Statistical and Machine Learning Methods for the Analysis of Summary Statistics Derived from Large Genomic Datasets

by

Kevin Stanford Liao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2023

Doctoral Committee:

Professor Sebastian Zöllner, Chair
Professor Veera Baladandayuthapani
Professor Jean Morrison
Professor Jonathan Terhorst
Professor William Wen

Kevin S. Liao

ksliao@umich.edu

ORCID iD: 0000-0003-4660-2480

Dedication

I dedicate this dissertation to all who have inspired my pursuit of scholastic knowledge in the field of statistics and genetics.

I am especially grateful for the support of my family for all their love, friends for their never-ending support, and Allison Joyce Reiner for everything else.

Acknowledgements

My Biostatistics PhD journey here at the University of Michigan has been a long, and at times, arduous road. Despite the difficulties and challenges along the way, the days I've spent in Ann Arbor have also been some of the best of my life. It is difficult to properly encapsulate the community of support and people who have made this dissertation possible, but I will try my best.

First and foremost, this dissertation and my growth as a scholar would not have been possible without the amazing guidance of my advisor Dr. Sebastian Zöllner. Your scientific enthusiasm, vast knowledge of (seemingly) anything genetics related, and patience working with me have been instrumental to me learning to conduct research and developing my individual interests. I want to further thank Dr. Jean Morrison for joining with and expanding the Zöllner-Morrison group and providing constant guidance, wisdom, and broadening my range of statistical genetics research. I also wish to thank my committee members Dr. William Wen, Dr. Veera Baladandayuthapani, and Dr. Jonathan Terhorst who have instructed me in Biostatistics courses, provided perceptive research feedback, and uniquely contributed research to the field of biostatistics and human genetics/genomics. Outside of my committee, I want to thank Dr. Michael Boehnke and the Genome Science Training Program, which generously supported me for two years during my PhD and allowed me the intellectual freedom to pursue and develop my own research interests that have formed the bulk of my dissertation. I want to further thank Dr. Michael Boehnke, Dr. Laura Scott, and Dr. Sarah Gagliano Taliun for the opportunity to work on

the inPSYght study as part of the WGSPD. Lastly, I want to thank Dr. Wei Wang and the statistical genetics team at 23andMe for the opportunity to intern during the summer of 2022, where my intern project formed the basis of the fourth chapter of my dissertation.

I won't try to thank all the amazing friends by name that have made my time both in and outside the Biostatistics department so enjoyable. First and foremost, thank you to present and past members of the Zöllner lab for engaging scientific conservations, the (self-proclaimed and pre-Covid) liveliest Biostatistics office space in SPH, and terrific end of semester BBQs. To all the friends within and outside the Biostatistics department, I'll fondly remember summer evenings at Bill's, tubing/kaying down the Huron, weekly board game nights and dinners, playing hooky on a weekday afternoon to play tennis/golf, and soccer games in the Ann Arbor Soccer Association. I even had the fortune to attend two Michigan Biostatistics weddings, where for one (shout out Speidi) I shared a dinner table with the chair of our own department. How many people can claim to have experienced that? As mentioned, some of the years in the program were the most challenging of my life. However, I wouldn't trade any of it because the challenges endured shaped the person I am today and forged a unique connection with others in the same boat, which ultimately allowed for such special connections and friendships.

I want to thank my parents, Chai-Ni and Tuan, for their love and never-ending support throughout my years of education. It's no coincidence that you retired from your careers as a Biostatistician and Software Engineer, and I ended up majoring in Biostatistics and Computer Science as an undergraduate. As a statistical geneticist, I love to joke with you both about how any intelligence or scholastic curiosity I have is solely thanks to the DNA you passed along. To my older sister Rachel and older brother Nolan, thanks to both of you for being some of the closest people in my life that have served as role models growing up. Lastly, I want to give a

special thank you to my amazing partner Allison Reiner. The last 1.5 years have seen us both go through major life changes, but at the center of it all you've been a constant that I can always turn to. Everyday I'm reminded of your selfless and kind spirit, proud of the career success you're having as a cancer researcher at Memorial Sloan Kettering, and grateful to have someone I can always be a goof around. Thank you for your being a constant source of motivation, and I'm looking forward to many more years of MM reminders, consumption of chocolates/sweets, and adventures in NYC.

Table of Contents

List of Tables

List of Figures

Abstract

Advancements in DNA sequencing over the past decade have transformed our ability to characterize genetic variation in large populations and study the genetics of many complex traits. For population geneticists, information on the genetic variation (i.e., which sites in the genome are mutated and at what frequency) alone is interesting as it allows for studying aspects of a population (e.g., demographic history, natural selection, and mutation rates). For statistical geneticists and genetic epidemiologists, the availability of phenotypic information in the same set of genetically sequenced individuals allow for studying the genetic basis of a complex trait. In this dissertation, I present three separate projects that leverage genetic information originating from DNA sequencing.

In the first project I focused on analyzing genetic variation without consideration of a phenotype, as is often done in the field of population genetics to make inferences on demographic history or natural selection. A commonly used summary statistic of genetic variation for population genetics inference is the allele frequency spectrum. However, methods based on the allele frequency spectrum make a simplifying assumption: all sites are interchangeable (i.e., an A->T mutation is the same as a C->T) mutation. In this project, I first extended previous literature to show heterogeneity in the allele frequency spectrum exists across mutation types at finer levels of resolution. I then illustrated how inferences of demographic history and natural selection are impacted by the violation of this assumption.

In the second project I focused on combining phenotypic information with genetic data through genome wide association studies (GWAS) and polygenic risk scores (PRS). GWAS estimate per-variant genetic effects on a complex trait, which can be used to summarize the genetic risk of that trait for an individual in PRS (constructed as the GWAS-weighted sum of their risk variants). However, PRS have a portability issue where phenotype predictions worsen as the ancestry of the target sample diverges from that of the GWAS sample. In admixed individuals, genome can be traced back to multiple ancestral populations and ancestry lies on a continuum. Such a continuum causes an ancestry dependence of PRS performance, as the PRS for samples whose ancestry better matches the external GWAS perform better. To help resolve this issue, I developed slaPRS, a stacking-based framework to integrate GWAS from multiple ancestral populations to construct polygenic risk scores (PRS) in admixed individuals. In simulations and real data, slaPRS performed well and reduced the ancestry dependence compared to existing approaches.

