current scHi-C datasets and the library size, our simulation analysis may be biased toward coarser Hi-C resolutions due to increasing sparsity at lower bin sizes. However, we find that THUNDER still performs exceedingly well in estimating true cell type proportions in our real data analysis even at a coarser 1Mb resolution. Second, the number of cell types and the overall coverage of the

genome with our synthetic bulk Hi-C data are both much lower than one would expect in a realistic sample of bulk Hi-C data. As more scHi-C data becomes available, we hope to continue to test THUNDER in different real-data based scenarios which may be more realistic in terms of Hi-C datas read-depth.

To summarize, we present THUNDER, an unsupervised deconvolution approach tailored to the unique challenges of deconvolving Hi-C data. THUNDER accurately estimates cell type proportions in bulk Hi-C data. THUNDERs biologically motivated feature selection approach performs well in all of our real data or real-data based simulations, including human cell lines, human cortex tissue, and human brain cells. We have demonstrated the practical utility of the method through our analysis of Hi-C data from Giusti et al. and the computational efficiency of the method through our analysis of 10Kb resolution Hi-C data. Finally, the estimated cell-type-specific chromatin interactome profiles are valuable for identifying bin-pairs which interact differentially across cell types.

Accurately estimating underlying cell type proportions via THUNDER should be the first step in any individual-level differential analysis of bulk Hi-C data to control for the almost inevitable confounding factor of underlying cell type proportions. Additionally, THUNDER provides a unique tool to identify differentially interacting bin-pairs at the cell-type-specific level which can be associated with disease or phenotypes of interest. An R package for running THUNDER can be downloaded from https://github.com/brycerowland/thundeR.git. We anticipate THUNDER to become a convenient and essential tool in future multi-sample Hi-C data analysis.

CHAPTER 3: ESTIMATING POLYGENIC RISK SCORES IN ADMIXED INDIVIDUALS WITH MODIFIED FUSION PENALTIES

3.1 Methods

3.1.1 GAUDI Framework

Consider a sample of $i = 1, ..., n$ admixed individuals from two ancestral populations, A and B, and the problem of estimating the PRS in all individuals. In further work, this phenotype model can be extended to an arbitrary number of ancestral populations, but for simplicity our model is presented with two ancestral populations. Let x_{ij1} and x_{ij2} denote the haplotype of individual i for SNP j on the paternal and maternal chromosome, respectively. Let l_{ij1} and l_{ij2} denote the local ancestry for individual i at the position of SNP j on the paternal and maternal chromosomes, respectively, taking values of A or B for the corresponding ancestral population.

Let $Y = (y_1, ..., y_n)$ be an n x 1 phenotype vector, where

$$y_i = \sum_{j=1}^{p} \beta_{A,j}(x_{ij1}I(l_{ij1} = A) + x_{ij2}I(l_{ij2} = A))$$
$$+ \beta_{B,j}(x_{ij1}I(l_{ij1} = B) + x_{ij2}I(l_{ij2} = B)) + \epsilon_i)$$

where p is the total number of SNPs, and $I(\cdot)$ is the indicator function. Some subset of the SNPs, p^*, are causal, meaning that the effect of the SNP on the phenotype is non-zero. Under this model $\beta_{A,j}$ is the population A specific effect of SNP j on the phenotype. With no local ancestry information, or regards to haplotype information, this collapses to the usual phenotype model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$$

Our model can be expressed in matrix form where we define $x_{ijP} = x_{ij1}I(l_{ij1} = P) + x_{ij2}I(l_{ij2} = P)$ such that

$$G_{n\times 2p} = \begin{pmatrix} x_{11A} & x_{11B} & x_{12A} & x_{12B} & \cdots & x_{1pA} & x_{1pB} \\ x_{21A} & x_{21B} & x_{22A} & x_{22B} & \cdots & x_{2pA} & x_{2pB} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1A} & x_{n1B} & x_{n2A} & x_{n2B} & \cdots & x_{npA} & x_{npB} \end{pmatrix}_{n\times 2p}$$

and

$$\beta_{2p\times 1} = \begin{pmatrix} \beta_{A,1} \\ \beta_{B,2} \\ \vdots \\ \beta_{A,n} \\ \beta_{B,n} \end{pmatrix}_{2p\times 1}$$

Thus, our phenotype model can be expressed as

$$Y = G\beta + \epsilon_{n\times 1}$$

The problem of PRS estimation under this model is equivalent to the problem of the accurate estimation of the population specific effects for ancestral populations A and B given the correct design matrix.

3.1.2 GAUDI Overview

Our PRS estimation method for admixed individuals is a modified fused lasso approach. Specifically, given genotype information for n admixed individuals at p SNPs, some subset of which are causal variants, p^*. We assume that for each individual we have also obtained local ancestry inference estimates via RFMix, which also involves haplotype inference such that we have

haplotypes for each of the n samples. As noted in Chapter 1, large reference panels in ancestral populations are not necesssary for accurate local ancestry inference with RFMix.

First, we use the P+T strategy described in Chapter 1 in the admixed samples to identify variants that are marginally associated with the trait of interest at k pre-specified p-value thresholds, $(t_1, ..., t_k)$. We then perform LD clumping in PLINK using options –indep-pairwise 500 5 0.5.[90] For each SNP achieving the p-value threshold, t, and passing LD clumping, p_t total SNPs, we use five-fold cross validation to estimate the tuning parameters using the following fused lasso objective function.

$$\underset{\beta_{2p_t}}{\operatorname{argmin}} \frac{1}{2}||Y_{n\times 1} - G_{n\times 2p_t}\beta_{2p_t\times 1}||_2^2 + \lambda||D_{3p_t\times 2p_t}\beta_{2p_t\times 1}||_1$$

where

$$D_{(3p_t)\times 2p_t} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \\ \gamma & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \gamma & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \gamma \end{pmatrix}$$

In this model, λ controls the overall strength of the penalty matrix, and γ controls the ratio of penalty between sparsity and fusion. In a similar way to elastic net trading off between the lasso and ridge penalty via a tuning parameter, γ trades off between sparsity and fusion penalties to model ancestry specific effects. One notable difference between our PRS estimation approach for admixed individuals, GAUDI, and traditional fused lasso is that only ancestry-specific effects for a given SNP are penalized with fusion, rather than all adjacent parameters. This penalization can easily be extended to consider more than two ancestral populations, but is not considered here.

The fitted PRS is thus,

$$Y = G\hat{\beta}_{t_i,\lambda,\gamma}$$

Cross-validated model performance for tuning parameters (the p-value threshold, λ, and γ) is optimized on the squared Pearson correlation coefficient between the phenotype and the fitted PRS.

In using the cross validation procedure to tune our model parameters, several practical model fitting considerations must be addressed. Since we include variants with low minor allele counts, it is possible that alleles become fixed in one of the training folds. Since our model fitting procedure drops fixed variants, we drop the population-specific term for the fixed variant, but may keep in the population-specific term for the other ancestral population given that there are greater than two variant calls. Additionally, we remove highly correlated population-specific variant columns that may not have been filtered out via LD pruning. Since LD pruning is agnostic to local ancestry, it is possible that the aggregate test for linkage equilibrium between all variants in the sample is satisfied, but underlying LD occurs due to patterns of genetic admixture. In this case, we remove any variants in with LD $r^2 > .95$ in model fitting, maximizing the total number of variants remaining in the model in linkage-equilibrium.

3.1.3 COSI Simulations

In order to simulate haplotypes of recent admixture, we used COSI to generate 500kb regions for 3,500 African American individuals.[97] We made two primary assumptions in generating our simulated haplotypes. First, we assumed that the global ancestry proportions of our African Americans were 80% African and 20% European ancestries, respectively. Second, using empirical estimates o f ancestral switch-points based on an analysis of TOPMed individuals, we assumed 4% of 500Kb regions would contain ancestry switchpoint events.[114] Thus, for 3,500 diploid individuals, 280 chromosomes will contain switch points (7,000 * 0.04 = 280). For each chromosome carrying an ancestry switch point, we generated one European and one African chromosome to simulate the admixture event at a random base-pair in the region. For the remaining 6,720 chromosomes

with no admixture events, we generated 80% African chromosomes (n = 5,376) and 20% European chromosomes (n = 1,344). Additionally, we simulated 5,000 European chromosomes and 5,000 African chromosomes to be used as reference for relevant methods.

3.1.4 Phenotype Simulations

We simulate phenotypes using 500kb regions generated from COSI for the 3,500 admixed individuals and the 2,500 reference AFR and EUR individuals. We considered three distinct sets of causal SNPs for inclusion in the simulated phenotypes in order to represent differing genetic architectures. First, we created the "Globally Common SNPs" genetic architecture; at a locus, we considered all SNPs that had AFR and EUR specific MAF $\geqslant 0.05$. Second, we created the "EUR Common SNPs" genetic architecture; we considered all SNPs that had AFR MAF <0.05 and EUR MAF $\geqslant 0.05$. Notably for estimating PRS in admixed populations, many discovered GWAS significant variants fall into this genetic architecture due to European-centric biases in GWAS results.[75, 76] However, the most common category of discovered GWAS variant is common across all ancestries. Third, we created the "AFR Common SNPs" genetic architecture; we considered all SNPs that had AFR MAF $\geqslant 0.05$ and EUR MAF < 0.05.

Across simulations we vary 5 values. First, we vary p_{causal} or the proportion of causal SNPs which takes on three possible values (1, 0.5, 0.05). Second, we vary p_{shared}, or the proportion of SNPs that have the same effect size across ancestry groups which took three possible values (1, 0.8, 0.5). Third, we varied heritability (h^2), or the proportion of variation explained by genetic effects, which took possible values of 0.2 or 0.6. Finally, we varied r^2, the maximum allowed correlation between SNPs in the phenotype, which took values of 0.2, 0.5, and 1. All results presented here have $r^2 = 0.5$. For varying the LD between SNPs in the phenotype, we performed LD pruning on the set of included SNPs within each genetic architecture using PLINK (–indep-pairwise 500 5 r^2).[90] We repeat each combination of the above four values via 10 replicates for each genetic architecture.

Thus, for a SNP j, we simulated effect sizes from the following distribution.

$$\begin{cases} \beta_{A,j} = \beta_{B,j} \sim N(0,1) & \text{w.p.} \quad p_{causal}p_{shared} \\ \beta_{A,j} \sim N(0,1), \beta_{B,j} \sim N(0,1) & \text{w.p.} \quad p_{causal}(1-p_{shared}) \\ 0 & \text{w.p.} \quad (1-p_{causal}) \end{cases}$$

We then estimated the variance explained by the causal SNPs, and simulated error terms using normally distributed errors such that the total heritability was equal to h2.

3.1.5 Prediction Accuracy Measurements

We assessed the performance of all PRS estimation methods by computing the squared Pearson correlation coefficient between the simulated phenotype and the estimated PRS in the held-out testing samples.

3.1.6 Real Data Analysis

For our real data analysis, we considered subjects from two studies: African American individuals from the WHI study genotyped on the MEGA array (n = 6,734), and European ancestry individuals from the WHI WHIMS sub-study (n = 5,681) downloaded from the dbGaP web site under phs000675.v4.p3.[117] For all models, we partitioned the WHI AA individuals into five non-overlapping testing folds consisting of 20% of the data, and the remaining 80% were used for model training. Thus, each training procedure was repeated five times across each of the training partitions. All WHI WHIMS European individuals were included for training in all models. Individuals were included in modeling if they had non-missing phenotype data as described below.

3.1.6.1 WHI Cohort Description

The Womens Health Initiative (WHI) is one of the largest (n=161,808) studies of womens health ever undertaken in the U.S. There are two major components of WHI: (1) a clinical trial

(CT) that enrolled and randomized 68,132 women ages 50−79 into at least one of three placebo control clinical trials (hormone therapy, dietary modification, and supplementation with calcium and vitamin D); and (2) an observational study (OS) that enrolled 93,676 women of the same age range into a parallel prospective cohort study.[4] A diverse population including 26,045 (17%) women from minority groups was recruited from 1993 to 1998 at 40 clinical centers across the U.S. Details on the study design, eligibility, recruitment, and the reliability of the baseline measures of demographic and health characteristics have been published elsewhere.[4, 58] Among the U.S. minority participants enrolled in WHI, 12,468 women (including 6,829 self-identified African American and 4,626 self-identified Hispanic subjects) consenting to genetic research were included in PAGE II for genotyping with the Multi-Ethnic Genotyping Array (MEGA).[9] Fasting blood samples were obtained from all participants at baseline and were analyzed for white blood cell count and platelet count by certified laboratories at each of the 40 clinical centers as part of a complete blood count.[58] Results were entered into the WHI database at each clinical center and were reviewed by clinical center staff .[28] In addition to the main WHI CT and OS, the WHIMS ancillary study contributed existing GWAS data on the HumanOmniExpressExome-8v1_B array and CBC measurements to this study.

3.1.6.2 Phenotype QC

We considered four blood cell phenotypes with low levels of missing data across the two cohorts: white blood cell count, platelet count, hematocrit, and hemoglobin. All phenotypes were adjusted by cohort for age, age^2, top 10 genotype PCs, center, genotyping array, and sex using linear regression models. White blood cell count values were log10(x + 1) transformed before regression. Residuals from the regression models were inverse normal transformed and serve as adjusted phenotypes.

3.1.6.3 Variant QC

For the GWAS association tests, we consider common variants (MAF > 0.01) with Rsq > 0.3, and we consider rare variants ($0.001 <$ MAF < 0.01) with Rsq > 0.6. Note that for our training samples, MAF = 0.001 corresponds to a MAC of approximately 10.

3.1.6.4 GWAS with REGENIE

We performed a GWAS on the four blood cell phenotypes and the quality controlled set of variants for each of the five training folds using REGENIE.[77] To fit the REGENIE null model accounting for cryptic relatedness, we used extremely-well imputed variants from REGENIE (MAF > 0.2, Rsq > 0.9999). We fit the four phenotypes simultaneously using the grouping options available in REGENIE, and set the number of blocks to be 1,000.

3.1.6.5 Local Ancestry Inference with RFMix

For the African American samples, we inferred local ancestry using RFMix using 1kGP phase 3 as a reference panel. We consider only two-way admixture between European and African ancestral populations, since in our downstream models we assume admixture between two continental ancestry groups. Before local ancestry inference, we filtered imputed SNPs by minor allele frequency $>$ 0.05 and constructed a reference panel. We included both 92 samples from European and African ancestry to make the reference panel balanced.

3.1.6.6 Variant Selection Experiments

We used two approaches to select variants for inclusion in the PRSs across the methodologies considered. First, we considered an unrestricted set of variants that passed the GWAS variant-filtering criteria. For each method, a different variable selection procedure is needed, and we compared PRS predictions using the best set of variants according to each method.

Second, we considered PRS fit on a common set of variants generated from one round of sequential conditional analysis. In brief, we grouped the REGENIE GWAS results into +/- 1Mb

windows surrounding the top GWAS SNP starting from the most significant GWAS locus and continuing until there are no more variants not in a locus with p-value $<$ 5e-6. Then, we used REGENIE to compute an association test for each variant at the locus conditioned on the sentinel variant at the locus. We then selected the most significant variant at each locus with p $<$ 5e-4 if one exists. This set of sentinel variants and conditional significant variants formed the basis of our PRS variant set. We then performed LD clumping by population group (50kb, r^2 = 0.8), then took the union of the variants across ancestry groups and ran LD clumping (250kb, r^2 = 0.5) using the initial GWAS p-values in African Americans. The output list of this procedure was our high quality variant set.

For both experiments, we summarized the performance of each method with the mean r^2 between the adjusted phenotype values and the PRS prediction across the 5 training partitions, and a range of the test r^2 values.

3.1.6.7 GAUDI

When analyzing real data, we included variants that had a minor allele count $>=$ 10 on at least once ancestral haplotype. If the variant was fixed in one ancestral population, we included only one ancestry-specific effect in the model. If the variant was observed in both populations, we included both ancestry specific effects in the model.

3.1.6.8 PRSice

PRSice is a popular software implementation of the P+T thresholding method.[30] We applied PRSice to the REGENIE summary statistics in both African American individuals and European individuals without using local ancestry information. We ran PRSice with default parameters.

3.1.6.9 Partial PRS

As discussed in Chapter 1, partial PRS (pPRS) is a method to incorporate local ancestry information in PRS estimation in admixed individuals.[73] Using our RFMix inferred local ancestry,

we applied the PRSice trained PRS in the training samples of African Americans and Europeans using the PRSice determined p-value thresholds.

3.2 Results

3.2.1 COSI Simulation Results

3.2.1.1 Simulated Phenotypes with no Ancestry Specific Effects

We first assessed the performance of GAUDI under the assumption that there are no ancestry specific effects for causal SNPs with LD between causal SNPs having a maximum value of $r^2 = 0.5$. While the shared ancestry effects assumption is an oversimplification, recent work has shown that there is almost always a positive correlation between effect sizes across global populations for most SNPs associated with complex traits.[110] We compared GAUDI to the P+T method as implemented in the PRSice software and the partial PRS detailed in Chapter 1.[73]

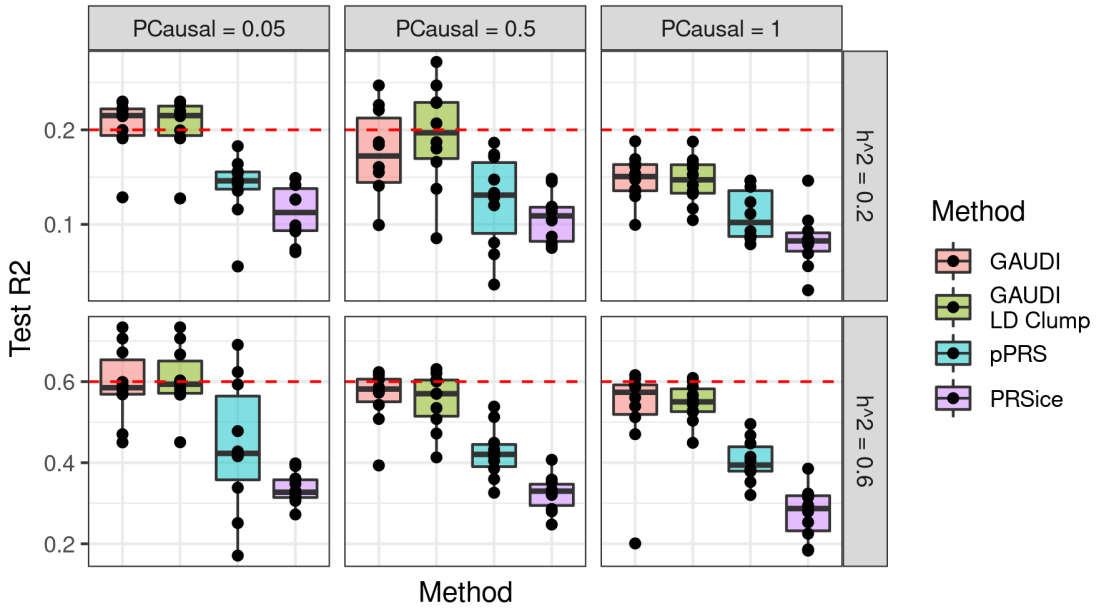

Figure 3.1: GAUDI Simulation Results - Shared Traits, Phenotype 3. Evaluation of PRS methods on COSI simulated admixed genotypes and phenotypes where causal effects are assumed to be shared across ancestral populations. All causal SNPs have MAF $\geqslant 0.05$ in the AFR reference data and MAF < 0.05 in the EUR reference data.