In the third project I focused on how genetic-phenotypic associations are shared across two more phenotypes through pleiotropy. Pleiotropy can be characterized at resolutions including genome wide, regionally, or at the SNP/gene-level. One approach to studying pleiotropy is local genetic correlation (LGC), which quantifies the extent of genetic sharing in a local region through the similarity in GWAS effect sizes. However, one problem of LGC is that it remains unable to identify SNP or gene-level pleiotropy, making it impossible to identify which variants or genes in a region drive a signal of LGC. To resolve this issue, I developed LDSC-MIX, a Bayesian mixture of regression method to infer latent groups of likely shared causal variants across two traits. In simulations and real data, LDSC-MIX identified SNP sets recovering the true LGC and tested whether genes in a region are enriched for such SNPs.

Chapter 1 Introduction

Since the introduction of next-generation DNA sequencing in 2005, advancements in the field of DNA-sequencing have allowed for genetic variation across the genome to be catalogued at resolutions and sample sizes previously not possible[1,2]. Large consortiums such as the 1000 Genomes Project[3], Trans-Omics for Precision Medicine (TOPMed)[4], and the Haplotype Reference Consortium[5] have performed short-read deep whole genome sequencing for thousands of samples across multiple ancestral populations to accurately call multiple types of genetic variation (e.g. single nucleotide polymorphisms, structural variation: indels/deletions, and more) down to the rarest frequencies. Genomic studies often further complement DNA-sequencing data with phenotype information on the same set of sampled individuals to identify genetic variants that contribute to disease risk. Large medical system-based cohort studies such as the UK Biobank project[6], Michigan Genomics Initiative[7], and All of Us[8] now routinely pair genetic information with a patient's medical profile, resulting in the pairing of genetic information with hundreds of phenotypic measurements collected through lab testing and provider diagnosis. The popularity of Biobanks globally for conducting genetic research has resulted in the Global Biobank Meta-analysis Initiative[9], which combines 23 biobanks across four continents to combine genotype-phenotype studies from more than 2.2 million consented individuals across diverse ancestries.

While we are now in an era of genomic research blessed with large sample sizes, current genomic studies now routinely have terabytes of genetic and phenotypic data which have spurred development of methods and strategies to summarize, manage, and analyze large data.

Summaries of data allow for efficient computation and alleviate privacy concerns for sharing individual-level genetic information across studies[10]. In the field of population genetics, which uses genetic information absent phenotype information to perform analysis such as inferring mutation rates, demographic history, and natural selection, one example of summarizing data is the allele frequency spectrum (AFS). The AFS is defined as the distribution of allele frequencies in a sample and serves as a summary statistic of the genetic variation in a population to effectively reduce terabytes of genetic data into a single distribution[11,12]. When phenotypic information is considered alongside genetic data, statistical geneticists and genetic epidemiologists frequently study genotype-phenotype associations using genome wide association studies (GWAS). GWAS identify risk variants for a phenotype through marginally testing and estimating the effect of each variant across the genome. The collection of GWAS estimated effect sizes, known as GWAS summary statistics, summarize genotype-phenotype associations without requiring individual-level genotype or phenotype information.

Over the past 15 years, GWAS have successfully implicated thousands of risk variants across hundreds of traits to usher in the post-GWAS era[13] where genetic associations are now used for functions such as genomic risk prediction[14], studying genetic architecture of traits[15], and integration with other omics data[16] to reveal biological function. Two uses of GWAS summary statistics that are of particular interest in the post-GWAS era are constructing polygenic risk scores (PRS) and studying pleiotropy. While a single GWAS risk variant may only explain a small percent of a trait's heritability, a sizable proportion of phenotypic variation can be explained by summarizing an individual's genetic risk for a given disease or trait in polygenic risk scores[17] (PRS). PRS are constructed as the weighted sum of risk variants with weights derived from GWAS summary statistics computed in an external GWAS sample. These PRS

have successfully been used[18,19] to identify individuals at high risk of disease, improve diagnostic accuracy, and allow for tailored personalized treatment for disease risk prediction in complex traits including coronary artery disease[20,21], type 1 and 2 diabetes[22,23], breast cancer[24,25], and more[26]. On the other hand, GWAS summary statistics can also be used to study pleiotropy[27,28] - the phenomenon in which genetic signals are shared across two or more phenotypes. Understanding of pleiotropy is important to reveal insights into the shared biological mechanism and pathways across traits and diseases. More importantly, as we enter the age of personalized medicine and genome editing, understanding of cross-trait genetics is imperative to avoid unexpected phenotypic effects[29].

While statistics summarizing genetic or genetic-phenotypic associations allow for convenient sharing and componential efficiency of methods, summary statistics present unique challenges as they often make simplifying assumptions or fail to capture all aspects of the original data. For example, the AFS (as well as most population genetics methods) makes a simplifying assumption that sites are interchangeable. In other words, an A->T mutation is equivalent to a C->G mutation. However, previous work has shown this assumption to be violated as there exists substantial heterogeneity in the AFS across sites due to unique evolutionary forces such as mutation rate heterogeneity and biased gene conversion. In the context of genotype-phenotype summaries, GWAS summary statistics report marginal associations that are affected by the correlation structure of the genome (i.e., linkage disequilibrium (LD)). In other words, SNPs across the genome are not independent causing SNPs to tag nearby variants, resulting in per-variant GWAS estimated effects sizes to be a function of nearby variant effects. Careful handling of LD in a sample/population thus becomes critical for genomic studies using GWAS summary statistics. An additional complication arises

when studying genetics across populations, as unique demographic histories have resulted in population differences in LD patterns and even genetic architecture which present further challenges. Nevertheless, GWAS summary statistics have become a common data input of methods development as they include a wealth of information on the genetic architecture of a trait and facilitate efficient computation without the need for individual level data.

In this dissertation, we propose novel analysis and statistical/machine learning methods utilizing summary statistics of genetic datasets. As the scale of genomic studies increases and the cost of generating genetic data decreases, the need for such studies increases. In Chapter 2, we begin by examining potential issues in current uses of the allele frequency spectrum to conduct population genetics inference using deep whole genome sequencing data. Current uses of the AFS assume interchangeability of sites and implicitly homogeneity in the AFS across different sites. However, previous work has shown heterogeneity in the allele frequency spectrum at the single base mutation level (A->T, C->G, etc), driven by evolutionary forces such as mutation rate heterogeneity and biased gene conversion. This finding raises the question of whether heterogeneity in the AFS persists at a finer resolution (i.e., 1-mers vs 3-mers) and the effect of this assumption violation on downstream inference.