Under the "AFR Common SNPs" genetic architecture, GAUDI outperformed PRSice and partial PRS across all simulated traits in the held-out testing data. GAUDI demonstrated better performance in settings with higher heritability and sparser phenotypes (Figure 3.1). However, GAUDI's out of sample performance was nearly equal to heritability in almost all simulated phenotypes. In this setting, we expected GAUDI to outperform alternative methods by borrowing information from the AFR segments of haplotypes to estimate the EUR effect. We also expected the partial PRS to outperform the PRSice PRS by accounting for local ancestry when applying the PRS, but we observed this only in near-omnigenic phenotypes.

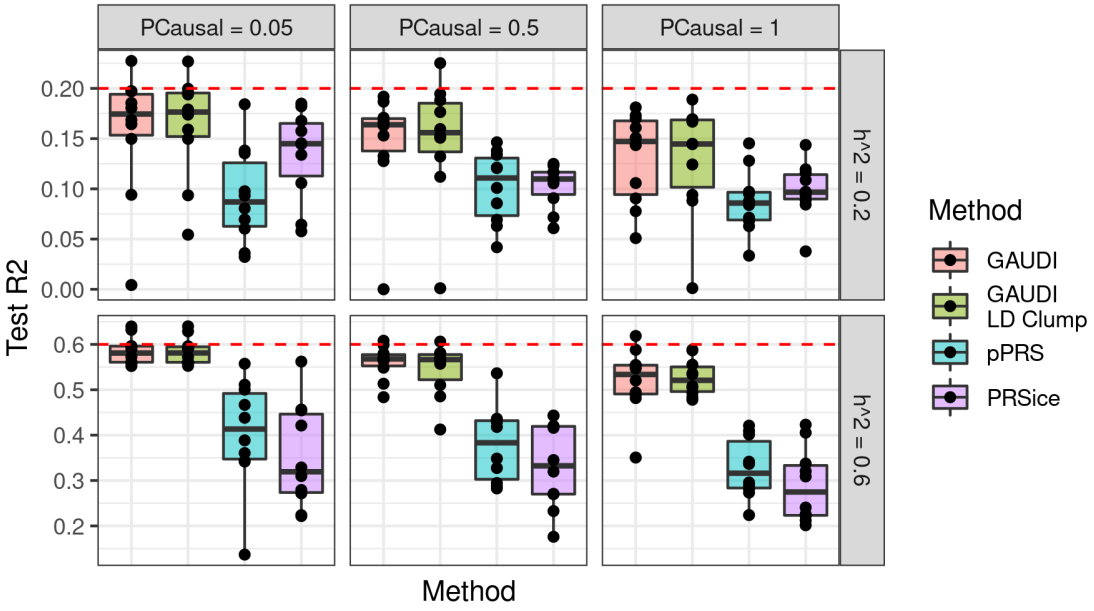

Figure 3.2: GAUDI Simulation Results - Shared Traits, Phenotype 2. Evaluation of PRS methods on COSI simulated admixed genotypes and phenotypes where causal effects are assumed to be shared across ancestral populations. All causal SNPs have MAF $\geqslant 0.05$ in the EUR reference data and MAF < 0.05 in the AFR reference data.

We simulated additional traits under the "EUR Common SNPs" genetic architecture. These phenotypes are comprised of variants reflecting the current bias toward common EUR variants in published GWAS results. Applying PRSice naively to admixed individuals in this setting resulted in predictions far below heritability (see Figure 3.2). Additionally, pPRS and PRSice performed equivalently in simulations with more causal variants, suggesting that pPRS peformance changes

based on the global ancestry proportions of admixed individuals and this interaction with the genetic architecture of a phenotype. Despite EUR segments comprising only 20% of the COSI admixed haplotypes on average, GAUDI demonstrated substantial gains by borrowing information across ancestral populations to estimate the PRS. The increase in predictive performance between PRSice and GAUDI suggests an immediate impact of applying GAUDI to estimate PRS in previously published PRS composed of variants discovered in large European cohorts.

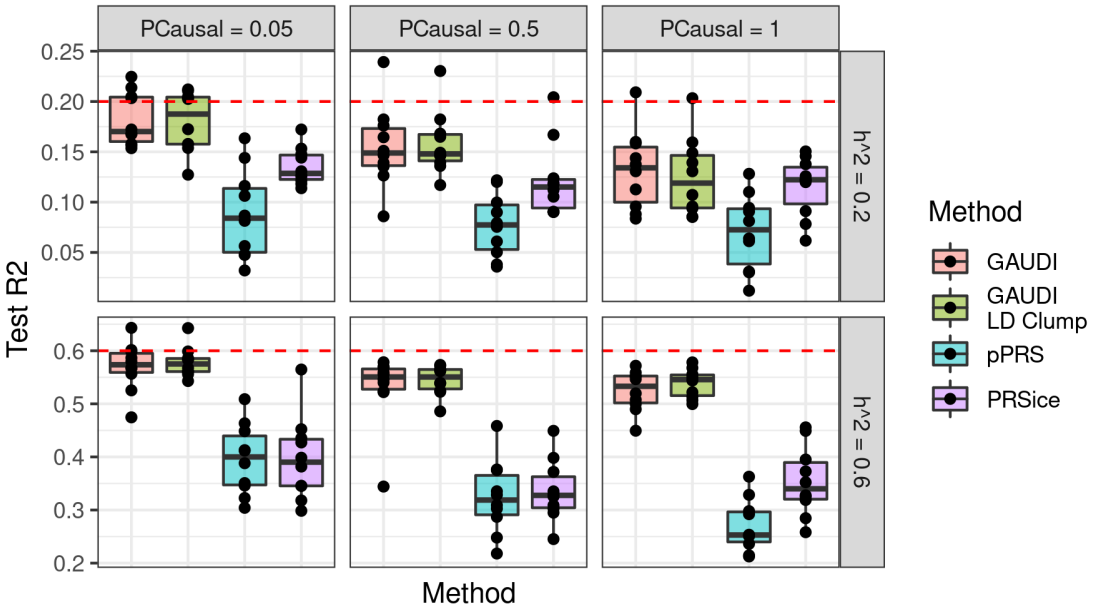

Figure 3.3: GAUDI Simulation Results - Shared Traits, Phenotype 1. Evaluation of PRS methods on COSI simulated admixed genotypes and phenotypes where causal effects are assumed to be shared across ancestral populations. All causal SNPs have MAF $\geqslant 0.05$ in the EUR reference data and MAF $\geqslant 0.05$ in the AFR reference data.

Finally, we simulated traits assuming the ”Globally Common SNPs” genetic architecture. GAUDI still performed the best under this simulation framework, however in situations of highly polygenic traits with low heritability the improvement due to penalized regression was less pronounced. These results demonstrate GAUDI's utility across a range of genetic architectures in admixed populations.

3.2.1.2 Simulated Phenotypes with Ancestry Specific Effects

We then simulated phenotypes where 50% of causal SNPs had ancestry specific effects and the remaining 50% had shared effect sizes across ancestral populations. Our results largely matched the above simulations. Surprisingly, in some scenarios the difference between GAUDI and competing methods was even higher with the introduction of ancestry specific effects.

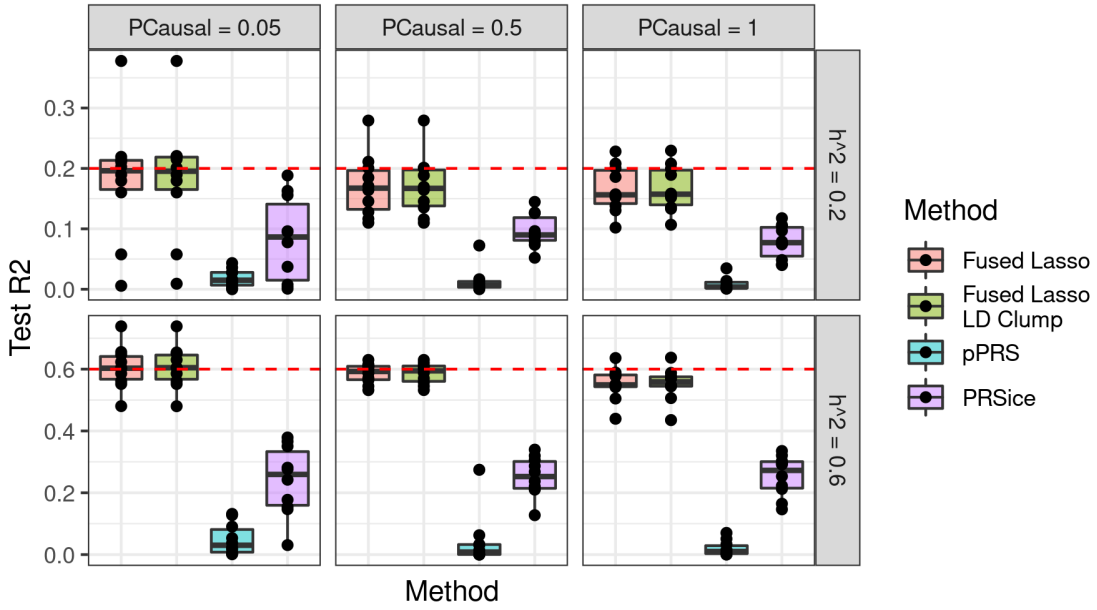

Figure 3.4: GAUDI Simulation Results - Specific Traits, Phenotype 1. Evaluation of PRS methods on COSI simulated admixed genotypes and phenotypes where 50% of causal effects are shared across ancestral populations and 50% are specific to the ancestral populations. All causal SNPs have MAF < 0.05 in the EUR reference data and MAF $\geqslant 0.05$ in the AFR reference data.

Under the genetic architecture "AFR Common SNPs", GAUDI estimated PRS captured almost all the heritability in testing samples in most phenotypes. Variability across simulations was reduced compared to GAUDI's performance in simulations with all ancestry shared effects. GAUDI PRS estimates capture the highest proportion of heritability in the "AFR Common SNPs" genetic architecture (Figure 3.4), and considerable gains are made relative to other competing methods in the EUR Common SNPs and "Globally Common SNPs" genetic architecture (Figures B.1, B.2). Under the Globally Common SNPs simulation framework, the difference between GAUDI and

PRSice is much larger when ancestry specific effects are introduced compared to the simulated phenotypes with no ancestry-specific effects (Figure B.2). These results demonstrate that GAUDI can accurately estimate both ancestry shared and ancestry specific effects to estimate polygenic risk in admixed individuals.

3.2.2 Real Data Analysis - All GWAS Variants

We applied GAUDI on a dataset of 6,734 African American individuals from the WHI study. Due to phenotype availability in WHI, we focused our study on white blood cell count (WBC#), platelet count (PLT#), hematocrit (HCT), and hemoglobin (HGB). We partitioned individuals into five equal folds, and estimated a PRS for the four phenotypes within each fold. Performance was assessed by the mean r^2 value across folds between the predicted PRS and the phenotype values in the held out samples. Additionally, we compared GAUDIs performance to PRSice and partial PRS (pPRS) incorporating local ancestry on the same training and testing data. For pPRS, we used a sample of 5,681 European ancestry individuals from the WHI WHIMS substudy as additional training data.

Across the four available phenotypes, only PLT# and WBC# have meaningfully non-zero test r^2 values. Given recent applications of PRS to blood cell traits in African American samples and our sample size comparatively, this relative order and magnitude of prediction accuracy was expected.[17] Incorporating local ancestry estimates for AA samples with GAUDI ($r^2 = 0.081$) improved mean cross-validated test r^2 by 48% compared to PRSice ($r^2 = 0.055$; Figure 3.5). This improvement is despite the fact that we restricted the variants considered by GAUDI to those with GWAS p-value less than $5 * 10^{-5}$ due to the sample size and PRSice considered all GWAS variants. GAUDI estimated WBC# PRSs contained an average of 2596 variants across all folds, while the PRSice score had an average of 497,200 variants. PRSice more accurately estimates the PLT# phenotype in test samples, likely due to limited power in the GAUDI model due to a doubled number of parameters. Additionally, GAUDI improved performance compared to pPRS ($r^2 = 0.039$) for WBC#, another method that also incorporates local ancestry when estimating a PRS. In fact, PRSice

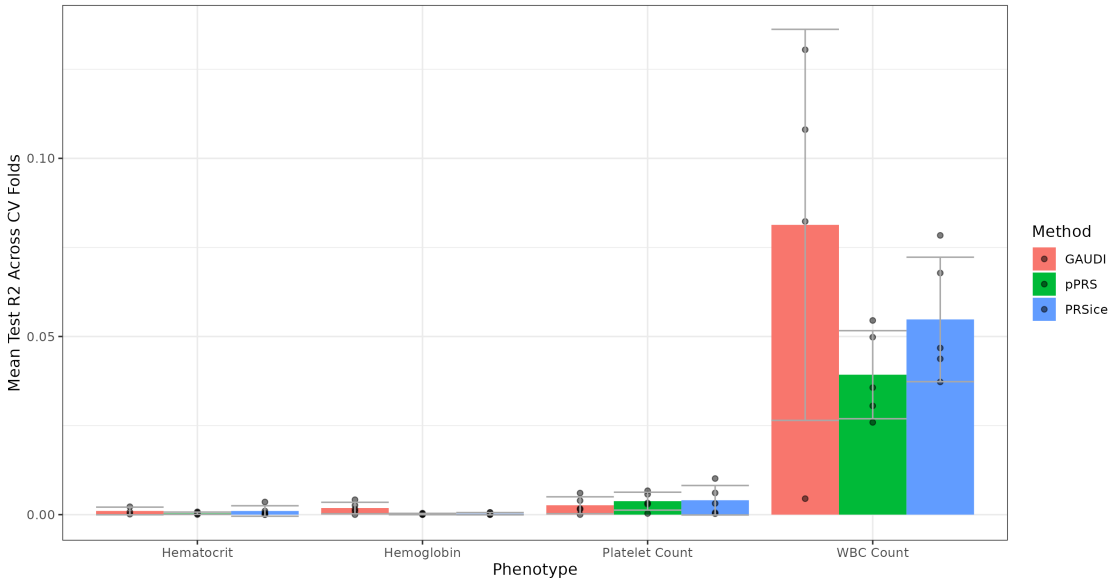

Figure 3.5: Method Comparison on WHI AAs - All GWAS Variants. GAUDI improves inference compared to alternative methods in WBC# phenotype prediction despite using fewer variants.

on only African American training samples outperformed pPRS. One consideration is that pPRS does not have a procedure to recalibrate the p-value thresholds derived from the ancestral samples on which pPRS is trained. The ideal p-value thresholds when training two PRS individually may not be the ideal p-value thresholds in joint modeling.

3.2.3 Real Data Analysis - Variants Derived from Conditional Analysis

Since the number of parameters to estimate in the GAUDI model is double that of traditional PRS methods, we estimated a PRS on a common set of variants determined from one round of sequential conditional analysis. We hypothesize that this set of variants will capture the majority of the heritability to be explained at the GWAS significant loci given the smaller sample size. One advantage of this approach is that the only difference between the PRSice and GAUDI estimates is the modleing of local ancestry in the PRS estimation, not differences in variant selection. Similar approaches to PRS construction using significant variants from GWAS conditional analysis

have recently demonstrated better performance to P+T or models that assume omnigenic genetic architectures.[21]

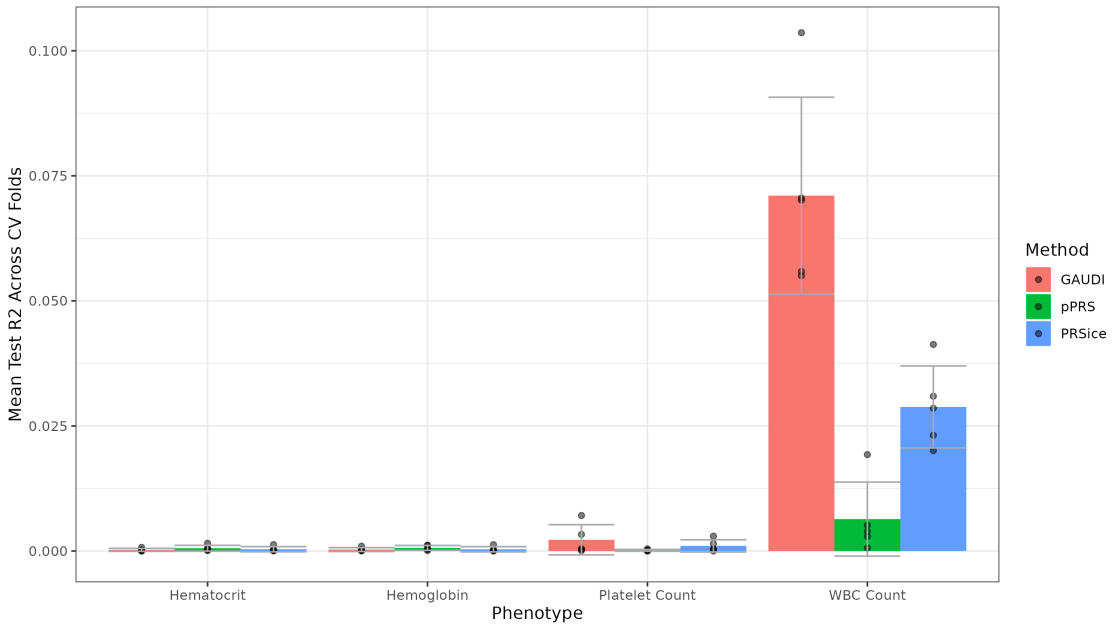

Figure 3.6: Method Comparison on WHI AAs - CA Variants. Using a common set of variants, GAUDI performs best on WBC# phenotype in out of sample prediction.