In this work we address this question by first showing the effects of evolutionary forces on the AFS persist when considering 3-mer mutation subtypes (e.g., A[T->G]C). We recognize evolutionary forces act on the AFS concurrently and further propose a novel D_{-2} statistic that removed the contribution of singletons and doubletons from the AFS. We present theoretical justification that our statistic is a difference in unbiased estimators of the population genetics parameter θ and derive the analytical variance. We then analyze the effect of AFS heterogeneity across mutation subtypes on downstream demographic inference and shaping the local AFS in a

region. We find AFS heterogeneity at the genome wide level is substantial enough to infer drastically different parameters across mutation subtypes under an exponential growth and bottleneck growth demographic model. In local patterns of variation, we find a combination of regional subtype composition and local genomic factors shape the regional AFS across regions and discuss implications on local tests of selection using the AFS.

In Chapter 3, we consider GWAS summary statistics in the context of constructing PRS (weighted sum of risk variants) for genomic risk prediction in admixed individuals. Admixed individuals (e.g African Americans and Hispanics) are people who have segments of their genome tracing back to multiple ancestries due to interbreeding between previously isolated populations. Constructing well-performing PRS in admixed individuals is problematic due to a known portability problem of PRS, in which PRS perform worse as the ancestry of the target sample deviates from the external GWAS used. Because the ancestry of admixed individuals lies on a continuum, this portability issue results in an equitability problem as admixed individuals whose ancestry better matches the external GWAS will unequally benefit from the PRS[30]. Recent approaches have proposed using multiple ancestry GWAS in the context of admixed PRS, though the best way to integrate ancestry specific GWAS while considering local ancestry is still unclear.

Here, we introduce a novel stacking-based method, stacking local ancestry PRS (slaPRS), that integrates information from multiple ancestry GWAS in a local framework while explicitly modeling local ancestry to construct PRS in admixed individuals. We assess the performance of slaPRS using simulated population-specific GWAS and admixed African American genotypes, derived from population genetics models that contain realistic population-specific patterns of genetic variation. We consider disease architectures across a range of heritability, number of

causal variants, transethnic genetic architecture, and underrepresented GWAS sample sizes. Across all settings, slaPRS improves PRS performance and reduces the PRS ancestry dependence when stratifying admixed African Americans by quantiles of European ancestry. In real data applications, we apply slaPRS to admixed African British from the UK Biobank[6] to predict lipid traits (total cholesterol, HDL, LDL) using population-specific European and African American GWAS of the respective traits from the Global Lipids Genetics Consortium[31]. In our analysis, we find that slaPRS performs best or second best across methods in all quantiles of European ancestry. However, we find in the studied lipid traits that slaPRS performs comparably to an approach that combines multiple population GWAS globally, potentially driven by the genetic architecture of the lipid traits.

In Chapter 4, we consider GWAS summary statistics across complex traits to study pleiotropy (association of a genetic signal with multiple phenotypes). Pleiotropy can be studied at differing resolutions including the genome, region, or SNP/gene-level[32,33]. For example, PRS discussed in Chapter 3 can be used to assess genome-level pleiotropy through associating estimated PRS across different traits[34]. Another common approach to studying pleiotropy is via genetic correlation, which aims to directly quantify the strength of similarity between genetic variant GWAS effects across two traits[35]. Genetic correlation can be estimated using both individual level data and summary statistics, with summary statistic based methods such as LD score regression being more popular due to computational efficiency and preventing the need for individual-level genotype/phenotype data access[36]. Originally genetic correlation was studied using all variants genome wide to study pleiotropy at the genome resolution, though methods have recently been proposed for local genetic correlation (LGC) which can now estimate genetic correlation in a genomic region to identify pleiotropy at the region level[37–39]. However, current

genetic correlation approaches are limited in their ability to identify SNPs and genes driving a signal of LGC, making biological interpretability difficult and limiting clinical use.

In this work, we propose LDSC-MIX, a novel mixture of cross trait LD score regressions framework to identify latent sets of variants or genes that drive signals of local genetic correlation. While methods such as colocalization[40] already exist for studying pleiotropy at the SNP-level through providing information on the set of likely shared causal variants in a region, such methods typically rely on the presence of GWAS hits and perform poorly in scenarios of weaker genetic effects. In simulations, we divide the genome into approximately independent LD blocks where in each local region we simulate effect sizes (GWAS) across two traits varying the local SNP-based heritability, genetic correlation, and number of causal variants. We find that LDSC-MIX typically outperforms colocalization in local regions of multiple weaker shared causal variant genetic signals that have few or no GWAS significant hits. However, in single shared causal variant or strong multiple shared causal variant scenarios colocalization was the preferred approach. In trait pairs for immune, cancer, and cholesterol phenotypes from the UK Biobank, LDSC-MIX identifies sets of potentially shared causal variants that recover the estimated LGC, as well as potential genes enriched for inferred genetically correlated SNPs.

The projects in this dissertation are a step towards understanding and presenting novel approaches for how summary statistics of genetic variation and genotype-phenotype associations can be used to study population genetics, perform genomic risk prediction, and interrogate pleiotropic effects. Our analysis of the AFS summary statistic reveals to the field how current uses of the AFS to conduct inference may be biased by failing to consider AFS-heterogeneity across subtypes. Our method slaPRS constructs PRS in admixed individuals across a range of ancestry quantiles to facilitate genomic risk prediction through integrating GWAS summary

statistics across multiple populations. Finally, our method LDSC-MIX identifies SNP and gene level pleiotropy in regions of local genetic correlation to provide biological interpretability on cross trait genetics and can potentially identify shared causal risk variants undetected by existing approaches.

Chapter 2 The Effect of Mutation Subtypes on the Allele Frequency Spectrum and Population Genetics Inference

2.1 Introduction

The advent of whole genome sequencing in the past decade has transformed the field of population genetics and allowed for a host of new analyses on genetic variation both within and between populations[3,4,41–43]. As a result, this abundance of information has allowed for a host of methods to infer population genetic parameters such as mutation rates, demographic history, natural selection, and more[44–48]. One class of methods, that in recent years have regained popularity for population genetics inference due to their computational tractability, are based on the allele frequency spectrum (AFS)[49–53]. The AFS is defined as the distribution of allele frequencies at segregating sites in a sample and serves as a summary statistic of the genetic variation within that population[11,12]. As the AFS ignores information on linkage between sites by simply capturing the frequency of derived alleles in a sample, it effectively reduces genome-wide data for large samples into a single distribution. As a result, population genetics methods based on the AFS allow for analyzing millions of variable sites in thousands of individuals[49,50].

Current AFS-based methods to test for selection (using the local AFS in a genomic region) and infer demographic history (using the genome-wide AFS) use a frequency spectrum constructed from all segregating sites in a sample. These AFS-based methods can generally be grouped into two categories 1) Methods which reduce the high-dimensional AFS to a one-dimensional summary statistic such as Tajima's D, Fu and Li's D and F, and Fay and Wu's H[54–

[57] and 2) Methods that model the full AFS such as $\delta a \delta i$, momi, and SFselect[49,50,58]. Each of these methods leverage that selection and demographic history affect the shape of the frequency spectrum. Thus, comparing the observed AFS to the expected AFS under a neutral population or a particular demographic model can be used to test for selection or estimate demographic model parameters.