As with the experiment using all GWAS variants to estimate PRS, we observed meaningful non-zero predictions only in WBC# and PLT#. After conditional analysis and LD clumping within and across ancestral populations by cross-validation fold, we established a set of variants of average size 361 and 139 variants for the WBC# and PLT# PRSs, respectively. Incorporating local ancestry information with GAUDI ($r^2 = 0.071$) on the same set of variants improved prediction accuracy in test samples compared to PRSice ($r^2 = 0.029$) by 145% (Figure 3.6). This performance gain was driven in large part by modeling the effect of the Duffy null variant, which has highly differentiated allele frequencies across populations (Figure B.3). pPRS again performs the worst out of the three methods considered here. Performance in PLT# was comparable across all methods.

3.3 Discussion

We present GAUDI, a PRS estimation method incorporating local ancestry to improve PRS estimation in admixed individuals. In simulations of various genetic architectures, GAUDI outperformed P+T PRS and pPRS methods. Further, we demonstrated the practical utility of GAUDI in data from the WHI study by estimating blood cell PRSs in admixed individuals. We captured differential effect sizes at variants with known differential allele frequencies across populations related to white blood cell traits (the Duffy null variant).[95, 94]

Since both GAUDI and PRSice outperformed pPRS for WBC# PRS, our results suggest that the inclusion of local ancestry to estimate a PRS must not be done without care. In addition to potential problems with the calibration of the p-value thresholds in the pPRS approach, variants which are rare in one population are only included in the PRS derived in that ancestral population. GAUDI uses fusion penalties to jointly model the SNP effects, even if one is rare on an ancestral haplotype. How to best incorporate each individual's unique mosaic of local ancestry in PRS training is an essential area of further study.

GAUDI currently has several limitations which could be addressed as future directions. Primarily, we believe that by adding an extra regression parameter per variant, GAUDIs prediction utility will be limited by a lack of power, especially while data is still limited in individuals with genetic admixture. First, we will explore using meta-analyses during variable selection to increase the sample size of our real data analysis. One possible alternative would be to extend GAUDI to a summary statistics based approach, but this path forward has several challenges as discussed below. Second, we do not consider three-way admixture in this work, but conceptually the GAUDI model is easily extended to three-way admixture with the construction of a new penalty matrix. Third, our current variant selection procedure for GAUDI does not incorporate local ancestry to identify plausible variants. One alternative would be to use methods such as TRACTOR to identify variants while accounting for the population structure induced by local ancestry mosaics.[5] However, our aim with this work has not been to identify the optimal set of causal variants to include in a

local-ancestry informed PRS, but how to use local ancestry to account for population structure in PRS estimation given a plausible set of variants. A final future direction of this work would be to extend our approach to a cross-population meta-analysis for variant selection (potentially using methods like TRACTOR to account for local ancestry).[17, 14] This approach would allow us to assess if using an independent set of European participants could improve European effect size estimation at inferred regions of European ancestry in African Americans.

As mentioned above, much work is necessary to extend summary statistics based methods for PRS estimation to admixed populations. Summary statistic-based methods are the most popular approach to estimating PRSs since GWAS summary statistics can be shared without prior data sharing agreements. Our simulation work demonstrates the shortcomings of the pPRS approach to train PRS for admixed individuals using summary statistics in the ancestral population.[73] Thus, I believe some level of incorporation of individual-level patterns of local ancestry, as with the GAUDI model, will be necessary for accurate PRS estimation in admixed individuals.

A second approach would be to apply or modify existing summary statistic based methods to train on admixed individuals. However, these summary statistic based methods rely on external LD reference panels to re-estimate variant effect sizes from GWAS summary statistics (see Chapter 1).[111, 123] This reliance on LD reference panels presents a host of challenges for admixed individuals. One challenge is defining a singular population of admixed individuals from a set of ancestral populations for which to generate an LD reference panel since global ancestry can vary continuously between 0 and 1 for each ancestral population within a sample of admixed individuals. Even if an appropriate sample could be defined using global ancestry thresholds (as with our simulation design), the relationship between admixture events and linkage between two alleles remains challenging to assess at the population level since LD is highly dependent on admixture dynamics.[87] Furthermore, population level summaries such as r^2 between two variants obscure the unique individual-level patterns of ancestral mosaics which may alter the effect of a variant given differing allele frequencies or a population-specific genetic effect. I believe this is a crucial road-block to advancing individualized genetic predictions from summary statistics data.

More methodological thought is needed to solve this problem. The most proactive solution is to normalize the release of an LD reference panel with all published summary statistics, as has been suggested elsewhere.[70] However, for admixed individuals, this may not be enough to disentangle the relationship between individual patterns of local ancestry and LD as mentioned above. Thus, the research community could continue development of appropriate summary statistics based methods for admixed individuals while promoting greater access to LD reference panels in which summary statistics were developed.

Polygenic risk score estimation methods could be further improved through the incorporation of functional information to better identify variants relevant to the prediction task at hand.[3, 65, 115, 18] The problem of variant selection for PRS estimation is the problem of identifying causal variants or their proxies, which we discuss more in depth in Chapter 5 in the context of non-European populations. Additionally, as clinical decision makers begin to incorporate PRS in practice, studies which collect both treatment information and genetic information could serve as valuable datasets for retrospective analyses to evaluate if the current methods of PRS are achieving the risk stratification they claim. Due to the ethical consequences of miscalibration, a PRS should not be treated like other variables to be stratified upon in a study design such as age or sex, but rather like a treatment that must demonstrate clinical benefit before widespread use.[76] These sorts of re-analyses of previous datasets are common in the clinical trials literature, and should be adopted to better understand the relationship between PRS estimation and clinical decision making.

As studies continue to enroll more participants with significant continental admixture, GAUDI will prove a valuable tool for PRS estimation. We have provided code to perform data analysis tasks related to PRS estimation with local ancestry as well as estimate parameters in the GAUDI model at GAUDI GitHub.

CHAPTER 4: SMA SILENT CARRIER SCREENING WITH MODIFIED PRS APPROACHES IN DIVERSE POPULATIONS

4.1 Introduction

In Chapter 4, we extend multi-variant prediction methods for polygenic scores to the challenging setting of predicting SMA Carrier status for silent carriers. This prediction problem is a pressing clinical and methodological challenge as no test exists that can reliably detect individuals with the SMA silent carrier haplotype. Currently, a SNP has been proposed to identify silent carriers for SMA, but has PPV $<2\%$ in AFR populations, the population in which the *SMN1* duplication allele is most prevalent.[19]

This application presents several methodological challenges currently unaddressed in PRS literature. First, the *SMN1/2* region is characterized by a complex pattern of gene duplications and inversions, which make applying traditional genetic prediction models based on genotypes and linkage disequilibrium challenging. Second, the SMA silent carrier haplotype is a relatively rare phenotype. In this setting, traditional approaches of variable selection in polygenic risk scores may not apply due to increased false positive rates if variants are selected based on p-values alone. Third, the SMA silent carrier allele has highly differential population frequencies, so questions regarding prediction transferability introduced in Chapter 3 are a concern here.

We used a multi-population meta-analysis approach to identify variants that were specific to SMA duplication events across 5 populations using WGS data from the 1kGP. We hypothesize that an appropriate multi-variant prediction model will perform better to identify SMA silent carriers than any single-variant predictors alone. Our method could enable population-wide silent carrier screening for WGS data, solving both methodological and pressing clinical challenges.

4.2 Methods

In this study, we estimated the sensitivity and specificity of single variants in the *SMN1/2* region to predict individuals with the *SMN1* duplication allele using cross-validation results from a multi-population meta-analysis. Further, we trained a multi-variant prediction model for the *SMN1* duplication allele by extending polygenic risk score techniques to the copy number variable setting.

4.2.1 1kGP Data Preprocessing

We will train and evaluate our model using WGS data from 1kGP samples from 5 population groups (AFR, AMR, EAS, EUR, and SAS) for (n = 2,504) individuals. Inclusion criteria for samples in the 1kGP dataset have been described previously.[19] For a subset of individuals, we had genotype information for parent-child trios (see below).

4.2.2 Cohort B

We also included samples from an internal Illumina dataset, referred to throughout as Cohort B. The majority of samples in the cohort were of European ancestry.

4.2.3 SMNCopyNumberCaller

We applied the SMNCopyNumberCaller tool on all samples in order to infer *SMN1* Copy Number for each individual.[19] Individuals with *SMN1* Copy Number > 2 (n = 532) were identified as having the *SMN1* duplication allele and were included as cases. For individuals in trios, we included parents in trios in which all three members have *SMN1* CN=2 (n = 1,747). Here, we assume that since the silent carrier (2+0) genotype frequency is rare in the population, most of the individuals with *SMN1* CN=2 will have the 1+1 genotype. For Cohort B in which no trio information was available, we included individuals with *SMN1* CN=2 as controls. Thus, our total analysis sample size was (n = 2,279).

4.2.4 Variant Calling

We extracted reads from the DRAGEN processed .bam files for the two regions corresponding to *SMN1* and *SMN2* in the hg38 reference genome separately for each population group and study. The two regions genomic coordinates in the hg38 reference genome are chr5:70864000-70970000 (*SMN1*) and chr5:69988913-70094555 (*SMN2*). We then aligned the *SMN1* and *SMN2* hg38-aligned reads to the *SMN1* region using bwa.[62] We called variants in the re-aligned file using Mutect2 in tumor-only mode.[7] While not analyzing tumor genomes, Mutect2 relaxes traditional germline variant calling assumptions, such as a ploidy of two and non-variable copy number, which are directly relevant to analyzing reads from the *SMN1/2* region. Preprocessing of the read data was based on the Mutect2 documentation.[34]

After calling variants, we include variants with vcf FLAGS PASS, germline, clustered_events, haplotype. Each of the germline , haplotype, and clustered_events flags are useful in the traditional somatic variant calling setting, but do not have much utility in our extension to the complexities of calling variants in the *SMN1/2* region. After filtering variants for each individual, we merged the variant calls by subject for each ancestry group. We removed variants with calls in fewer than 5% of samples within each ancestry group. Since our variant calling is agnostic to the total *SMN1* and *SMN2* copy number for each subject, we can not use the allele fraction or genotypes provided by Mutect2 to compute allele frequencies. Since we are using *SMN1* and *SMN2* copy number information in association testing, we chose to not use this information during variant calling.

4.2.5 Assessing Predictive Utility of *SMN1* associated variants

We used 5-fold cross validation to assess the predictive utility of the single variants identified by the multi-population meta-analysis and the multi-variant prediction model. For each cross-validated fold, we performed a multi-population meta-analysis for the *SMN1* duplication allele using binary variant calls (call/no call). We first conducted population specific association testing using ordinary least squares separately for each study. We then used the METAL meta-analysis formulas to transform the marginal variant effect sizes into multi-population effect sizes and the

associated standard errors and p-values.[116] We used the Bonferroni adjusted threshold based on the number of variants to identify variants significantly associated with the *SMN1* duplication allele within each fold. For single variants, we predicted *SMN1* duplication alleles based on the direction of the METAL multi-population effect size estimate. Predictive performance is assessed on cross-validated sensitivity, specificity, and positive predictive value using the population specific silent carrier frequency.[44] Additionally, we only considered variants that were significant in each of the five cross-validated folds. We compared our results to the standard of care variant, NM_000344.3:c.*3+80T>G (also referred to as g.27134T>G), which has been identified previously in the literature as correlated with the *SMN1* duplication allele.

4.2.6 Multi-variant prediction model using P+T Scoring.

Using the summary statistics from the meta-analysis, we fit a multi-variant prediction model similar to P+T scoring commonly used in the polygenic risk score literature.[118] Rather than selecting variants based on their meta-analysis p-values, we include variants in the multi-variant model based on a specificity threshold. Two different variant-level specificity measures were considered: cross-validated population specificity and the cross-validated minimum population-specific specificity. For the inner-loop cross validation, we trained our multi-variant models over a grid of correlation values (0.5, 0.7, 0.9) and variant-level specificity values (0.97, 0.98, 0.99, 0.995). To establish a cutoff to binarize the predictions of our multi-variant score, for each set of tuning parameters, we specified a false-positive rate for the score prediction of 0.5% and selected the score cutoff that maximized sensitivity. Commonly, ROC curves are used for this end in PRS research, and the area under the ROC curve is used as an optimizing value. In the case of silent carrier detection, we may value a low false positive rate over a high true positive rate, so we fixed our false positive rate when designing our model fitting procedure. This sort of argument is much more common when designing statistical tests for null-hypothesis significance testing. Finally, we applied the score threshold on the inner loop testing data, and computed the population sensitivity and specificity. The tuning parameters and associated thresholds chosen in the final model are those

which achieve the specificity threshold of 0.5% and maximize sensitivity. One final consideration: when applying our model on population WGS data, we may not have ancestry information for each sample. Therefore, we can use ancestry specific sensitivity and specificity when training the model, but choose to not use it when evaluating the model.

4.3 Results

4.3.1 Population Allele Frequencies Inform Modeling Choices

Below, we simulate ideal tests to illustrate our hypothesis that selecting variants agnostic to their predictive utility will fail at the task of SMA silent carrier prediction. We used population allele frequencies for silent carriers of SMA in order to estimate the positive predictive value of silent carrier detection tests as a function of sensitivity and specificity. Assuming a test with 80% specificity, tests with perfect sensitivity fail to achieve a positive predictive value above 2% for silent carriers in AFR individuals in which silent carriers are most prevalent (Figure 4.1). The remaining populations do achieve a positive predictive value above 0.5% even with a perfectly sensitive test. However, for tests with sensitivity of 80%, positive predictive values >50% were achieved when specificity was extremely close to 1 (Figure 4.2). For AFR individuals, a test with 80% sensitivity requires a specificity of 99.6% to achieve a PPV of 50%, and for EUR individuals a similar test requires 99.9% specificity. These calculations informed our future modeling choices to fix specificity at 99.5% when fitting our multi-variant model and then maximize sensitivity.

4.3.2 Single Variant Association Study Identifies *SMN1* Duplication Allele Specific Variants

We estimated the sensitivity and specificity of single variants in the *SMN1/2* region to predict individuals with the *SMN1* duplication allele using cross-validation results from a multi-population meta-analysis. For each fold, we estimated the marginal effect of each variant on the *SMN1* duplication allele. We used the fold-specific marginal effect sizes from the multi-population meta-analysis to predict *SMN1* duplication allele carrier status. We identified variants that were significant

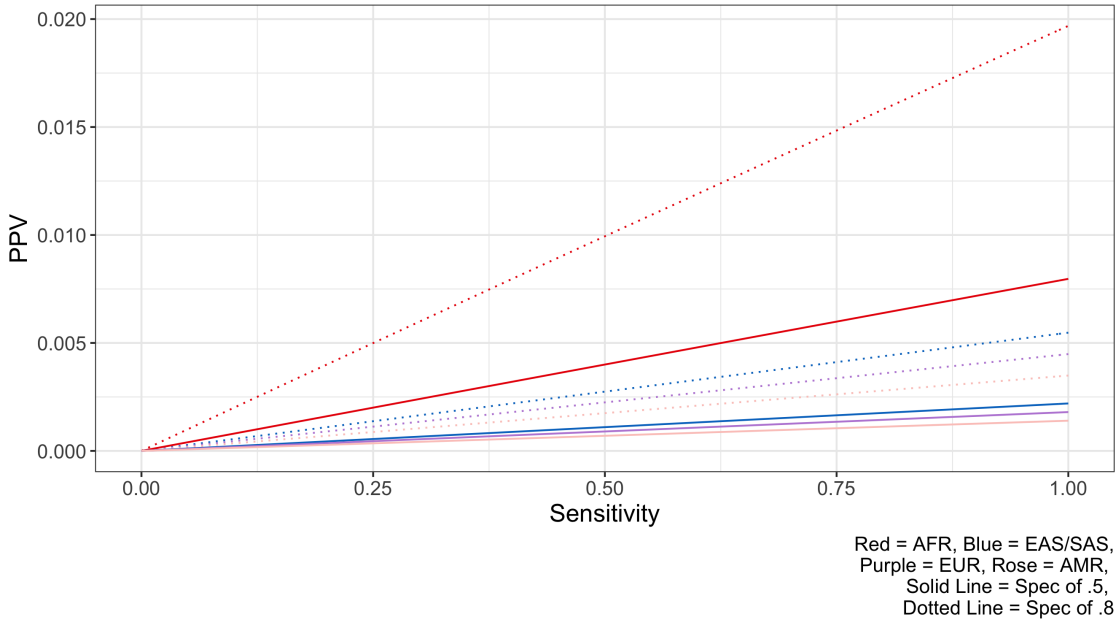

Figure 4.1: PPV as a Function of Sensitivity for SMA Silent Carrier Tests. Even with a perfectly sensitive test, PPV of tests are still <2% to detect SMA silent carriers in AFR population where prevalence of silent carriers is the highest. Different colors represent different assumptions of population frequencies of silent carriers. Dotted and dashed lines represent tests with different specificity values.

at the Bonferroni adjusted threshold in all five cross-validation folds, including the previously reported standard-of-care variant, g.27134T>G. Using the held out data for each fold, we estimated the single-variant sensitivity and specificity for each variant significant in all 5 folds using a binary prediction rule. Our 5-fold sensitivity and specificity estimates largely match the previously reported frequency estimates for g.27134T>G in carriers and non-carriers, confirming the utility of our variant calling and prediction procedure (Figures C.1 and C.2).[19] The cross-validated population sensitivity and specificity are misleading performance estimates, as they are heavily influenced by population sample size, as seen by the disparate estimates when stratified by population below. Further, the PPV of this variant is extremely low in all populations, most notably in African ancestry individuals (Figure C.3). Given a positive variant call for g.27134T>G, the cross-validated PPV estimate implies that there is less than a 2% chance of having the *SMN1* duplication allele, limiting clinical utility of the test.

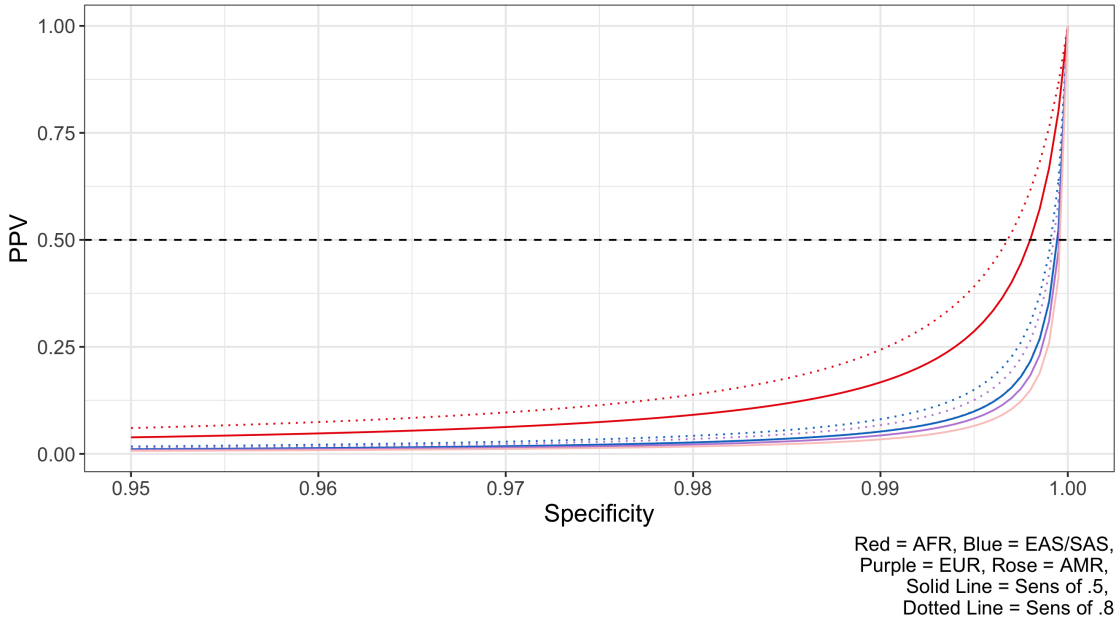

Figure 4.2: PPV as a Function of Specificity for SMA Silent Carrier Tests. PPV is an exponential function of specificity. A test to detect silent carriers much have extremely low false positive rate to yield clinically actionable test results. Different colors represent different assumptions of population frequencies of silent carriers. Dotted and dashed lines represent tests with different sensitivity values.

We identified 9 variants with cross-validated PPV greater than 25% in AFR individuals, 23 variants in AMR individuals, and 9 in SAS individuals. The majority of these variants are only found in carriers of a single population, and are thus only predictive in that population (see two examples in Figure C.4 and Figure C.5). Despite sample size, our results suggest that due to the rare nature of the silent carrier phenotype, the population specific allele frequency is an upper limit to the clinical utility of using single variants to identify individuals with the *SMN1* duplication allele using PCR based methods. These results motivate the use of a multi-variant prediction model, combining predictive single variants to increase predictive accuracy in population screening.

4.3.3 Multi-population PRS

We fit a multi-population multi-variant risk score for the *SMN1* duplication allele and assessed the procedure using 5-fold cross validation of our training data. The mean cross-validated false-

positive rate was <0.01%, and the mean cross-validated sensitivity was 49.6%. As a result of no false positives in our cross-validation results, the PPV for the *SMN1* duplication allele in AFR and AMR individuals was estimated to be 1. Notably, our multi-variant risk score achieves >30% sensitivity in three populations while single-variant predictions usually were only meaningfully sensitive in one population if they controlled the false positive rate.

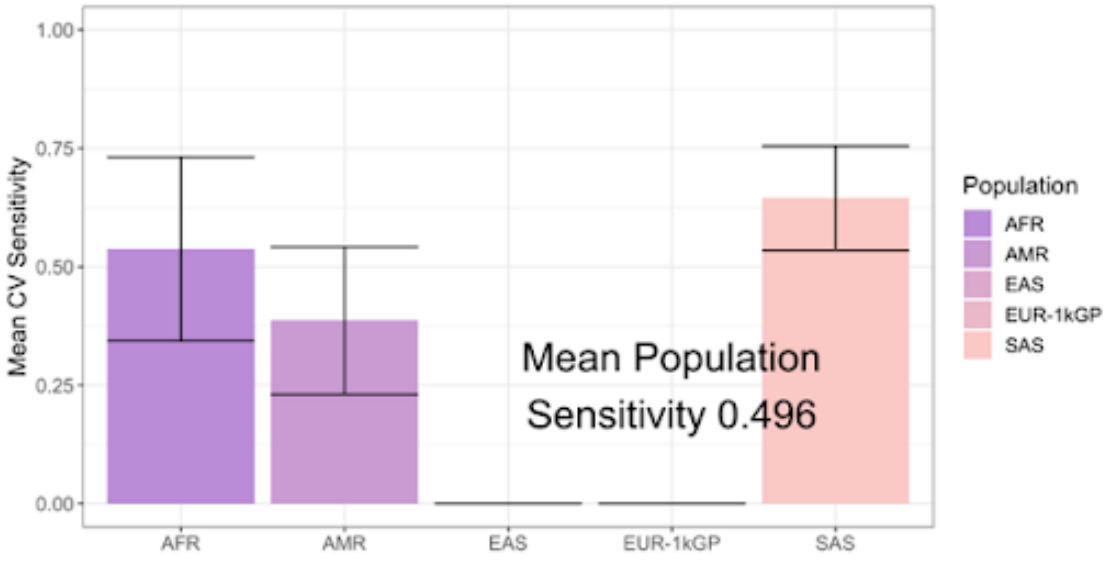

Figure 4.3: Cross-validated Sensitivity of Multi-variant Prediction Score for *SMN1* Duplication. We use a PRS-like approach to select variants with high specificity to predict the *SMN1* duplication allele. In 5-fold cross-validation experiments, our score out-performs the standard of care variant, g.27134T>G, in both mean population sensitivity while achieving no false positives in the cross-validation.

After assessing the performance of the multi-variant risk score via cross validation, we fit the same procedure on the full set of training data from 1kGP and the Internal Illumina dataset. Our final model consists of 27 variants. In the training dataset, we identified three individuals with *SMN1* CN=2 but who were predicted to have the *SMN1* duplication allele implying that these individuals may be silent carriers. In particular, one SAS subject has variant calls for 9 of the total 27 variants comprising the *SMN1* duplication allele risk score, 6 of which are nearly specific to SAS individuals

with the *SMN1* duplication allele. Additionally, the standard of care variant g.27134T>G was not called, likely due to its low frequency in South Asian individuals generally.[19, 44]

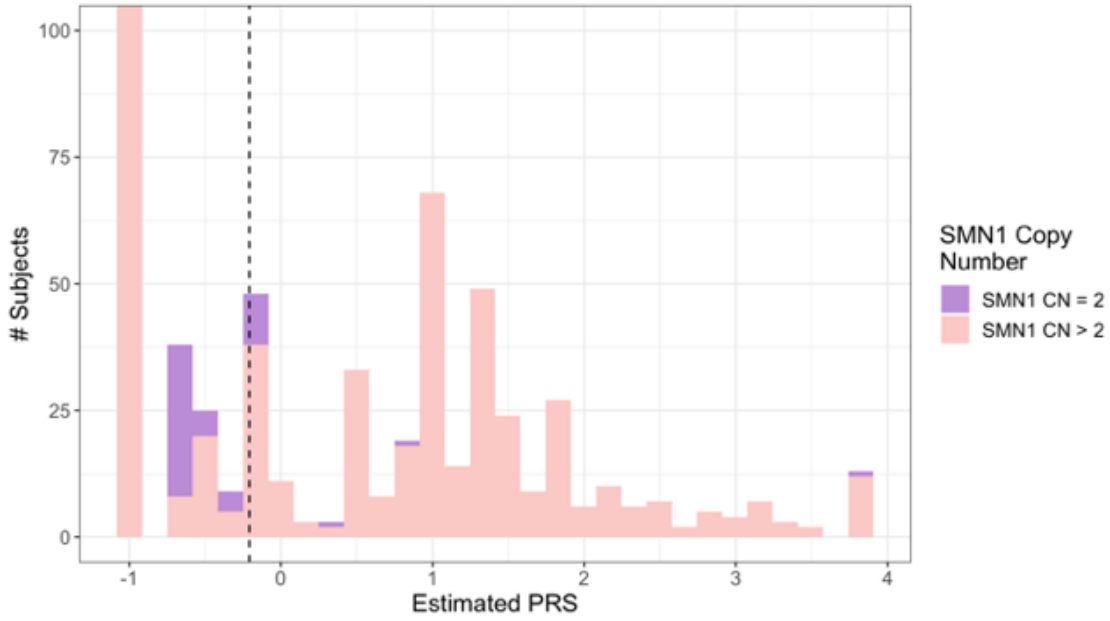

Figure 4.4: Estimated Distribution of *SMN1* Duplication Allele Risk. We applied our estimation procedure to the full training sample to assess the risk distribution for an *SMN1* duplication. In particular, we discover 3 individuals who have *SMN1* CN=2 and a positive prediction for an *SMN1* duplication allele, making them prime candidates for follow up as SMA silent carriers. In particular, one subject with *SMN1* CN=2 has one of the highest scores in the training sample for a duplication event.

4.4 Discussion

In this chapter, I presented our approach to identifying SMA silent carriers using multi-variant prediction models extended to the variable copy-number setting. We applied somatic variant calling approaches and multi-population meta-analyses to construct a predictor that controls the false positive rate while remaining sensitive to *SMN1* duplication. Importantly, this multi-variant approach is able to predict duplication rates across global populations. Since population allele frequencies of the *SMN1* copy number differ across population, it is essential that whole-genome sequencing-based methods that perform disease screening are effective across all populations that may be screened with the method.

One important innovation of our approach was the importance of considering measures other than variant p-values for inclusion in the score. We developed a variant selection procedure based primarily on population-specific specificity based on a binary prediction rule. We hypothesize that this approach works better than traditional p-value variable selection from the P+T PRS literature summarized in Chapter 1 because of the rare nature of the event. A future direction of this work would be to formally compare this variant selection procedure to others via simulation of other rare event phenotypes. This methodology could be useful for more polygenic and non-Mendelian diseases that are still rare binary events.

We hope to further this work through the application of the score derived in 1kGP data to other diverse cohorts with whole-genome sequencing data.

CHAPTER 5: CONCLUSION

In this dissertation, we present statistical research that attempts to address challenging biological problems unified by the theme of improving the clinical utility of omics data: inferring cell type proportions from bulk Hi-C data, incorporating mosaics of local ancestry in admixed individuals, and detection of silent carriers of a Mendelian disease. An important motif in this theme is the need to increase the ancestral diversity of participants in genetic studies. As participants in genetic studies become increasingly diverse, biostatisticians must adapt our tools to make sure that our inferences are still accurate and that we are answering useful questions. At the same time, we must be advocates for increasing diversity in genetic studies, which means working with other public health professionals to address the many barriers to individuals of non-European ancestry to participating in genetic research. With these charges in mind, we here discuss conclusions and future directions for the work presented in the previous chapters.

In chapter 2, we presented THUNDER, which solves a practical problem in the analysis of Hi-C data, namely, confounding due to variation in underlying cell type proportions. With inferred cell type proportion estimates from THUNDER, a genome-wide scan for variants associated with the 3D-interactome is now possible without confounding as Hi-C data on more individuals is generated. Additionally, THUNDER estimated cell type specific profiles can be used as cell-type-specific functional annotations to increase the interpretability of GWAS loci while single-cell data on many cell types are unavailable. However, datasets where genotype and Hi-C data are paired for individuals are extremely rare. Such paired datasets for other forms of omics data exist, but mostly in European ancestry individuals.[6, 41, 55] As we have noted in other chapters, research disparities in genetics have public health consequences. Since these pitfalls have been reported widely by

other disciplines in genetics, it is imperative that future studies of chromatin interactions include individuals of diverse ancestry.[76, 55] There is a ripe opportunity to establish a shared data resource to examine the impact of genetic variation on spatial chromatin interaction and gene regulation across ancestral populations.[63] Additionally, the functional annotations derived from the data gathered by such a data resource would be a boon to moving from associations to causal variants and genes in ancestrally diverse populations.

In chapter 3, we presented GAUDI, a polygenic risk score estimation method which incorporates local ancestry to improve prediction in admixed individuals. We must take seriously the health disparities in the field of statistical genetics before we are swept away by promises of precision medicine; with current analyses and methods, the promises will only be actualized for some, not for all. As representation of admixed individuals increases in population genetics research, better statistical tools must be developed to account for the mosaic-like genomes of these individuals. One consequence of the increase in admixed individuals in genetic research is the potential need to re-think assumptions regarding the use of LD reference panels in the use of PRS methods relying on GWAS summary statistics. While it is a common assumption that it is essential to match an LD reference panel as closely as possible to the ancestry of the individuals in your study, there is still no consensus as to the impact of individual-level admixture on LD (see Discussion Chapter 4). If current approaches are left inadequate, it may be most appropriate to pivot to analyses using individual-level data until an equitable solution can be achieved. Several challenges must be overcome if this direction were to be chosen: data privacy agreements must be established and respected for research participants, computationally efficient methods must be developed, and data storage must not be cost-prohibitive for a global community of researchers. Through each iteration of PRS methods research, collaboration with public health practitioners to ensure the accurate interpretation and application of PRS in ancestrally diverse populations is paramount.[2]

In Chapter 4, we presented a polygenic risk score-like approach to a non-polygenic genetic disorder, spinal muscular atrophy. We saw the boon of using multiple alleles to predict the *SMN1* duplication allele in the presence of copy number variation in a highly complex genomic region.

Through close collaboration with genetic counselors, statistical geneticists, and physicians, there is much fruitful work to be done in undiagnosed genetic disorders, early disease detection, and population screening. In line with the motif introduced earlier this chapter, our research must be done equitably across populations especially if the screening tool is to be deployed publicly with no access to self-reported race-ethnicity data or inferred ancestry. Even if that information were available, many individuals with recent population admixture enrolling in genetic studies require that WGS screening tools be effective across populations. For an effective population screening method for SMA silent carrier status, it was imperative that we trained and tested our method on ancestrally diverse samples since *SMN1* and *SMN2* copy number distributions vary by global ancestry. Additionally, the choice of metrics to evaluate the performance of our method in a population-specific manner was essential to ensuring biases were addressed in the development of our approach. We hope that our method development continues to improve identification of SMA silent carriers, as well as inspiring future application in WGS-based population screening.

APPENDIX A: ADDITIONAL RESULTS FOR CHAPTER 2

A.1 Feature Selection Details

Let $W_1(i,j)$ denote the element in the i^{th} row and j^{th} column of the cell-type specific profile matrix W_1. Let S_{intra} denote the set of intrachromosomal bin-pairs. The derivation below is for intrachromosomal bin-pairs, but the feature selection algorithm is the same for interchromosomal bin-pairs.

Standard deviation across cell types for bin-pair i is defined as,

$$SD_i = \frac{1}{k-1}\sum_{j=1}^{k}(W_1(i,j) - \frac{1}{k}W_1(i,\cdot))^2$$

Feature score across cell types for bin pair i is defined as follows,

$$FS_i = 1 + \frac{1}{log_2(k)}\sum_{j=1}^{k} p(i,j)log_2(p(i,j))$$

where $p(i,\Omega)$ is the probability that the i^{th} pairwise bin contributes to cell type Ω, i.e.,

$$p(i,\Omega) = \frac{W_1(i,\Omega)}{\sum_{j=1}^{k} W_1(i,j)}$$

Feature scores range from $[0,1]$ with higher scores representing bin-pairs with higher cell-type specificity. We further define,

$$\hat{\mu}_{SD,intra} = \frac{1}{|S_{intra}|}\sum_{i:i\in S_{intra}} SD_i$$

$$\hat{\mu}_{FS,intra} = \frac{1}{|S_{intra}|}\sum_{i:i\in S_{intra}} FS_i$$

$$\hat{\sigma}_{SD,intra} = \frac{1}{|S_{intra}|-1}\sum_{i:i\in S_{intra}} (SD_i - \hat{\mu}_{SD,intra})^2$$

$$\hat{\sigma}_{FS,intra} = \frac{1}{|S_{intra}| - 1} \sum_{i:i \in S_{intra}} (FS_i - \hat{\mu}_{FS,intra})^2$$

$$\hat{m}_{SD,intra} = \text{median}_{\{i \in S_{intra}\}}(SD_i)$$

$$\hat{m}_{FS,intra} = \text{median}_{\{i \in S_{intra}\}}(FS_i)$$

$$\hat{s}_{SD,intra} = \text{median}_{\{i \in S_{intra}\}}(\hat{m}_{SD,intra} - SD_i)$$

$$\hat{s}_{FS,intra} = \text{median}_{\{i \in S_{intra}\}}(\hat{m}_{FS,intra} - FS_i)$$

Let $FF_i = \frac{\mu_i}{\sigma_i}$ be the Fano Factor for the i^{th} row of the mixture matrix. Let e_i denote the row-wise matximum for the i^{th} row of the cell type profile matrix. Let $\hat{m}_{CTP}$ denote the median value for all elements of the cell-type profile matrix, W_1.

Table A.1 describes all methods tested in the THUNDER feature selection simulations. The *intra* subscript is dropped for clarity.

A.2 Supplemental Tables

Feature Selection Method	Mathematical Definition
CTS or ACV	$FS_i > \hat{\mu}_{FS} + 3\hat{\sigma}_{FS}$ OR $SD_i > \hat{\mu}_{SD} + 3\hat{\sigma}_{SD}$
CTS and ACV	$FS_i > \hat{\mu}_{FS} + 3\hat{\sigma}_{FS}$ AND $SD_i > \hat{\mu}_{SD} + 3\hat{\sigma}_{SD}$
CTS or ACV - Median	$FS_i > \hat{m}_{FS} + 3\hat{s}_{FS}$ OR $SD_i > \hat{m}_{SD} + 3\hat{s}_{SD}$
CTS and ACV - Median	$FS_i > \hat{m}_{FS} + 3\hat{s}_{FS}$ AND $SD_i > \hat{m}_{SD} + 3\hat{s}_{SD}$
CTS	$FS_i > \hat{\mu}_{FS} + 3\hat{\sigma}_{FS}$
ACV	$SD_i > \hat{\mu}_{SD} + 3\hat{\sigma}_{SD}$
CTS - Median	$FS_i > \hat{m}_{FS} + 3\hat{s}_{FS}$
ACV - Median	$SD_i > \hat{m}_{SD} + 3\hat{s}_{SD}$
Top 1000 FF	Select top 1000 rows based on FF_i
Top 100 FF	Select top 100 rows based on FF_i
Kim-Park	$FS_i > \hat{\mu}_{FS} + 3\hat{\sigma}_{FS}$ AND $e_i > \hat{m}_{CTP}$

Table A.1: Defining Feature Selection Methods for THUNDER simulations

Table A.2: Mixing proportions for GM12878, HAP1, and HeLa mixtures.