Typical construction of the AFS to summarize data and conduct inference treats all sites equally. However, the AFS can differ between sites due to heterogeneity in evolutionary forces. Across sites, mutation rates vary driven primarily by immediate surrounding sequence context and local genomic factors[47,48,59] (e.g. CpG TpG sites have orders of magnitude higher rates due to methylation). For fast mutating sites, recurrent mutations (i.e., multiple independent mutation events) violate the infinite sites model (assumes each site is equally likely to mutate and will only mutate once) and lead to multiple carriers of the same allele. Thus, in large samples, fast mutating sites [60–63] have a general shift away from rarer frequencies as two or more lower count mutations occurring at the same position are evaluated as a single higher count mutation (e.g. two singletons treated as one doubleton). While it is possible for the opposite scenario where an additional mutation reverses the original, such backwards mutations occur at a much lower rate. Moreover, empirical findings[61] suggest the scenario of shifting to higher allele counts is more prevalent in shaping the AFS. Another factor non-uniformly shaping the AFS across sites is biased gene conversion (gBGC), which occurs during recombination and is the process in which A/T – G/C heterozygotes have preferential transmission of the G/C allele. In highly recombining regions, gBGC leads to increases in the allele frequencies of A/T -> G/C mutation types[64–66] and is known to mimic selection. Due to these evolutionary forces uniquely shaping the AFS across different sites, combining all sites into a single overall AFS (as is typically done) may bias

inference as signals of selection or demographic history in the overall AFS are confounded by its composition of sites. This confounding will likely increase as genetic sample sizes grow larger and any AFS-heterogeneity across sites is exacerbated. However, the impact of this potential confounding on population genetics inference is currently unknown.

In this work we combined and extended these known drivers to study how the AFS varies across mutation subtypes and the downstream implications on population genetics inference. We used a collection of 53,133,922 SNPs from 3556 sequenced individuals from the Bipolar Research in Deep Genome and Epigenome Sequencing (BRIDGES) study[47]. Each variant was classified into one of 96 mutation subtype 3-mers defined by the specific point mutation and its immediate adjacent bases. For each subtype, we constructed an AFS to effectively partitioning the overall AFS into 96 distinct frequency spectra. Under the infinite sites model, these 96 AFSs should differ only due to sampling variation. We showed that the AFS differs widely between subtypes, even outside CpG TpG sites, with much of these differences being driven by mutation rate heterogeneity and biased gene conversion, two factors previously implicated to shape the AFS at the single base level. Signals of gene conversion on the AFS may be confounded by mutation rates (that primarily affect the extremely rare variant counts). To disentangle the two factors, we further further derived a novel Tajima's D type statistic D_{-2} that removes the singleton and doubleton contribution while retaining the same functional form for interpretability. As a result of AFS heterogeneity across subtypes, theoretical inference of demographic history using the full genome wide AFS for a single subtype under a growth and three-epoch model varied drastically among the subtypes. Similarly, in local genomic regions we found both the local composition of subtypes and genomic factors, such as recombination rate, to be significant predictors of the regional AFS across the genome.

2.2 Materials and Methods

2.2.1 Comparison of the AFS across Mutation Subtypes to Identify Signals of Evolutionary Forces Driving AFS Heterogeneity

We analyzed whole genome sequencing data from the Bipolar Research in Deep Genome and Epigenome Sequencing (BRIDGES) study of unrelated individuals of European ancestry. The samples were aggregated from a variety of studies that each collected cases and controls of individuals with Bipolar Disorder[47]. Sequencing was performed, per the Illumina protocol on Build GRCh37, to generate our final dataset with a mean coverage of 9.6x across individuals. After filtering out samples with high contamination, case misspecification, ancestry outliers, and relatedness our final sample included 3556 unrelated European individuals with a total of 56,482,865 variants.

Each of the single nucleotide polymorphisms (SNPs) observed in our dataset was classified as one of six mutation types, determined by the ancestral allele and the derived allele: A->C, A->G, A->T, C->A, C->G, C->T. We determined the ancestral allele using the 1000 Genomes ancestral alleles for Build 37 and annotated via bcftools[3,67]. Note that each mutation type is defined to account for the complementary nature of DNA, and thus we ignored which strand the mutation occurs on (e.g., a SNP with ancestral allele T and derived allele G corresponds to an A->C SNP on the opposite strand and vice versa, so for brevity, we classify both as A->C mutations). Each mutation type was further refined into 3-mer mutation subtypes by considering the surrounding immediate nucleotides using the GRCh37 human reference sequence. This resulted in 96 distinct mutation subtypes for our analysis (4 possible bases downstream * 6 mutation types * 4 possible bases upstream = 96 mutation subtypes). In previous

work[47] we estimated relative mutation rates for each 3-mer using singletons from the same BRIDGES dataset.

Comparison of the AFS across Mutation Subtypes to Identify Signals of Evolutionary Forces Driving AFS Heterogeneity

We first constructed a distinct unfolded AFS for each of the 96 mutation subtypes. For a given haploid sample of size n, let η_i be the number of segregating sites in the sample in which exactly i individuals have the derived allele. The AFS is then defined by the vector $(\eta_1, \eta_2, \cdots, \eta_{n-1})$.

We summarized and compared the 96 subtype AFSs using the ratio of singletons to doubletons and Tajima's D[54]. The ratio of singletons to doubletons ($\frac{\eta_1}{\eta_2}$) was used to identify signals of recurrent mutations lowering the singleton count[61] for sites with higher mutation rates, where the ratio reflected any reduction in singletons and increase in doubletons. While recurrent mutations work to shift the frequency for many of the rarest frequency variants in high mutation rate sites, their effect is most prevalent in shifting two singletons to a single doubleton count[61–63]. Tajima's D was used to identify signals of biased gene conversion, where Tajima's D is a summary statistic of the high-dimensional AFS computed by comparing two unbiased estimates of $\theta = 4N\mu$ (N is effective diploid population size and μ is mutation rate) under a neutral population model: Watterson's estimator θ_W and Mean Pairwise Difference θ_π:

$$D = \frac{\theta_\pi - \theta_W}{\sqrt{var(\theta_\pi - \theta_S)}}$$

$$\theta_\pi = \binom{n}{2}^{-1} \left[\sum_{i=1}^{n-1} \eta_i i(n-i) \right]$$

$$\theta_W = \frac{S}{h_n}$$

Here S is the number of segregating sites, n is the haploid sample size, and $h_n = \sum_{i=1}^{n-1} \frac{1}{i}$. θ_π assigns more weight to alleles segregating at intermediate counts compared to θ_W, which weights all allele counts equally[68]. As a result, an excess of rare or intermediate frequency alleles in the AFS can be observed in the sign of Tajima's D. As gBGC skews allele frequencies towards intermediate frequency variants for weak to strong mutation types (A/T -> C/G) and towards rare variants for strong to weak mutation types[64], we expect weak to strong mutation types to have a more positive Tajima's D and vice versa for strong to weak types.