Sample Number	GM12878	HAP1	HeLa
1	1.00	0.00	0.00
2	0.00	1.00	0.00
3	0.00	0.00	1.00
4	0.70	0.25	0.05
5	0.70	0.05	0.25
6	0.05	0.70	0.25
7	0.05	0.25	0.70
8	0.25	0.05	0.70
9	0.25	0.70	0.05
10	0.45	0.45	0.10
11	0.45	0.10	0.45
12	0.10	0.45	0.45

Table A.3: Mixing proportions for Lee *et al.* Mixtures.

Sample Number	ODC	Astro	MG	Endo	OPC	NeuronMix
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	1	0
6	0	0	0	0	0	1
7	0.1	0.1	0.1	0.1	0.1	0.1
8	0.1	0.49	0.1	0.1	0.1	0.1
9	0.1	0.1	0.49	0.1	0.1	0.1
10	0.1	0.1	0.1	0.49	0.1	0.1
11	0.1	0.1	0.1	0.1	0.491	0.1
12	0.1	0.1	0.1	0.1	0.1	0.49
13	0	0.2	0.2	0.2	0.2	0.2
14	0.	0	0.2	0.2	0.2	0.2
15	0.2	0.2	0	0.2	0.2	0.2
16	0.2	0.2	0.2	0	0.2	0.2
17	0.2	0.2	0.2	0.2	0	0.2
18	0.2	0.2	0.2	0.2	0.2	0

Table A.4: Computational Performance on 3 YRI Samples of 10Kb Resolution Hi-C Data.

Chr	Duration (h)	Memory (GB)	Bin-Pairs	Selected Bin-Pairs
THUNDER Step 1				
chr1	4.7	85.1	2,898,116	57,611
chr2	3.7	96.7	3,211,342	63,152
chr3	4.5	70.3	2,742,971	54,380
chr4	4.4	73.1	2,653,508	52,331
chr5	4.0	73.1	2,457,251	48,239
chr6	3.8	54.6	2,334,790	46,153
chr7	2.5	52.1	1,972,510	39,288
chr8	2.3	51.9	1,926,609	38,021
chr9	1.5	44.0	1,443,866	28,720
chr10	3.2	48.5	1,681,822	33,231
chr11	3.7	51.5	1,727,598	34,530
chr12	3.6	51.7	1,752,414	35,044
chr13	2.5	37.6	1,364,537	26,453
chr14	1.3	41.2	1,170,924	23,367
chr15	1.0	31.0	996,396	20,132
chr16	1.5	29.3	832,338	16,141
chr17	1.8	29.6	841,985	17,422
chr18	0.9	32.2	1,025,512	20,399
chr19	0.9	22.1	468,526	9,114
chr20	0.9	26.4	730,999	14,688
chr21	0.7	22.8	441,879	8,468
chr22	0.3	19.5	339,585	6,887
THUNDER Step 2				
Genome-Wide	0.7	16.2	693,771	NA

Table A.5: Computational Performance on 5 YRI Samples of 10Kb Resolution Hi-C Data.

Chr	Duration (h)	Memory (GB)	Bin-Pairs	Selected Bin-Pairs
THUNDER Step 1				
chr1	5.4	97.3	3,174,333	65,265
chr2	7.2	103.2	3,507,051	68,506
chr3	6.3	92.4	2,988,427	61,700
chr4	5.1	89.1	2,893,327	59,152
chr5	4.0	85.7	2,677,352	53,739
chr6	6.9	81.2	2,545,972	50,167
chr7	6.1	69.6	2,163,857	44,196
chr8	3.3	59.1	2,105,779	41,971
chr9	2.5	49.0	1,580,008	32,001
chr10	3.2	59.4	1,848,124	36,801
chr11	3.1	57.5	1,896,637	37,615
chr12	2.9	62.2	1,931,846	38,602
chr13	2.3	49.2	1,483,383	29,353
chr14	1.9	45.3	1,278,870	25,480
chr15	2.1	35.7	1,094,740	22,468
chr16	2.4	34.5	924,458	17,667
chr17	3.8	37.3	935,763	19,203
chr18	1.5	42.7	1,120,797	21,741
chr19	1.5	23.5	524,543	9,540
chr20	1.3	30.0	802,176	15,743
chr21	1.1	25.2	485,244	9,076
chr22	0.6	21.1	380,611	7,713
THUNDER Step 2				
Genome-Wide	2.5	18.2	767,700	NA

Table A.6: Computational Performance on 10 YRI Samples of 10Kb Resolution Hi-C Data.

Chr	Duration (h)	Memory (GB)	Bin-Pairs	Selected Bin-Pairs
THUNDER Step 1				
chr1	12.6	110.5	3,538,635	68,709
chr2	15.4	160.2	3,903,593	78,890
chr3	20.3	103.8	3,306,003	68,256
chr4	17.4	106.6	3,201,164	65,000
chr5	11.0	114.4	2,967,778	58,602
chr6	9.8	86.9	2,818,330	56,838
chr7	12.9	93.5	2,426,518	50,105
chr8	11.8	89.6	2,343,805	46,794
chr9	6.6	73.7	1,761,728	34,939
chr10	11.4	80.4	2,070,850	41,651
chr11	9.5	85.1	2,117,742	42,870
chr12	7.6	84.6	2,129,312	41,521
chr13	12.1	63.5	1,638,718	31,997
chr14	5.2	62.1	1,431,155	28,821
chr15	6.3	53.3	1,234,104	24,959
chr16	5.1	47.6	1,060,098	20,798
chr17	3.7	43.9	1,075,081	21,785
chr18	4.8	54.0	1,245,372	25,415
chr19	3.6	27.9	616,312	10,771
chr20	4.7	39.9	900,160	18,390
chr21	3.8	28.1	545,011	10,082
chr22	1.5	27.1	440,207	8,700
THUNDER Step 2				
Genome-Wide	4.1	28.6	855,893	NA

A.3 Supplemental Figures

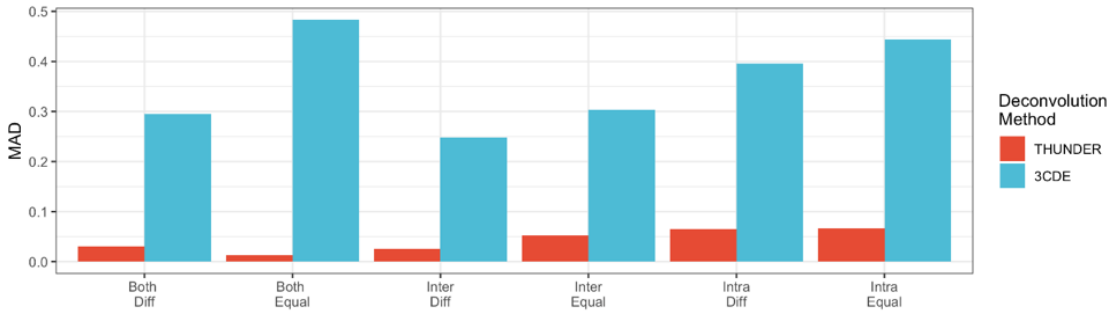

Figure A.1: Performance of THUNDER and 3CDE on HAP1 and HeLa Simulated Mixtures. We see that in several simulations, 3CDE achieves near the maximum mean absolute deviation from true cell type proportions (0.5).We do not test 3CDE in further simulations because of its inability to handle multiple Hi-C samples simultaneously.

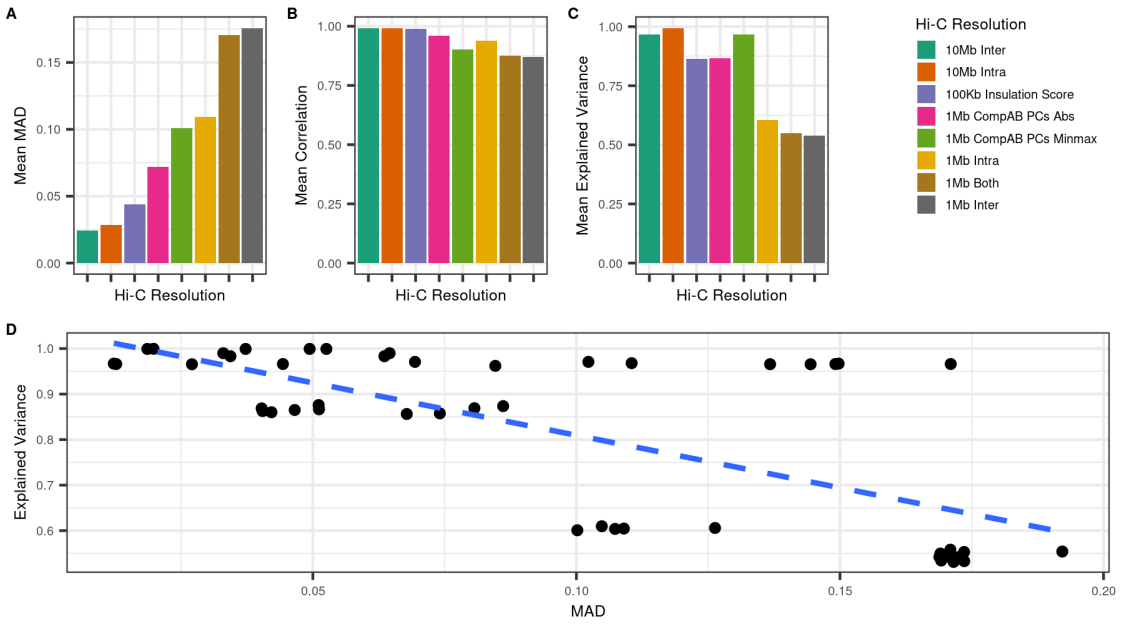

Figure A.2: Simulation Results Supporting the THUNDER Procedure of Choosing k. (a-c) We see that the relative rankings of the different Hi-C data resolutions are nearly the same for all three measurements. Since explained variance is independent of knowing the true cell type proportions, we propose it as a method of evaluating goodness of fit for THUNDER. (d) Explained variance is negatively correlated with MAD across all simulations. Blue dotted line is the linear regression line through the points.

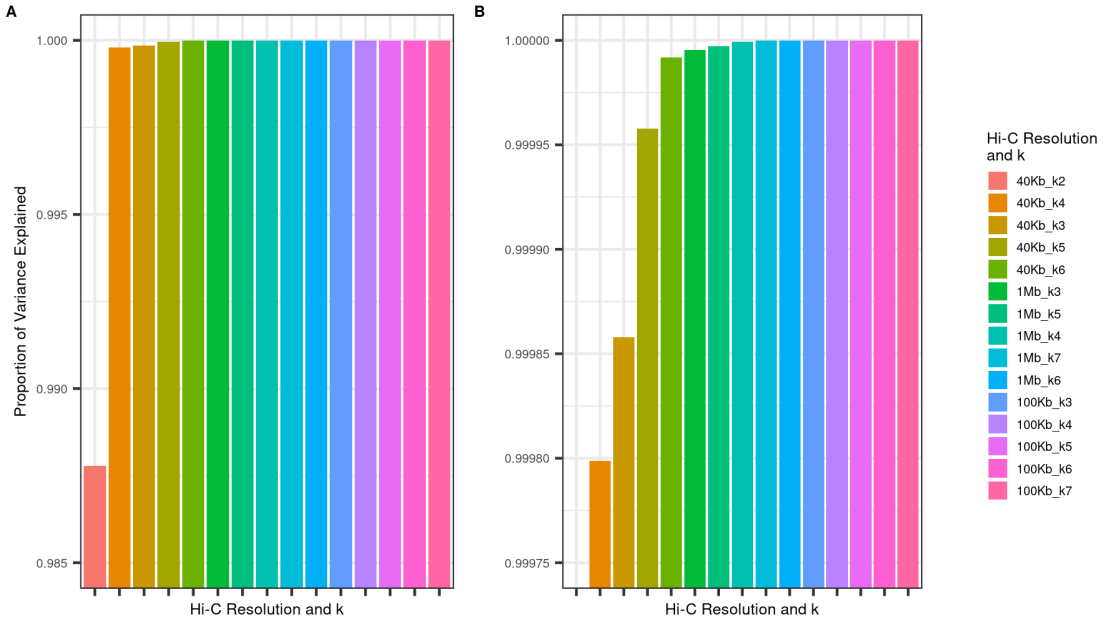

Figure A.3: Explained Variance Estimates for a range of k and Hi-C data resolutions for deconvolution of Giusti-Rodriguez et al. Hi-C data. (a-b)Proportion of variance explained for all combinations of k and Hi-C data resolution tested for deconvolution of the Giusti *et al.* dataset. (b) is zoomed in on the y-axis. Since many features equal 1, we choose the resolution with the largest bin-pair resolution to match our simulation results.

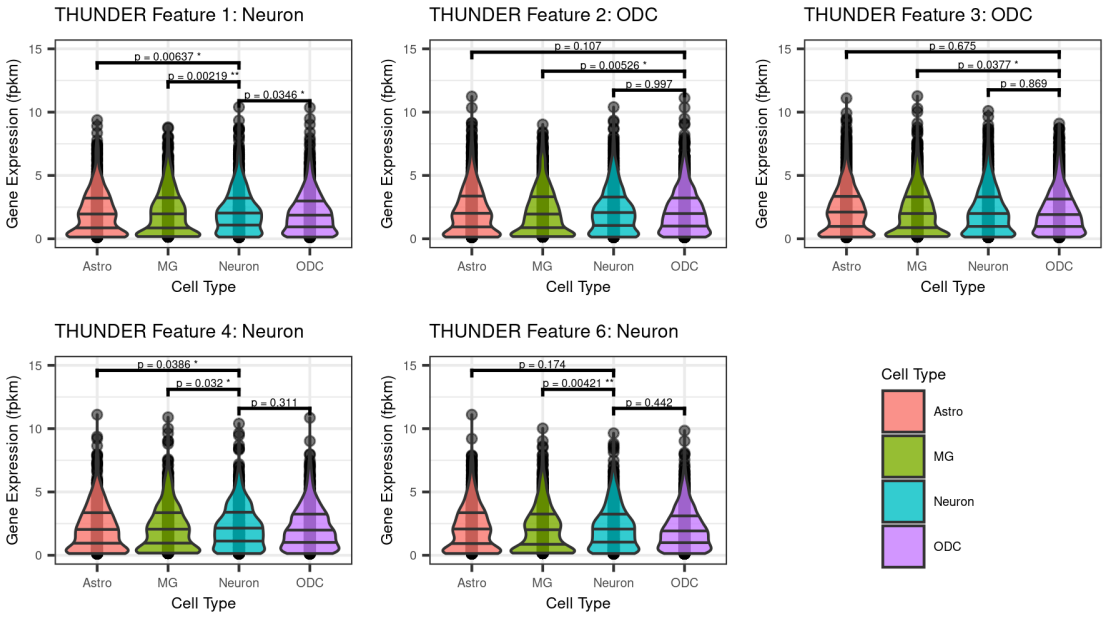

Figure A.4: Gene Expression Enrichment Tests in THUNDER bins. We tested for gene expression enrichment in bins identified by THUNDER to be specific to cell types after deconvolution. For each THUNDER feature, we compared the distributions of gene expression of cell type specifically expressed genes contained in the THUNDER bins.

APPENDIX B: ADDITIONAL RESULTS FOR CHAPTER 3

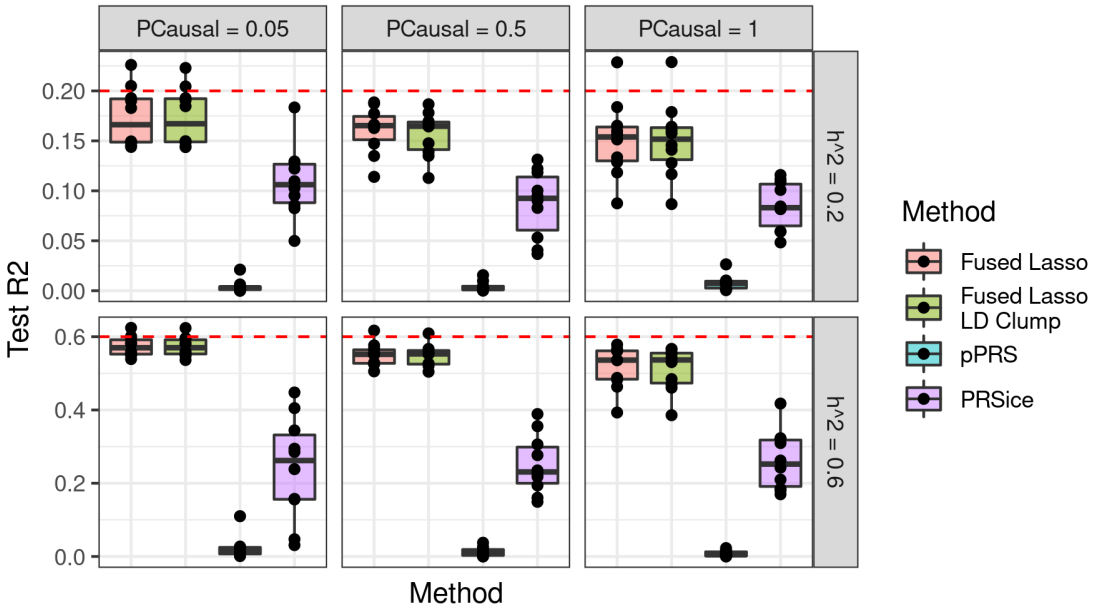

Figure B.1: GAUDI Simulation Results - Specific Traits, Phenotype 2. Evaluation of PRS methods on COSI simulated admixed genotypes and phenotypes where 50% of causal effects are shared across ancestral populations and 50% are specific to the ancestral populations. All causal SNPs have MAF $\geqslant 0.05$ in the EUR reference data and MAF < 0.05 in the AFR reference data.