As Tajima's D is strongly influenced by the singleton and doubleton count, we derived a novel D statistic, which we call D_{-2}, that removes the singleton η_1 and doubleton η_2 contribution to Tajima's D:

$$D_{-2} = \frac{\theta_{\pi-2} - \theta_{W-2}}{\sqrt{var(\theta_{\pi-2} - \theta_{w-2})}}$$

$$\theta_{\pi-2} = \frac{n(n-1)}{(n-2)(n-3)} * \binom{n}{2}^{-1} \left\{ \left[\sum_{i=1}^{n-1} \eta_i i(n-i) \right] - \eta_1(n-1) - 2\eta_2(n-2) \right\}$$

$$\theta_{W-2} = \frac{S - \eta_1 - \eta_2}{h_n - \frac{3}{2}}$$

In the numerator, $\theta_{\pi-2}$ and θ_{W-2} are the Mean Pairwise Difference and Watterson's estimators with η_1 and η_2 removed and then reweighted to ensure both estimators stay unbiased for θ under a neutral population. Thus, the D_{-2} statistic allowed us to summarize the AFS for all other allele counts in a familiar form to investigate the effects of gene conversion on the AFS without potential confounding of the singleton/doubleton count driven by recurrent mutations in sites with higher mutation rates. We derived the analytical form for the variance of the difference in

$\theta_{\pi-2}$ and θ_{W-2} using covariance derivations for linear combinations of the AFS[69,70] (Supplementary). To check the behavior of our statistic under the null, we simulated neutral frequency spectra for two separate subtype's estimate of θ using fastsimcoal2[71] (Supplementary). Similar to Tajima's D, comparisons of our D_{-2} statistic across subtypes were used to interrogate potential signals of gBGC.

2.2.2 Effect of Heterogeneity in the Genome Wide AFS Across Subtypes on Demographic Inference

To assess the effect of AFS heterogeneity across mutation subtypes on demographic inference we used the method $\delta a \delta i$[49], which takes a diffusion approach to simulate the AFS for a predefined demographic model. Once simulated, comparisons to the observed genome-wide AFS allow $\delta a \delta i$ to infer parameters for the given demographic model[49]. Here, we inferred demographic model parameters separately for each of the 96 genome-wide AFS across mutation subtypes to assess systematic differences. For each $\delta a \delta i$ run, we considered two models of population history. In the first model, we modeled a population undergoing exponential growth (Figure 2.1), with a constant ancestral effective population size N_e that started exponentially growing at some time T_0 in the past to a present population size λN_e while mutations accumulate with rate μ. In the second model, we considered a modified three-epoch model (Figure 1), a more natural model for the human population that allows for two changes of population size. We model a population with an ancestral population size N_e that at time T_0 contracts in a bottleneck of length T_1 to size $N_e\lambda_1$ and recovers to relative size $N_e\lambda_2$ during time T_2. Under both models, we inferred 1) the compound parameter $\theta = 4N_e\mu$, 2) times T_0 (both models) and T_1, T_2(three epoch model) in generations, and 3) the ratios between the ancestral and post-change population sizes λ (growth), λ_1, λ_2 (three-epoch).

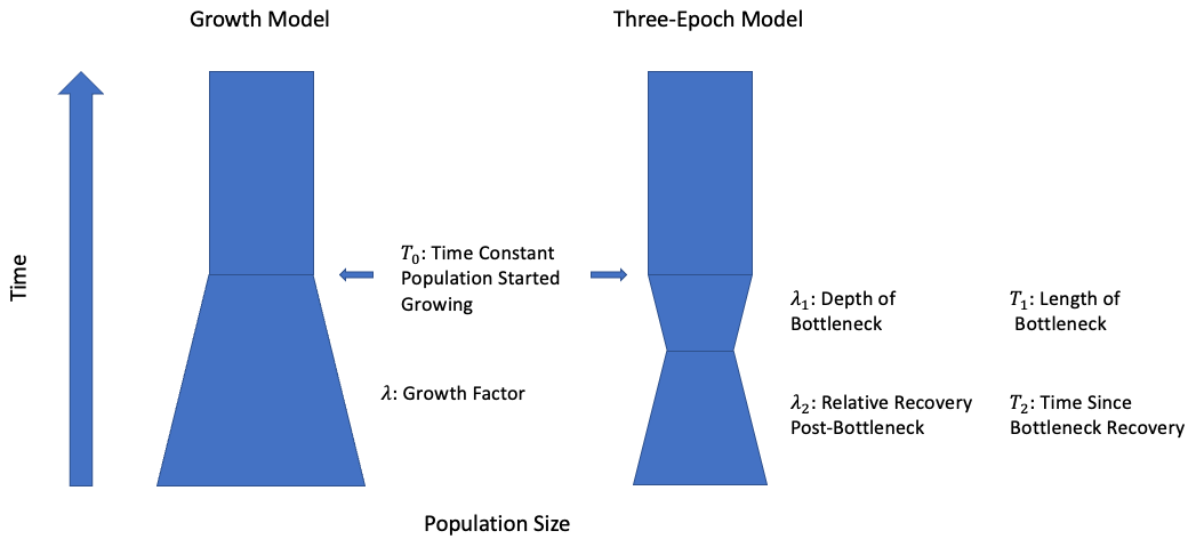

Figure 2.1 Diagram of growth and three-epoch demographic models fit in $\delta a \delta i$ Parameters of growth model include time since ancestral constant size population started growing and growth factor. Parameters of three-epoch model include length of bottleneck, time since bottleneck recovery as well as bottleneck depth and recovery. Each model fit separately using the 96 distinct subtype AFS.

We computed the ancestral effective population size solving $N_e = \frac{\theta}{4\mu}$ where the absolute mutation rate μ was derived from extremely rare variants in the same BRIDGES dataset[47]. To derive an absolute per-site, per-generation mutation rate from the relative mutation rate, we assume 60 de novo mutations per generation (a value typically observed in trio studies)[72,73]:

$$\mu_{subtype} = Rel\ Mut\ Rate_{subtype} * \frac{60}{\sum_{subtypes} \#\ Motifs_{subtype} * Rel\ Mut\ Rate_{subtype}}$$

2.2.3 Effect of Heterogeneity in the Local Composition of Mutation Subtypes on the Regional AFS

Practical applications of AFS-based statistics (such as Tajima's D) often use the combined local allele frequency spectrum (i.e., derived from all segregating sites, regardless of subtype) as it varies over non-overlapping windows across the genome. We evaluated how the local AFS could be shaped by heterogeneity in its composition of mutation subtypes. To reduce

the potential confounding of selection when assessing the relationship between local region subtype composition and the AFS, we subset sites to only intergenic sites (assuming limited selection on intergenic regions[74,75]) annotated using EPACTs[76]. We partitioned the remaining genome in 100kb windows and computed in each window from the local AFS 1) Tajima's D, 2) the proportion of overall singletons, doubletons, and tripletons and 3) the counts/proportion of each of the 96 subtypes comprising the local AFS. For each subtype, we then ranked windows according to proportion constituting the local AFS and classified a window as being "abundant" in that subtype if its proportion fell in the top 10% of windows (Figure 2.2).

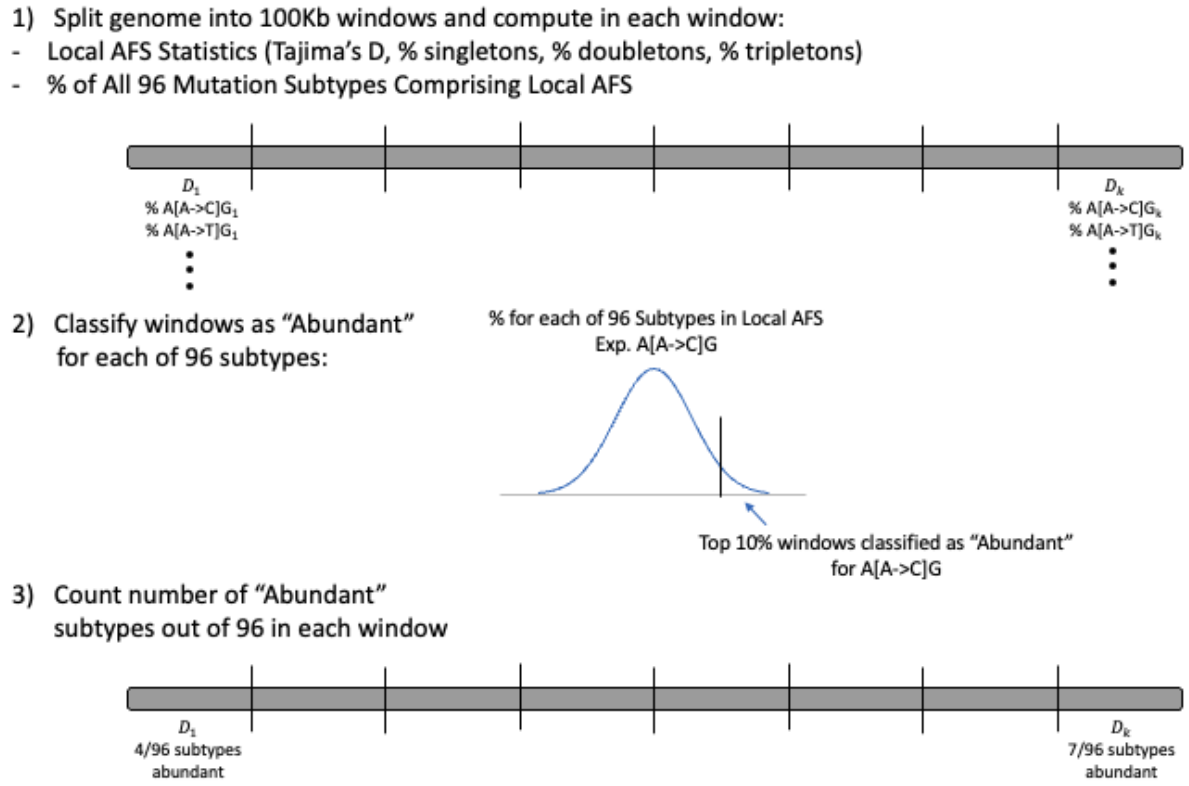

Figure 2.2 Diagram of analysis to assess local mutation subtype composition and regional AFS. In 100kb windows we compute 1) local AFS statistics (Tajima's D, % singletons, etc), 2) the counts and proportions comprising the local AFS for each of 96 mutation subtypes. Windows in the top 10% of subtype proportion across the genome are classified as "abundant" for that subtype and we count the number of abundant subtypes in each window.

We first evaluated whether the distribution of Tajima's D across these windows was independent of the distribution of variants by stratifying windows into 10 quantiles by Tajima's D and finding the mean number of abundant subtypes out of 96 in each quantile. Under the null expectation that the composition of mutation subtypes in a local AFS is independent of the Tajima's D in a window, we would expect the number of abundant subtypes to be constant across quantiles. We further stratified abundant subtypes by whether these subtypes were 1) low vs high mutation rate using the median mutation rate across subtypes as the separator and 2) direction of gene conversion by WS, SW, and Neutral to interrogate whether mutation rate heterogeneity or biased gene conversion drove the dependence in the distribution of local Tajima's D and variants.

We quantified the overall contribution of local subtype heterogeneity on the observed local AFS statistics (Tajima's D, % singletons, etc.) by computing the expected value of each statistic. Using the count of each subtype as a weight, we computed the expected local AFS statistic in a window as a weighted mean of the 96 subtype's genome wide observed value. For example, the expected Tajima's D in a window is:

$$D_{expected} = \frac{\sum_{i=1}^{96} D_{GW}^{i} * w_i}{\sum_{i=1}^{96} w_i}$$

Where D_{GW}^{i} is Tajima's D using the genome wide AFS for subtype i and the weights w_i are the # of subtypes i in the window. If local heterogeneity in mutation subtypes fully explained regional differences in a statistic, we would expect observed and expected statistic values to be equal. We fit a multivariate generalized estimating equation (GEE) linear model with a working exchangeable correlation structure for each chromosome:

$$D_{obs,i} = \beta_0 + \beta_{exp} D_{exp,i} + \beta_{RR} RR_i + \beta_{GC} GC_i$$

$$R(\rho): \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & 1 & \vdots \\ \rho & \cdots & 1 \end{pmatrix}$$

Where $\beta_{exp}, \beta_{RR}, \beta_{GC}$ are effects of the expected statistic, recombination rate, and GC content in a window. We adjust for recombination rate and GC content in a window because recombination rate is known to confound selection[77,78], and GC content affects germline mutation rates[79]. The GEE framework was used to produce robust β standard errors because neighboring genomic regions have correlated statistics (i.e Tajima's D) and thus would affect the standard error estimates of an ordinary linear regression.