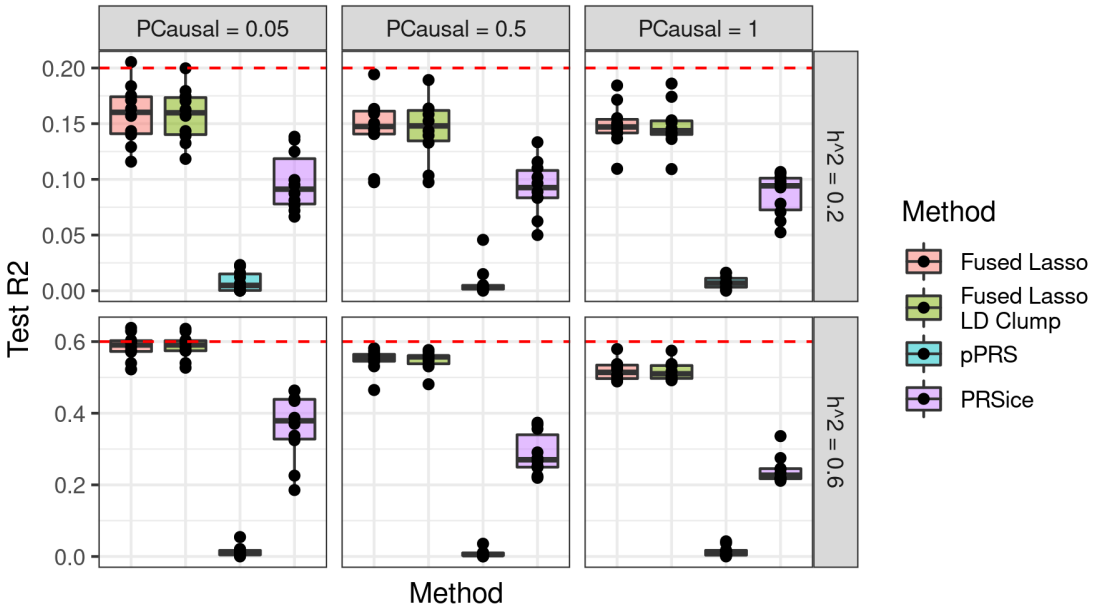

Figure B.2: GAUDI Simulation Results - Specific Traits, Phenotype 1. Evaluation of PRS methods on COSI simulated admixed genotypes and phenotypes where 50% of causal effects are shared across ancestral populations and 50% are specific to the ancestral populations. All causal SNPs have MAF $\geqslant 0.05$ in the EUR reference data and MAF $\geqslant 0.05$ in the AFR reference data.

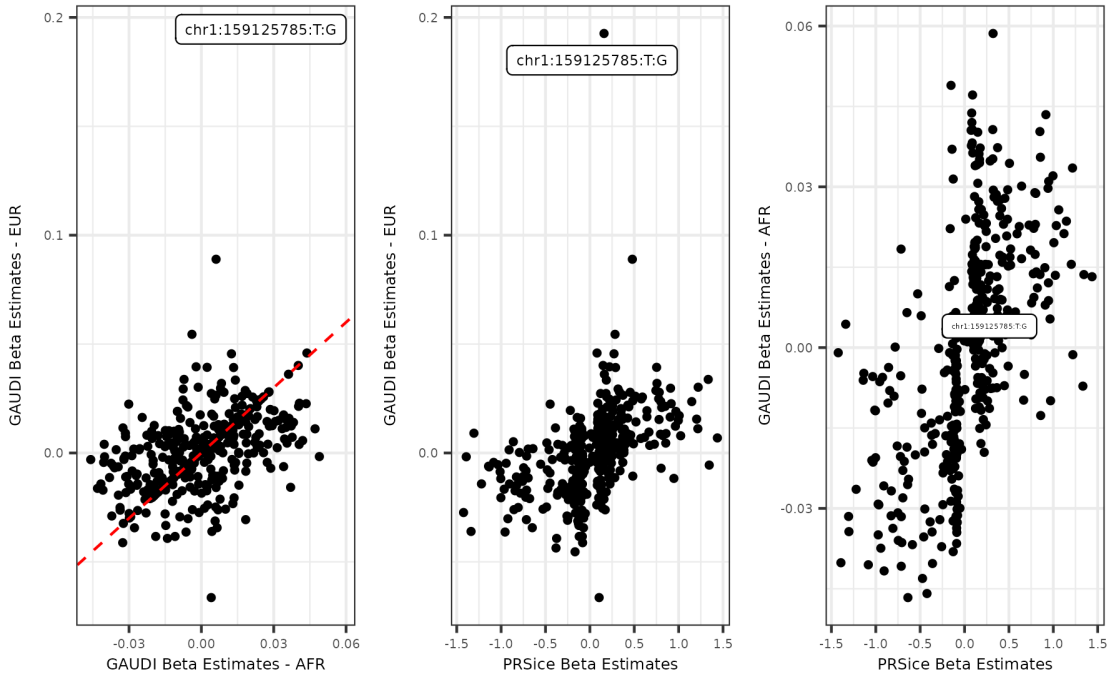

Figure B.3: Differential Effect Size at LD Proxy for Duffy Null Variant. (a) For GAUDI effect sizes on CA variants, we observe that many effect sizes are jointly estimated to be similar for AFR and EUR effect sizes. However, one variant, chr1:159125785:T:G, has a large EUR effect size estimate $\beta_{EUR} = 0.193$ and a small AFR effect size estimate $\beta_{AFR} = 4.71 * 10^{-4}$. Red dotted line is the 45 degree line. (b-c) When comparing GAUDI and PRSice effects on the same variants, the primary difference between the two scores is at chr1:159125785:T:G. There are not many sign disagreements between the two scores, suggesting that incorporating local ancestry estimates does not necessarily change the direction of interpreted effect, but more accurately refines the estimated effect size for a variant.

APPENDIX C: ADDITIONAL RESULTS FOR CHAPTER 4

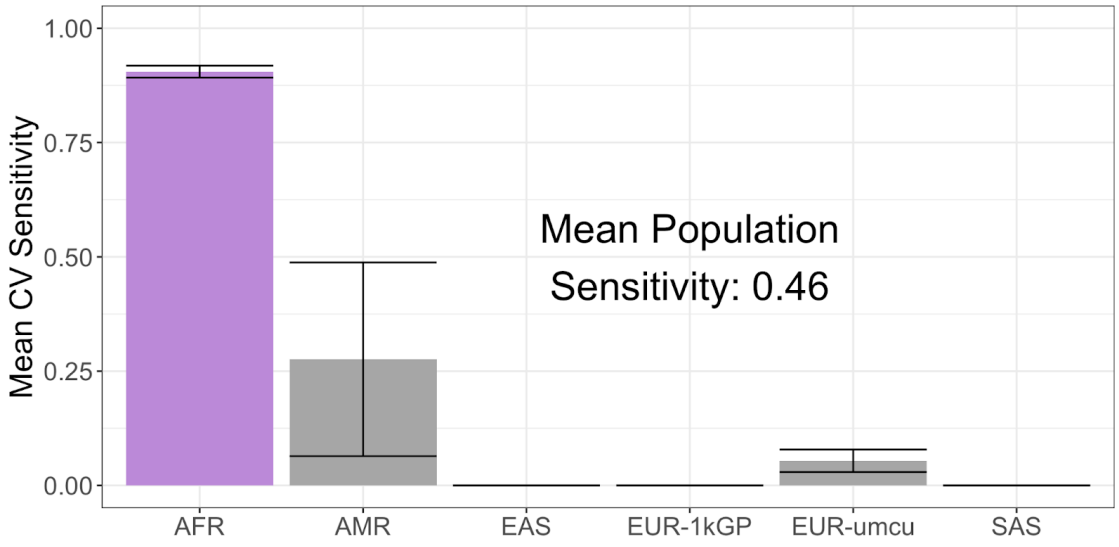

Figure C.1: Standard of Care Variant - CV Sensitivity. Population-specific sensitivity cross-validated estimates for the standard of care variant to identify silent carriers, g.27134T>G. This measure corresponds to the cross-validated estimate of allele frequency in *SMN1* duplication allele carriers.

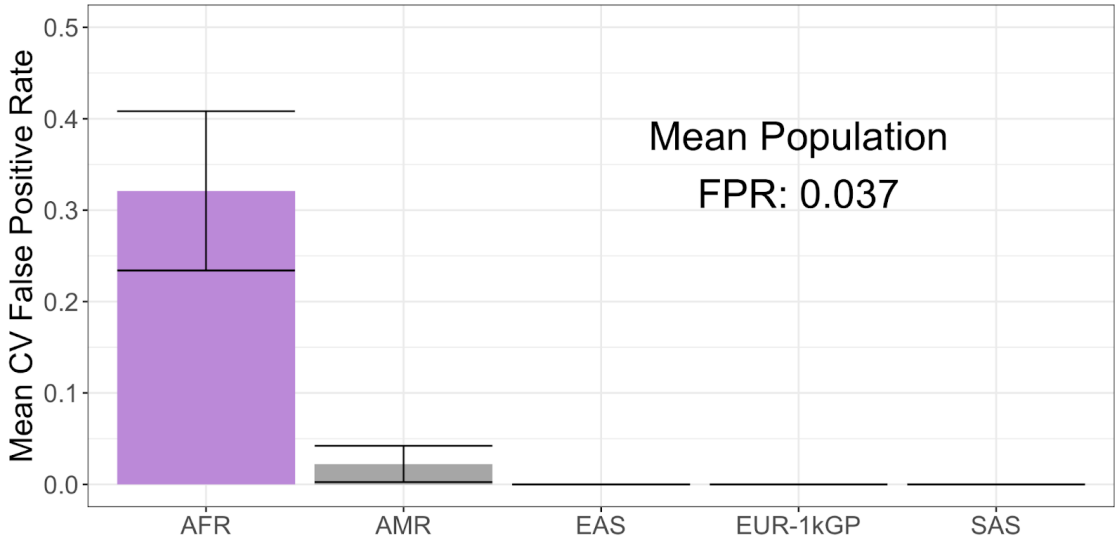

Figure C.2: Standard of Care Variant - CV Specificity. Population-specific false positive rate cross-validated estimates for the standard of care variant to identify silent carriers, g.27134T>G. The standard of care variant is present in many individuals with no *SMN1* duplication allele, especially in AFR individuals.

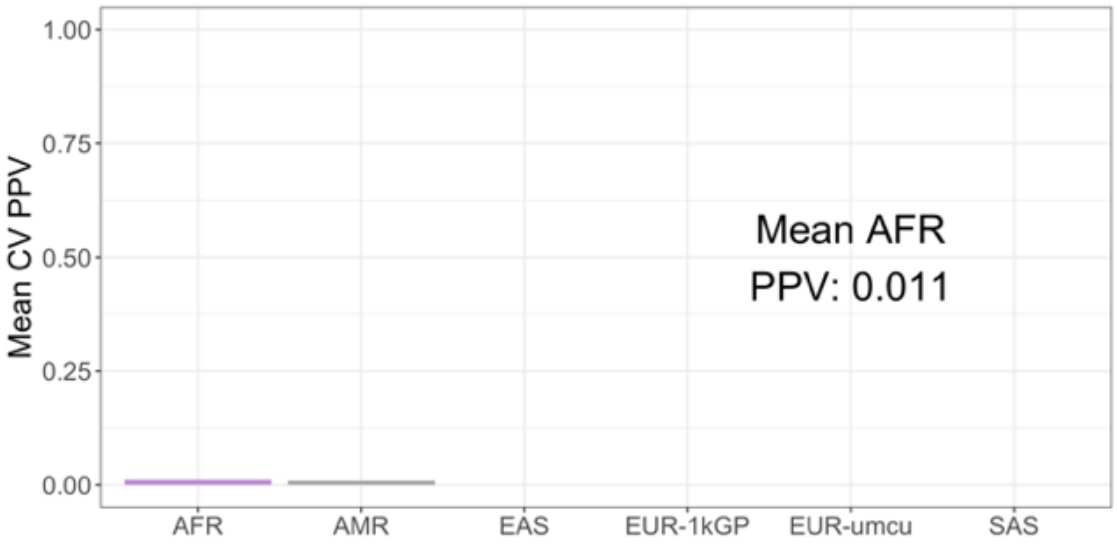

Figure C.3: Standard of Care Variant - CV PPV. Population-specific positive predictive value (PPV) cross-validated estimates for the standard of care variant to identify silent carriers, g.27134T>G. As a result of the rare prevalence of the *SMN1* duplication allele, as well as the high false positive rates in AFR, the positive predictive value of the associated test to detect *SMN1* duplication is extremely low. See Figures 4.1 and 4.2 for this calculation.

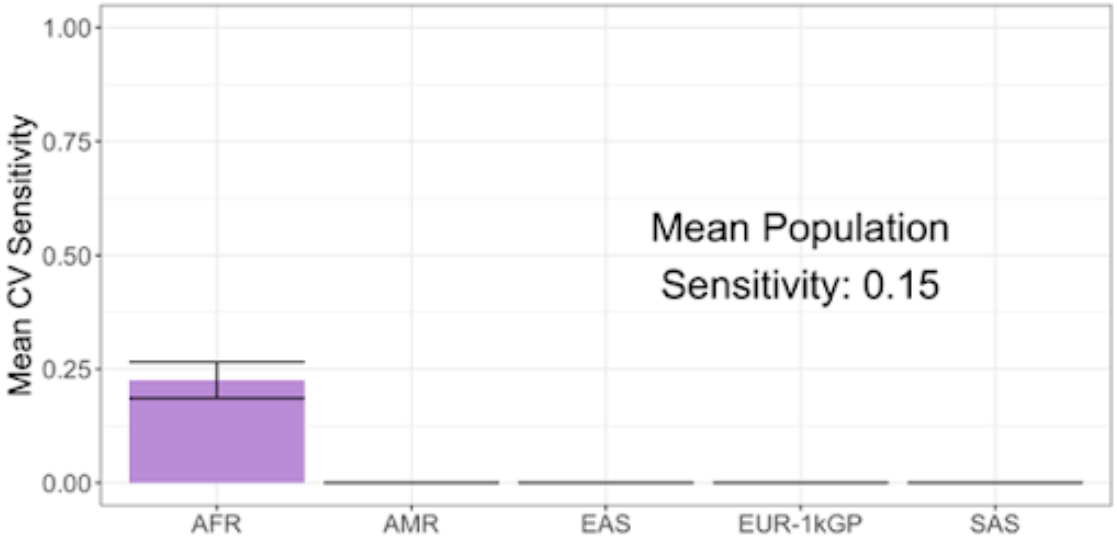

Figure C.4: High PPV AFR Variant Discovered - Variant A. Population-specific sensitivity cross-validated estimate for Variant A, which has a controlled false positive rate in all populations. This variant is present in 25% of AFR individuals with the *SMN1* duplication allele.

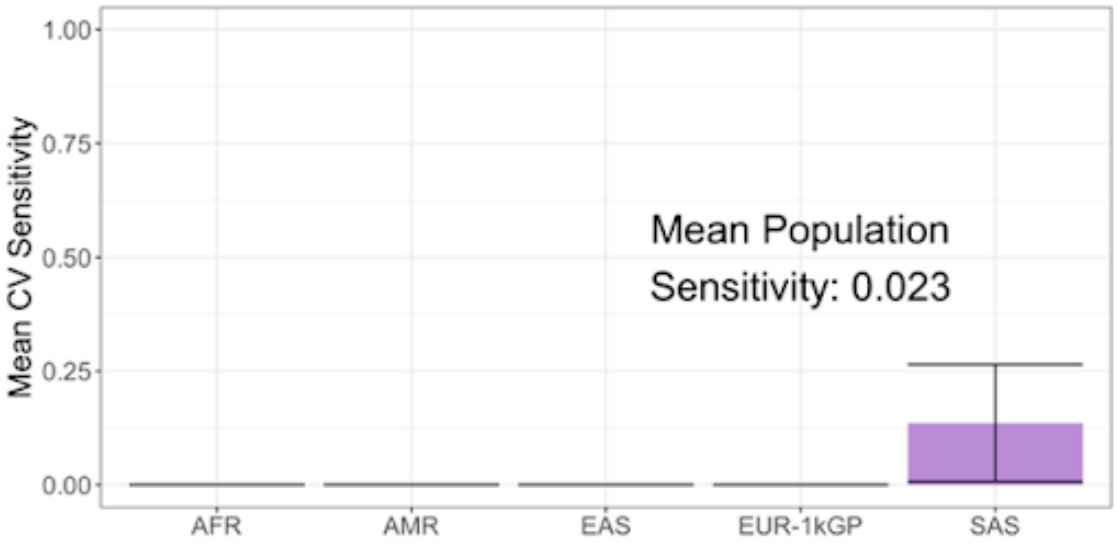

Figure C.5: High PPV SAS Variant Discovered - Variant B. Population-specific sensitivity cross-validated estimate for Variant A, which has a controlled false positive rate in all populations. This variant is present in 25% of AFR individuals with the *SMN1* duplication allele. Together with Figure C.4, these variants suggest a multi-variant prediction model would be a helpful solution to the prediction task.

REFERENCES

[1] K. C. Akdemir, V. T. Le, S. Chandran, Y. Li, R. G. Verhaak, R. Beroukhim, P. J. Campbell, L. Chin, J. R. Dixon, P. A. Futreal, PCAWG Structural Variation Working Group, and PCAWG Consortium. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.*, 52(3):294–305, Mar. 2020.

[2] I. C. D. Alliance, A. Adeyemo, M. K. Balaconis, and others. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nature*, 2021.

[3] T. Amariuta, K. Ishigaki, H. Sugishita, T. Ohta, M. Koido, K. K. Dey, K. Matsuda, Y. Murakami, A. L. Price, E. Kawakami, C. Terao, and S. Raychaudhuri. Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nat. Genet.*, 2020.

[4] G. L. Anderson, J. Manson, R. Wallace, B. Lund, D. Hall, S. Davis, S. Shumaker, C.-Y. Wang, E. Stein, and R. L. Prentice. Implementation of the women's health initiative study design. *Ann. Epidemiol.*, 13(9 Suppl):S5–17, Oct. 2003.

[5] E. G. Atkinson, A. X. Maihofer, M. Kanai, A. R. Martin, K. J. Karczewski, M. L. Santoro, J. C. Ulirsch, Y. Kamatani, Y. Okada, H. K. Finucane, K. C. Koenen, C. M. Nievergelt, M. J. Daly, and B. M. Neale. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat. Genet.*, 53(2):195–204, Feb. 2021.

[6] A. Battle, S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, C. McCormick, C. D. Haudenschild, K. B. Beckman, J. Shi, R. Mei, A. E. Urban, S. B. Montgomery, D. F. Levinson, and D. Koller. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.*, 24(1):14–24, Jan. 2014.