2.3 Results

We leveraged large sample whole genome sequencing to first evaluate patterns in the overall AFS and assess potential heterogeneity in the AFS across mutation subtypes. In the overall genome wide AFS, among the 56,482,865 total SNPs across N = 3556 samples, we observed an excess of rare variation shown through the proportion of singletons (60.3%), doubletons (9.91%), and tripletons (4.01%) (Table 2.1). This excess of rare variation is consistent with exponential and accelerating faster than exponential population growth in recent human demographic history[80–82]. When partitioning the overall AFS by the six mutation types, the proportion of singletons varied greatly with C->A mutation types having the highest singleton proportion (63.7%) and C->T mutation types having the lowest (58.7%) (Table 2.1). We observed additional variation in the AFS when considering mutation types on a more granular scale through stratifying by flanking nucleotides. For example, we found the A[C->T]G mutation subtype, a CpG TpG site with an outlier mutation rate, had considerably lower singleton and higher doubleton proportions (53.82%, 13.68%) when compared to A[C->T]A (61.93%, 9.67%), A[C->T]C (59.29%, 9.53%), and A[C->T]T (60.86%, 9.51%). Notably, even

outside the CpG TpG subtype, some singleton proportions of the same mutation type differ by >2% across subtypes when altering the +1 base (C vs A). Due to our large sample sizes, differences in the listed singleton and proportions across the four A[C->T]X subtypes were highly significant after adjusting for multiple comparisons ($p < 0.001$). Counts and proportions for the other 92 3-mer subtypes can be found in the appendix.

Tajima's D, a summary statistic of the entire high-dimensional genome-wide AFS, ranged from -2.19 to -1.50 across the 96 subtypes (Figure 2.7). Uniformly negative values reflected the excess of rare variation observed. Even within a single mutation type there was variation in Tajima's D when further considering adjacent nucleotides. For example, among C->G mutation types, Tajima's D ranged from -2.19 to -1.50, with three CpG subtypes G[C->G]G (-1.50), A[C->G]G (-1.52), and T[C->G]G (-1.74) having clear lower outlier Tajima's D values. Similar to comparisons in the singleton proportions, even outside CpG subtypes there existed variation in Tajima's D (e.g., A->G mutation types ranged from -2.09 to -1.83). Substantial differences in both the proportions of singleton-tripleton proportions and Tajima's D across the 96 3-mer subtypes summarizes heterogeneity in the AFS across sites.

MT	# Sites	Singletons (%)	Doubletons (%)	Tripletons (%)
A->C	4,187,865	61.4	9.29	3.77
A->G	16,132,320	60.2	9.45	3.85
A->T	4,020,474	61.5	9.35	3.86
C->A	5,944,362	63.7	9.11	3.68
C->G	4,898,261	62.3	9.43	3.82
C->T	22,678,843	58.7	10.7	4.31
Overall	53,133,922	60.3	9.91	4.01

Mutation Subtype	# Sites	Singletons (%)	Doubletons (%)	Tripletons (%)
A[C->T]A	1,636,849	61.93	9.67	3.91
A[C->T]C	1,026,598	59.29	9.53	4.01
A[C->T]G	2,240,435	53.82	13.68	5.38
A[C->T]T	1,170,191	60.86	9.51	3.88

Table 2.1 Genome-wide counts and proportion of singletons, doubletons, and tripletons. a) Counts and proportions for the six main mutation types and b) Counts and proportions for A[C->T]X mutation subtypes varying the base downstream.

2.3.1 A Comparison of the AFS across Mutation Subtypes to Identify Evolutionary Forces Driving AFS Heterogeneity

When assessing the relationship between mutation rates and the singleton to doubleton ratio, we found the ratio of singletons to doubletons was highly negatively correlated with the estimated singleton-derived mutation rates[3] across the 96 mutation subtypes ($\rho = -0.84,\ p = 2.2e^{-16}$) (Figure 2.3). As previously mentioned, sites with higher mutation rates are more susceptible to recurrent mutations. To assess whether the signal was driven by the CpG TpG sites with their order of magnitude higher mutation rates, we repeated the analysis after removing the 4 CpG TpG subtypes and found the ratio of singletons to doubletons was still negatively correlated with the estimated singleton-derived mutation rates[47] across the 96 mutation subtypes

($\rho = -0.35,\ p = 6.1e^{-4}$). We further stratified subtypes by the six mutation types, observing consistently negative correlations between $\rho = -0.21$ and $\rho = -0.97$ (Table 2.3). Four of the six correlations (A->G, C-> A, C->G and C->T) were statistically significant even though each correlation is based on only 16 observations.