[7] D. Benjamin, T. Sato, K. Cibulskis, G. Getz, C. Stewart, and L. Lichtenstein. Calling somatic SNVs and indels with mutect2. Dec. 2019.

[8] K. Bibbins-Domingo, D. C. Grossman, and S. J. Curry. The US preventive services task force 2017 draft recommendation statement on screening for prostate cancer: An invitation to review and comment. *JAMA*, 317(19):1949–1950, May 2017.

[9] S. A. Bien, G. L. Wojcik, N. Zubair, C. R. Gignoux, A. R. Martin, J. M. Kocarnik, L. W. Martin, S. Buyske, J. Haessler, R. W. Walker, I. Cheng, M. Graff, L. Xia, N. Franceschini, T. Matise, R. James, L. Hindorff, L. Le Marchand, K. E. North, C. A. Haiman, U. Peters, R. J. F. Loos, C. L. Kooperberg, C. D. Bustamante, E. E. Kenny, C. S. Carlson, and PAGE Study. Strategies for enriching variant coverage in candidate disease loci on a multiethnic genotyping array. *PLoS One*, 11(12):e0167758, Dec. 2016.

[10] B. D. Bitarello and I. Mathieson. Polygenic scores for height in admixed populations. *G3*, 10(11):4027–4036, Nov. 2020.

[11] E. E. Blue, A. R. V. R. Horimoto, S. Mukherjee, E. M. Wijsman, and T. A. Thornton. Local ancestry at APOE modifies alzheimer's disease risk in caribbean hispanics. *Alzheimers. Dement.*, 2019.

[12] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.*, 2004.

[13] A. Buniello, J. A. L. Macarthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, 2019.

[14] M. Cai, J. Xiao, S. Zhang, X. Wan, H. Zhao, G. Chen, and C. Yang. A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *Am. J. Hum. Genet.*, 108(4):632–655, Apr. 2021.

[15] S. Carstens, M. Nilges, and M. Habeck. Inferential structure determination of chromosomes from Single-Cell Hi-C data. *PLoS Comput. Biol.*, 2016.

[16] N. Chatterjee, J. Shi, and M. García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.*, 17(7):392–406, July 2016.

[17] M.-H. Chen, L. M. Raffield, A. Mousas, B.-C. C. (bcx2), A. D. Johnson, A. P. Reiner, P. Auer, and G. Lettre. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *bioRxiv*, 2020.

[18] T.-H. Chen, N. Chatterjee, M. T. Landi, and J. Shi. A penalized regression framework for building polygenic risk models based on summary statistics from genome-wide association studies and incorporating external information. *J. Am. Stat. Assoc.*, 116(533):133–143, 2021.

[19] X. Chen, A. Sanchis-Juan, C. E. French, A. J. Connell, I. Delon, Z. Kingsbury, A. Chawla, A. L. Halpern, R. J. Taft, D. R. Bentley, M. E. R. Butchbach, F. L. Raymond, and M. A. Eberle. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genet. Med.*, 2020.

[20] D. Clyde. Making the case for more inclusive GWAS, 2019.

[21] D. V. Conti, B. F. Darst, L. C. Moss, E. J. Saunders, X. Sheng, A. Chou, F. R. Schumacher, A. A. A. Olama, S. Benlloch, T. Dadaev, M. N. Brook, A. Sahimi, T. J. Hoffmann, A. Takahashi, K. Matsuda, Y. Momozawa, M. Fujita, K. Muir, A. Lophatananon, P. Wan, L. Le Marchand, L. R. Wilkens, V. L. Stevens, S. M. Gapstur, B. D. Carter, J. Schleutker, T. L. J. Tammela, C. Sipeky, A. Auvinen, G. G. Giles, M. C. Southey, R. J. MacInnis, C. Cybulski, D. Wokołorczyk, J. Lubiński, D. E. Neal, J. L. Donovan, F. C. Hamdy, R. M. Martin, B. G. Nordestgaard, S. F. Nielsen, M. Weischer, S. E. Bojesen, M. A. Røder, P. Iversen, J. Batra, S. Chambers, L. Moya, L. Horvath, J. A. Clements, W. Tilley, G. P. Risbridger,

H. Gronberg, M. Aly, R. Szulkin, M. Eklund, T. Nordström, N. Pashayan, A. M. Dunning, M. Ghoussaini, R. C. Travis, T. J. Key, E. Riboli, J. Y. Park, T. A. Sellers, H.-Y. Lin, D. Albanes, S. J. Weinstein, L. A. Mucci, E. Giovannucci, S. Lindstrom, P. Kraft, D. J. Hunter, K. L. Penney, C. Turman, C. M. Tangen, P. J. Goodman, I. M. Thompson, Jr, R. J. Hamilton, N. E. Fleshner, A. Finelli, M.-É. Parent, J. L. Stanford, E. A. Ostrander, M. S. Geybels, S. Koutros, L. E. B. Freeman, M. Stampfer, A. Wolk, N. Håkansson, G. L. Andriole, R. N. Hoover, M. J. Machiela, K. D. Sørensen, M. Borre, W. J. Blot, W. Zheng, E. D. Yeboah, J. E. Mensah, Y.-J. Lu, H.-W. Zhang, N. Feng, X. Mao, Y. Wu, S.-C. Zhao, Z. Sun, S. N. Thibodeau, S. K. McDonnell, D. J. Schaid, C. M. L. West, N. Burnet, G. Barnett, C. Maier, T. Schnoeller, M. Luedeke, A. S. Kibel, B. F. Drake, O. Cussenot, G. Cancel-Tassin, F. Menegaux, T. Truong, Y. A. Koudou, E. M. John, E. M. Grindedal, L. Maehle, K.-T. Khaw, S. A. Ingles, M. C. Stern, A. Vega, A. Gómez-Caamaño, L. Fachal, B. S. Rosenstein, S. L. Kerns, H. Ostrer, M. R. Teixeira, P. Paulo, A. Brandão, S. Watya, A. Lubwama, J. T. Bensen, E. T. H. Fontham, J. Mohler, J. A. Taylor, M. Kogevinas, J. Llorca, G. Castaño-Vinyals, L. Cannon-Albright, C. C. Teerlink, C. D. Huff, S. S. Strom, L. Multigner, P. Blanchet, L. Brureau, R. Kaneva, C. Slavov, V. Mitev, R. J. Leach, B. Weaver, H. Brenner, K. Cuk, B. Holleczek, K.-U. Saum, E. A. Klein, A. W. Hsing, R. A. Kittles, A. B. Murphy, C. J. Logothetis, J. Kim, S. L. Neuhausen, L. Steele, Y. C. Ding, W. B. Isaacs, B. Nemesure, A. J. M. Hennis, J. Carpten, H. Pandha, A. Michael, K. De Ruyck, G. De Meerleer, P. Ost, J. Xu, A. Razack, J. Lim, S.-H. Teo, L. F. Newcomb, D. W. Lin, J. H. Fowke, C. Neslund-Dudas, B. A. Rybicki, M. Gamulin, D. Lessel, T. Kulis, N. Usmani, S. Singhal, M. Parliament, F. Claessens, S. Joniau, T. Van den Broeck, M. Gago-Dominguez, J. E. Castelao, M. E. Martinez, S. Larkin, P. A. Townsend, C. Aukim-Hastie, W. S. Bush, M. C. Aldrich, D. C. Crawford, S. Srivastava, J. C. Cullen, G. Petrovics, G. Casey, M. J. Roobol, G. Jenster, R. H. N. van Schaik, J. J. Hu, M. Sanderson, R. Varma, R. McKean-Cowdin, M. Torres, N. Mancuso, S. I. Berndt, S. K. Van Den Eeden, D. F. Easton, S. J. Chanock, M. B. Cook, F. Wiklund, H. Nakagawa, J. S. Witte, R. A. Eeles, Z. Kote-Jarai, and C. A. Haiman. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat. Genet.*, 53(1):65–75, Jan. 2021.

[22] C. Crowley, Y. Yang, Y. Qiu, B. Hu, H. Won, B. Ren, M. Hu, and Y. Li. FIREcaller: an R package for detecting frequently interacting regions from Hi-C data. *bioRxiv*, 2019.

[23] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, Feb. 2002.

[24] K. Devarajan. Nonnegative matrix factorization: An analytical and interpretive tool in computational biology, 2008.

[25] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 2012.

[26] M. Dong, A. Thennavan, E. Urrutia, Y. Li, C. M. Perou, F. Zou, and Y. Jiang. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.*, 22(1):416–427, Jan. 2021.

[27] F. Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.*, 2013.

[28] C. B. Eaton, A. Young, M. A. Allison, J. Robinson, L. W. Martin, L. H. Kuller, K. C. Johnson, J. D. Curb, L. Van Horn, A. McTiernan, S. Liu, and J. E. Manson. Prospective association of vitamin D concentrations with mortality in postmenopausal women: results from the women's health initiative (WHI). *Am. J. Clin. Nutr.*, 94(6):1471–1478, Dec. 2011.

[29] P. Essletzbichler, T. Konopka, F. Santoro, D. Chen, B. V. Gapp, R. Kralovics, T. R. Brummelkamp, S. M. B. Nijman, and T. Bürckstümmer. Megabase-scale deletion using CRISPR/-Cas9 to generate a fully haploid human cell line. *Genome Res.*, 24(12):2059–2065, Dec. 2014.

[30] J. Euesden, C. M. Lewis, and P. F. O'Reilly. PRSice: Polygenic risk score software. *Bioinformatics*, 2015.

[31] R. Fang, M. Yu, G. Li, S. Chee, T. Liu, A. D. Schmitt, and B. Ren. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq, 2016.

[32] H. Feng, P. Jin, and H. Wu. Disease prediction by cell-free DNA methylation. *Brief. Bioinform.*, 20(2):585–597, Apr. 2018.

[33] A. Frigyesi and M. Höglund. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. *Cancer Inform.*, 6: 275–292, May 2008.

[34] GATK Team. Map and clean up short read sequence data efficiently. `https://gatk.broadinstitute.org/hc/en-us/articles/360039568932--How-to-Map-and-cleanup-`.

[35] R. Gaujoux and C. Seoighe. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, 2010.

[36] G. Genovese, D. J. Friedman, M. D. Ross, L. Lecordier, P. Uzureau, B. I. Freedman, D. W. Bowden, C. D. Langefeld, T. K. Oleksyk, A. L. Uscinski Knob, A. J. Bernhardy, P. J. Hicks, G. W. Nelson, B. Vanhollebeke, C. A. Winkler, J. B. Kopp, E. Pays, and M. R. Pollak. Association of trypanolytic ApoL1 variants with kidney disease in african americans, 2010.

[37] E. Geza, J. Mugo, N. J. Mulder, A. Wonkam, E. R. Chimusa, and G. K. Mazandu. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. *Brief. Bioinform.*, 2018.

[38] P. Giusti-Rodríguez, L. Lu, Y. Yang, C. A. Crowley, X. Liu, I. Juric, J. S. Martin, A. Abnousi, S. Colby Allred, N. Ancalade, N. J. Bray, G. Breen, J. Bryois, C. M. Bulik, J. J. Crowley, J. Guintivano, P. R. Jansen, G. J. Jurjus, Y. Li, G. Mahajan, S. Marzi, J. Mill, M. C. O'Donovan, J. C. Overholser, M. J. Owen, A. F. Pardiñas, S. Pochareddy, D. Posthuma, G. Rajkowska, G. Santpere, J. E. Savage, N. Sestan, Y. Shin, C. A. Stockmeier, J. T. R. Walters, S. Yao, Bipolar Disorder Working Group of the Psychiatric Genomics Consortium, Eating Disorders Working Group of the Psychiatric Genomics Consortium, G. E. Crawford,

F. Jin, M. Hu, Y. Li, and P. F. Sullivan. Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. Jan. 2019.

[39] I. J. Good and Y. Mittal. The amalgamation and geometry of Two-by-Two contingency tables. *Ann. Stat.*, 1987.

[40] D. U. Gorkin, Y. Qiu, M. Hu, K. Fletez-Brant, T. Liu, A. D. Schmitt, A. Noor, J. Chiou, K. J. Gaulton, J. Sebat, Y. Li, K. D. Hansen, and B. Ren. Common DNA sequence variation influences 3-dimensional conformation of the human genome. *Genome Biol.*, 2019.

[41] 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, Lead analysts:, Laboratory, Data Analysis &Coordinating Center (LDACC):, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, A. Battle, C. D. Brown, B. E. Engelhardt, and S. B. Montgomery. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, Oct. 2017.

[42] Y. Guan. Detecting structure of haplotypes and local ancestry. *Genetics*, 2014.

[43] D. Gurdasani, I. Barroso, E. Zeggini, and M. S. Sandhu. Genomics of disease risk in globally diverse populations, 2019.

[44] B. C. Hendrickson, C. Donohoe, V. R. Akmaev, E. A. Sugarman, P. Labrousse, L. Boguslavskiy, K. Flynn, E. M. Rohlfs, A. Walker, B. Allitto, C. Sears, and T. Scholl. Differences in SMN1 allele frequencies among ethnic groups within north america, 2009.

[45] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, Feb. 1970.

[46] E. A. Houseman, J. Molitor, and C. J. Marsit. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439, May 2014.

[47] E. A. Houseman, M. L. Kile, D. C. Christiani, T. A. Ince, K. T. Kelsey, and C. J. Marsit. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics*, 2016.

[48] L. Hsu, J. Jeon, H. Brenner, S. B. Gruber, R. E. Schoen, S. I. Berndt, A. T. Chan, J. Chang-Claude, M. Du, J. Gong, T. A. Harrison, R. B. Hayes, M. Hoffmeister, C. M. Hutter, Y. Lin, R. Nishihara, S. Ogino, R. L. Prentice, F. R. Schumacher, D. Seminara, M. L. Slattery, D. C. Thomas, M. Thornquist, P. A. Newcomb, J. D. Potter, Y. Zheng, E. White, and U. Peters. A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology*, 148(7):1330–1339.e14, June 2015.

[49] T. H. Hui Zou. Regularization and variable selection via the elastic net. In *Journal of the Royal Statistical Society, Series B*, 2005.

[50] L. N. Hutchins, S. M. Murphy, P. Singh, and J. H. Graber. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, 24(23):2684–2690, Dec. 2008.

[51] International Schizophrenia Consortium, S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, and P. Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, Aug. 2009.

[52] A. E. Jaffe and R. A. Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.*, 2014.

[53] I. Junier, Y. G. Spill, M. A. Marti-Renom, M. Beato, and F. Le Dily. On the demultiplexing of chromosome capture conformation data, 2015.

[54] I. Juric, M. Yu, A. Abnousi, R. Raviram, R. Fang, Y. Zhao, Y. Zhang, Y. Qiu, Y. Yang, Y. Li, B. Ren, and M. Hu. Maps: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLoS Comput. Biol.*, 2019.

[55] K. L. Keys, A. C. Y. Mak, M. J. White, W. L. Eckalbar, A. W. Dahl, J. Mefford, A. V. Mikhaylova, M. G. Contreras, J. R. Elhawary, C. Eng, D. Hu, S. Huntsman, S. S. Oh, S. Salazar, M. A. Lenoir, J. C. Ye, T. A. Thornton, N. Zaitlen, E. G. Burchard, and C. R. Gignoux. On the cross-population generalizability of gene expression prediction models, 2019.

[56] A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor, and S. Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations, 2018.

[57] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 2007.

[58] R. D. Langer, E. White, C. E. Lewis, J. M. Kotchen, S. L. Hendrix, and M. Trevisan. The women's health initiative observational study: baseline characteristics of participants and reliability of baseline measures. *Ann. Epidemiol.*, 13(9 Suppl):S107–21, Oct. 2003.

[59] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999.

[60] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, 2001.

[61] D. S. Lee, C. Luo, J. Zhou, S. Chandran, A. Rivkin, A. Bartlett, J. R. Nery, C. Fitzpatrick, C. O'Connor, J. R. Dixon, and J. R. Ecker. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat. Methods*, 2019.

[62] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform, 2009.

[63] Y. Li, M. Hu, and Y. Shen. Gene regulation in the 3D genome, 2018.

[64] Z. Li and H. Wu. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol.*, 20(1):1–17, 2019.

[65] Y. Liang, M. Pividori, A. Manichaikul, A. A. Palmer, N. J. Cox, H. Wheeler, and H. K. Im. Polygenic transcriptome risk scores improve portability of polygenic risk scores across ancestries.

[66] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct. 2009.

[67] S. Limou, G. W. Nelson, J. B. Kopp, and C. A. Winkler. APOL1 kidney risk alleles: population genetics and disease associations. *Adv. Chronic Kidney Dis.*, 21(5):426–433, Sept. 2014.

[68] L. R. Lloyd-Jones, J. Zeng, J. Sidorenko, L. Yengo, G. Moser, K. E. Kemper, H. Wang, Z. Zheng, R. Magi, T. Esko, A. Metspalu, N. R. Wray, M. E. Goddard, J. Yang, and P. M. Visscher. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nat. Commun.*, 2019.

[69] M. R. Lunn and C. H. Wang. Spinal muscular atrophy, 2008.

[70] Y. Luo, X. Li, X. Wang, S. Gazal, J. M. Mercader, 23 and Me Research Team, SIGMA Type 2 Diabetes Consortium, B. M. Neale, J. C. Florez, A. Auton, A. L. Price, H. K. Finucane, and S. Raychaudhuri. Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. *Hum. Mol. Genet.*, 30(16):1521–1534, July 2021.

[71] D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of Gene-Enhancer interactions. *Cell*, 161(5):1012–1025, May 2015.

[72] B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.*, 2013.

[73] D. Marnetto, K. Pärna, K. Läll, L. Molinaro, F. Montinaro, T. Haller, M. Metspalu, R. Mägi, K. Fischer, and L. Pagani. Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat. Commun.*, 2020.

[74] C. Márquez-Luna, P.-R. Loh, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, and A. L. Price. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.*, 41(8):811–823, Dec. 2017.

[75] A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.*, 100(4):635–649, Apr. 2017.

[76] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.*, 2019.

[77] J. Mbatchou, L. Barnard, J. Backman, A. Marcketta, J. A. Kosmicki, A. Ziyatdinov, C. Benner, C. O'Dushlaine, M. Barber, B. Boutkov, L. Habegger, M. Ferreira, A. Baras, J. Reid, G. Abecasis, E. Maxwell, and J. Marchini. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.*, 53(7):1097–1103, July 2021.

[78] J. L. Mega, N. O. Stitziel, J. G. Smith, D. I. Chasman, M. J. Caulfield, J. J. Devlin, F. Nordio, C. L. Hyde, C. P. Cannon, F. M. Sacks, N. R. Poulter, P. S. Sever, P. M. Ridker, E. Braunwald, O. Melander, S. Kathiresan, and M. S. Sabatine. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet*, 385(9984):2264–2271, June 2015.

[79] E. Mercuri, E. Bertini, and S. T. Iannaccone. Childhood spinal muscular atrophy: controversies and challenges. *Lancet Neurol.*, 11(5):443–452, May 2012.

[80] M. Mumbach, A. Rubin, R. Flynn, C. Dai, P. Khavari, W. Greenleaf, and H. Chang. HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *bioRxiv*, 2016.

[81] P. Natarajan, R. Young, N. O. Stitziel, S. Padmanabhan, U. Baber, R. Mehran, S. Sartori, V. Fuster, D. F. Reilly, A. Butterworth, D. J. Rader, I. Ford, N. Sattar, and S. Kathiresan. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*, 135(22): 2091–2101, May 2017.

[82] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, 2015.

[83] A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, and A. A. Alizadeh. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.*, 37(7):773–782, July 2019.

[84] A. Nott, I. R. Holtman, N. G. Coufal, J. C. M. Schlachetzki, M. Yu, R. Hu, C. Z. Han, M. Pena, J. Xiao, Y. Wu, Z. Keulen, M. P. Pasillas, C. O'Connor, C. K. Nickl, S. T. Schafer, Z. Shen, R. A. Rissman, J. B. Brewer, D. Gosselin, D. D. Gonda, M. L. Levy, M. G. Rosenfeld, G. McVicker, F. H. Gage, B. Ren, and C. K. Glass. Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science*, 366(6469):1134–1139, Nov. 2019.

[85] N. Pashayan, S. W. Duffy, D. E. Neal, F. C. Hamdy, J. L. Donovan, R. M. Martin, P. Harrington, S. Benlloch, A. Amin Al Olama, M. Shah, Z. Kote-Jarai, D. F. Easton, R. Eeles, and P. D. Pharoah. Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. *Genet. Med.*, 17(10):789–795, Oct. 2015.

[86] R. A. Patel, S. A. Musharoff, J. P. Spence, H. Pimentel, C. Tcheandjieu, H. Mostafavi, N. Sinnott-Armstrong, S. L. Clarke, C. J. Smith, VA Million Veteran Program, P. P. Durda, K. D. Taylor, R. Tracy, Y. Liu, C. W. Johnson, F. Aguet, K. G. Ardlie, S. Gabriel, J. Smith, D. A. Nickerson, S. S. Rich, J. I. Rotter, P. S. Tsao, T. L. Assimes, and J. K. Pritchard. Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. Mar. 2022.

[87] C. L. Pfaff, E. J. Parra, C. Bonilla, K. Hiester, P. M. McKeigue, M. I. Kamboh, R. G. Hutchinson, R. E. Ferrell, E. Boerwinkle, and M. D. Shriver. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.*, 68(1):198–207, Jan. 2001.

[88] T. W. Prior. Perspectives and diagnostic considerations in spinal muscular atrophy. *Genet. Med.*, 12(3):145–152, Mar. 2010.

[89] T. W. Prior and Professional Practice and Guidelines Committee. Carrier screening for spinal muscular atrophy. *Genet. Med.*, 10(11):840–842, Nov. 2008.

[90] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, Sept. 2007.

[91] E. Rahmani, R. Schweiger, L. Shenhav, T. Wingert, I. Hofer, E. Gabel, E. Eskin, and E. Halperin. BayesCCE: a bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. *Genome Biol.*, 19(1):141, Sept. 2018.

[92] V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, and J. Shendure. Massively multiplex single-cell Hi-C. *Nat. Methods*, 2017.

[93] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 2014.

[94] N. Rappoport, A. J. Simon, N. Amariglio, and G. Rechavi. The duffy antigen receptor for chemokines, ACKR1,- 'jeanne DARC' of benign neutropenia. *Br. J. Haematol.*, 184(4): 497–507, Feb. 2019.

[95] D. Reich, M. A. Nalls, W. H. L. Kao, E. L. Akylbekova, A. Tandon, N. Patterson, J. Mullikin, W.-C. Hsueh, C.-Y. Cheng, J. Coresh, E. Boerwinkle, M. Li, A. Waliszewska, J. Neubauer, R. Li, T. S. Leak, L. Ekunwe, J. C. Files, C. L. Hardy, J. M. Zmuda, H. A. Taylor, E. Ziv, T. B. Harris, and J. G. Wilson. Reduced neutrophil count in people of african descent is due

to a regulatory variant in the duffy antigen receptor for chemokines gene. *PLoS Genet.*, 5(1): e1000360, Jan. 2009.

[96] B. Rowland, R. Huh, Z. Hou, C. Crowley, J. Wen, Y. Shen, M. Hu, P. Giusti-Rodríguez, P. F. Sullivan, and Y. Li. THUNDER: A reference-free deconvolution method to infer cell type proportions from bulk Hi-C data. *PLoS Genet.*, 18(3):e1010102, Mar. 2022.

[97] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, 2005.

[98] A. D. Schmitt, M. Hu, I. Jung, Z. Xu, Y. Qiu, C. L. Tan, Y. Li, S. Lin, Y. Lin, C. L. Barr, and B. Ren. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.*, 2016.

[99] E. Sefer, G. Duggal, and C. Kingsford. Deconvolution of ensemble chromatin interaction data reveals the latent mixing structures in cell subpopulations. *J. Comput. Biol.*, 2016.

[100] T. M. Seibert, C. C. Fan, Y. Wang, V. Zuber, R. Karunamuni, J. K. Parsons, R. A. Eeles, D. F. Easton, Z. Kote-Jarai, A. A. Al Olama, S. B. Garcia, K. Muir, H. Grönberg, F. Wiklund, M. Aly, J. Schleutker, C. Sipeky, T. L. Tammela, B. G. Nordestgaard, S. F. Nielsen, M. Weischer, R. Bisbjerg, M. A. Røder, P. Iversen, T. J. Key, R. C. Travis, D. E. Neal, J. L. Donovan, F. C. Hamdy, P. Pharoah, N. Pashayan, K.-T. Khaw, C. Maier, W. Vogel, M. Luedeke, K. Herkommer, A. S. Kibel, C. Cybulski, D. Wokolorczyk, W. Kluzniak, L. Cannon-Albright, H. Brenner, K. Cuk, K.-U. Saum, J. Y. Park, T. A. Sellers, C. Slavov, R. Kaneva, V. Mitev, J. Batra, J. A. Clements, A. Spurdle, M. R. Teixeira, P. Paulo, S. Maia, H. Pandha, A. Michael, A. Kierzek, D. S. Karow, I. G. Mills, O. A. Andreassen, A. M. Dale, and PRACTICAL Consortium*. Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ*, 360:j5757, Jan. 2018.

[101] S. S. Shen-Orr and R. Gaujoux. Computational deconvolution: extracting cell type-specific information from heterogeneous samples, 2013.

[102] S. S. Shen-Orr, R. Tibshirani, P. Khatri, D. L. Bodian, F. Staedtler, N. M. Perry, T. Hastie, M. M. Sarwal, M. M. Davis, and A. J. Butte. Cell type-specific gene expression differences in complex tissues. *Nat. Methods*, 2010.

[103] T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, A. O'Shaughnessy-Kirwan, J. Cramard, A. J. Faure, M. Ralser, E. Blanco, L. Morey, M. Sansó, M. G. S. Palayret, B. Lehner, L. Di Croce, A. Wutz, B. Hendrich, D. Klenerman, and E. D. Laue. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 2017.

[104] L. Tan, D. Xing, C. H. Chang, H. Li, and X. S. Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 2018.

[105] M. Thomas, L. C. Sakoda, M. Hoffmeister, E. A. Rosenthal, J. K. Lee, F. J. B. van Duijnhoven, E. A. Platz, A. H. Wu, C. H. Dampier, A. de la Chapelle, A. Wolk, A. D. Joshi, A. Burnett-Hartman, A. Gsur, A. Lindblom, A. Castells, A. K. Win, B. Namjou, B. Van Guelpen, C. M. Tangen, Q. He, C. I. Li, C. Schafmayer, C. E. Joshu, C. M. Ulrich, D. T. Bishop, D. D. Buchanan, D. Schaid, D. A. Drew, D. C. Muller, D. Duggan, D. R. Crosslin, D. Albanes, E. L. Giovannucci, E. Larson, F. Qu, F. Mentch, G. G. Giles, H. Hakonarson, H. Hampel, I. B. Stanaway, J. C. Figueiredo, J. R. Huyghe, J. Minnier, J. Chang-Claude, J. Hampe, J. B. Harley, K. Visvanathan, K. R. Curtis, K. Offit, L. Li, L. Le Marchand, L. Vodickova, M. J. Gunter, M. A. Jenkins, M. L. Slattery, M. Lemire, M. O. Woods, M. Song, N. Murphy, N. M. Lindor, O. Dikilitas, P. D. P. Pharoah, P. T. Campbell, P. A. Newcomb, R. L. Milne, R. J. MacInnis, S. Castellví-Bel, S. Ogino, S. I. Berndt, S. Bézieau, S. N. Thibodeau, S. J. Gallinger, S. H. Zaidi, T. A. Harrison, T. O. Keku, T. J. Hudson, V. Vymetalkova, V. Moreno, V. Martín, V. Arndt, W. Q. Wei, W. Chung, Y. R. Su, R. B. Hayes, E. White, P. Vodicka, G. Casey, S. B. Gruber, R. E. Schoen, A. T. Chan, J. D. Potter, H. Brenner, G. P. Jarvik, D. A. Corley, U. Peters, and L. Hsu. Genome-wide modeling of polygenic risk score in colorectal cancer risk. *Am. J. Hum. Genet.*, 2020.

[106] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.*, 58(1): 267–288, Jan. 1996.

[107] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 2005.

[108] R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *aos*, 39(3): 1335–1371, June 2011.

[109] A. Torkamani, N. E. Wineinger, and E. J. Topol. The personal and clinical utility of polygenic risk scores, 2018.

[110] Y. Veturi, G. de Los Campos, N. Yi, W. Huang, A. I. Vazquez, and B. Kühnel. Modeling heterogeneity in the genetic architecture of ethnically diverse groups using random effect interaction models. *Genetics*, 211(4):1395–1407, Apr. 2019.

[111] B. J. Vilhjálmsson, J. Yang, H. K. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P. R. Loh, G. Bhatia, R. Do, T. Hayeck, H. H. Won, B. M. Neale, A. Corvin, J. T. R. Walters, K. H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. K. Chan, R. Y. L. Chen, E. Y. H. Chen, W. Cheng, E. F. C. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, L. E. Delisi, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, L. Georgieva, E. S. Gershon, I. Giegling, P. Giusti-Rodrguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, J. Grove, L. De Haan,

C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julia, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, B. J. Kelly, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. A. Kucinskiene, H. Kuzelova-Ptackova, A. K. Kahler, C. Laurent, J. L. C. Keong, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K. Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lnnqvist, M. Macek, P. K. E. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, P. B. Mortensen, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Mller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S. Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietilinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H. C. So, C. C. A. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Sderman, S. Thirumalai, D. Toncheva, P. A. Tooney, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. M. Wong, B. K. Wormley, J. Q. Wu, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, R. Adolfsson, O. A. Andreassen, P. M. Visscher, D. H. R. Blackwood, E. Bramon, J. D. Buxbaum, A. D. Børglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jonsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, B. J. Mowry, M. M. Nthen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St. Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, M. C. O'Donovan, P. Kraft, D. J. Hunter, M. Adank, H. Ahsan, K. Aittomäki, L. Baglietto, S. Berndt, C. Blomquist, F. Canzian, J. Chang-Claude, S. J. Chanock, L. Crisponi, K. Czene, N. Dahmen, I. Dos Santos Silva, D. Easton, A. H. Eliassen, J. Figueroa, O. Fletcher, M. Garcia-Closas, M. M. Gaudet, L. Gibson, C. A. Haiman, P. Hall, A. Hazra, R. Hein, B. E. Henderson, J. L. Hopper, A. Irwanto, M. Johansson, R. Kaaks, M. G. Kibriya, P. Lichtner, E. Lund, E. Makalic, A. Meindl, H. Meijers-Heijboer, B. Müller-Myhsok, T. A. Muranen, H. Nevanlinna, P. H. Peeters, J. Peto, R. L. Prentice, N. Rahman, M. J. Sánchez, D. F. Schmidt, R. K. Schmutzler, M. C. Southey, R. Tamimi, R. Travis, C. Turnbull, A. G. Uitterlinden, R. B. Van Der Luijt, Q. Waisfisz, Z. Wang, A. S. Whittemore, R. Yang, W. Zheng, S. Kathiresan, M. Pato, C. Pato, E. Stahl, N. Zaitlen, B. Pasaniuc, E. E. Kenny, M. H. Schierup, P. De Jager, N. A. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, N. Patterson, and A. L. Price. Modeling linkage disequilibrium increases accuracy of polygenic risk scores.

Am. J. Hum. Genet., 2015.

[112] X. Wang, J. Park, K. Susztak, N. R. Zhang, and M. Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.*, 10(1):380, Jan. 2019.

[113] Y. Wang, J. Guo, G. Ni, J. Yang, P. M. Visscher, and L. Yengo. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.*, 11(1):3865, July 2020.

[114] D. Wegmann, D. E. Kessner, K. R. Veeramah, R. A. Mathias, D. L. Nicolae, L. R. Yanek, Y. V. Sun, D. G. Torgerson, N. Rafaels, T. Mosley, L. C. Becker, I. Ruczinski, T. H. Beaty, S. L. R. Kardia, D. A. Meyers, K. C. Barnes, D. M. Becker, N. B. Freimer, and J. Novembre. Recombination rates in admixed individuals identified by ancestry-based inference. *Nat. Genet.*, 2011.

[115] O. Weissbrod, M. Kanai, H. Shi, S. Gazal, W. J. Peyrot, A. V. Khera, Y. Okada, A. R. Martin, H. Finucane, A. L. Price, and The Biobank Japan Project. Leveraging fine-mapping and non-european training data to improve cross-population polygenic risk scores. Jan. 2021.

[116] C. J. Willer, Y. Li, and G. R. Abecasis. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 2010.

[117] G. L. Wojcik, M. Graff, K. K. Nishimura, R. Tao, J. Haessler, C. R. Gignoux, H. M. Highland, Y. M. Patel, E. P. Sorokin, C. L. Avery, G. M. Belbin, S. A. Bien, I. Cheng, S. Cullina, C. J. Hodonsky, Y. Hu, L. M. Huckins, J. Jeff, A. E. Justice, J. M. Kocarnik, U. Lim, B. M. Lin, Y. Lu, S. C. Nelson, S.-S. L. Park, H. Poisner, M. H. Preuss, M. A. Richard, C. Schurmann, V. W. Setiawan, A. Sockell, K. Vahi, M. Verbanck, A. Vishnu, R. W. Walker, K. L. Young, N. Zubair, V. Acuña-Alonso, J. L. Ambite, K. C. Barnes, E. Boerwinkle, E. P. Bottinger, C. D. Bustamante, C. Caberto, S. Canizales-Quinteros, M. P. Conomos, E. Deelman, R. Do, K. Doheny, L. Fernández-Rhodes, M. Fornage, B. Hailu, G. Heiss, B. M. Henn, L. A. Hindorff, R. D. Jackson, C. A. Laurie, C. C. Laurie, Y. Li, D.-Y. Lin, A. Moreno-Estrada, G. Nadkarni, P. J. Norman, L. C. Pooler, A. P. Reiner, J. Romm, C. Sabatti, K. Sandoval, X. Sheng, E. A. Stahl, D. O. Stram, T. A. Thornton, C. L. Wassel, L. R. Wilkens, C. A. Winkler, S. Yoneyama, S. Buyske, C. A. Haiman, C. Kooperberg, L. Le Marchand, R. J. F. Loos, T. C. Matise, K. E. North, U. Peters, E. E. Kenny, and C. S. Carlson. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762):514–518, June 2019.

[118] N. R. Wray, M. E. Goddard, and P. M. Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.*, 17(10):1520–1528, Oct. 2007.

[119] J. Wu, Y. Liu, and Y. Zhao. Systematic review on local ancestor inference from a mathematical and algorithmic perspective. *Front. Genet.*, 12:639877, May 2021.

[120] Y. Xu, D. Vuckovic, S. C. Ritchie, P. Akbari, T. Jiang, J. Grealey, A. S. Butterworth, W. H. Ouwehand, D. J. Roberts, E. D. Angelantonio, J. Danesh, N. Soranzo, and M. Inouye. Learning polygenic scores for human blood cell traits. *bioRxiv*, 2020.

[121] Z. Xu, G. Zhang, F. Jin, M. Chen, T. S. Furey, P. F. Sullivan, Z. Qin, M. Hu, and Y. Li. A hidden markov random field-based bayesian method for the detection of long-range chromosomal interactions in Hi-C data. *Bioinformatics*, 2016.

[122] Z. Xu, G. Zhang, C. Wu, Y. Li, and M. Hu. FastHiC: A fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. In *Bioinformatics*, 2016.

[123] S. Yang and X. Zhou. Accurate and scalable construction of polygenic scores in large biobank data sets. *Am. J. Hum. Genet.*, 2020.

[124] P. Yousefi, K. Huen, H. Quach, G. Motwani, and others. Estimation of blood cellular heterogeneity in newborns and children for epigenomewide association studies. *Environmental and*, 2015.

[125] Y. Zhang, S. A. Sloan, L. E. Clarke, C. Caneda, C. A. Plaza, P. D. Blumenthal, H. Vogel, G. K. Steinberg, M. S. B. Edwards, G. Li, J. A. Duncan, 3rd, S. H. Cheshier, L. M. Shuer, E. F. Chang, G. A. Grant, M. G. H. Gephart, and B. A. Barres. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron*, 89(1):37–53, Jan. 2016.

[126] S. C. Zheng, S. Beck, A. E. Jaffe, D. C. Koestler, K. D. Hansen, A. E. Houseman, R. A. Irizarry, and A. E. Teschendorff. Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses, 2017.

[127] S. C. Zheng, C. E. Breeze, S. Beck, and A. E. Teschendorff. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat. Methods*, 2018.