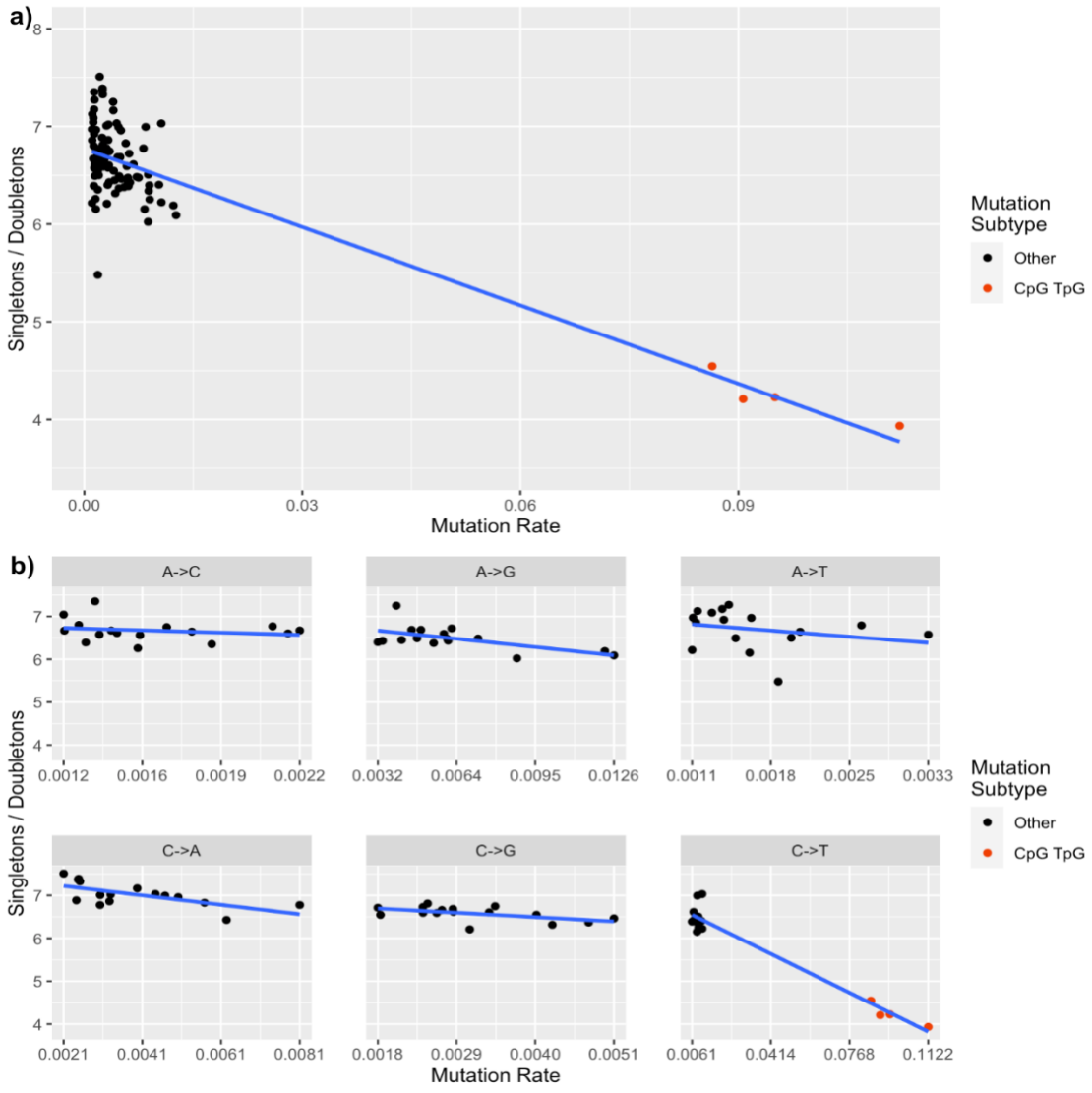

Figure 2.3 Correlation between the ratio of singletons to doubletons by the estimated mutation rate from extremely rare variants. Black points represent a mutation subtype, red points represent the four CpG TpG sites, and the blue line is the least squares regression line. a) Correlation of all 96 mutation subtypes b) Correlation by six mutation types.

When comparing Tajima's D and our D_{-2} statistic across subtypes to identify signals of gene conversion on the AFS, Tajima's D ranged from [-2.19, -1.84] while our D_{-2} statistic ranged from [-0.44, 0.10], with a strong correlation between the two statistics across subtypes ($\rho = 0.76, p = 2.2e-16$). Under the null, we observed that D_{-2} was asymmetric with a heavier positive tail suggesting that utility of this statistic beyond summarizing the AFS shape without singletons and doubletons may be limited (Figure 2.8). We grouped non-CpG mutation subtypes into weak to strong (WS), strong to weak (SW), and indifferent variants with each category having 32 subtypes and compared the mean Tajima's D between groups (Table 2.2). We observed the smallest average Tajima's D in SW (-2.071), followed by indifferent (-2.048) and WS (-2.004). Comparing each category against the mean of the other two categories, we found a statistically significant difference in means for WS vs SW and indifferent (p = 2.2e-4, t-test) and SW vs WS and indifferent (p=2.2e-3, t-test). The "more positive" mean Tajima's D for WS subtypes compared to non-WS indicated an excess of intermediate frequency variants, which is consistent with a model of gBGC where low-frequency S alleles get transmitted more often than expected and thus reach intermediate allele frequency more frequently. Similarly consistent with expectations under gBGC, the mean Tajima's D was "more negative" for SW subtypes compared to non-SW.

As Tajima's D is strongly dependent on the number of singletons and doubletons, the effect of gBGC may be confounded by mutation rate heterogeneity distorting primarily the singleton and doubleton counts. To limit the effect of mutation rate heterogeneity, we repeated the analysis using our D_{-2} statistic that ignores all singletons and doubletons in its calculation. We again observed the smallest Tajima's D in SW followed by indifferent and WS (though indifferent and WS had very similar values), which was consistent with the hypothesized effect

of gBGC on the AFS. When comparing each category against the mean of the other two categories, we observed a weaker signal of gBGC as only the SW comparison had a borderline statistically significant difference in mean Tajima's D (p=0.080, t-test) compared with WS and indifferent.

Mutation Group	Tajima's D No CpGs			D_{-2} Estimator No CpGs		
	Group Mean	Non-Group Mean	P-value	Group Mean	Non-Group Mean	P-value
Weak to Strong (A->C, A->G)	-2.004	-2.058	2.2e-4	-0.243	-0.253	0.399
Indifferent: (A->T, C->G)	-2.048	-2.033	0.387	-0.243	-0.252	0.450
Strong to Weak (C->A, C->T)	-2.071	-2.024	2.2e-3	-0.264	-0.243	0.080

Table 2.2 Mean Tajima's D (left) and D-2 estimator (right) for subtypes in each mutation group against subtypes not in group. P-values computed from two-sample t-test. CpG subtypes were excluded from analysis.

2.3.2 Effect of Heterogeneity in the Genome Wide AFS Across Subtypes on Demographic Inference

Inferred demographic parameters varied drastically when running $\delta a \delta i$ separately across the 96 distinct mutation subtype-specific genome-wide AFS. Under the simple exponential growth model (Figure 2.1), estimates of the ancestral effective population size derived from the inferred population genetics parameter θ varied two-fold from 5062.99 to 10518.87 across the 96 subtype's genome wide AFS. Similarly, estimates of the time at which the ancestral constant population size started growing varied from 96.53 to 206.02 generations across subtypes (see Appendix). There was a strong correlation (ρ = 0.70, p = 2.06e-15) between the genome-wide

proportion of singletons for a given subtype and its inferred relative growth (Figure 2.4). This is expected since the singleton count should be highly informative of growth rate and time since growth under an exponential growth model of human populations[81,83]. CpG TpG sites have an overall lower proportion of singletons due to their higher mutation rate causing recurrent mutations, and thus had smaller inferred relative growth. However, two non-CpG subtype A[A->G]G and T[A->T]A had a similarly low proportion of singletons (< 57%) and smaller inferred relative growth (114.07 and 76.44). In addition, the subtype G[A->G]G had an outlying higher inferred relative growth (211.71) given its proportion of singletons (0.60) as compared to the trend of other subtypes.

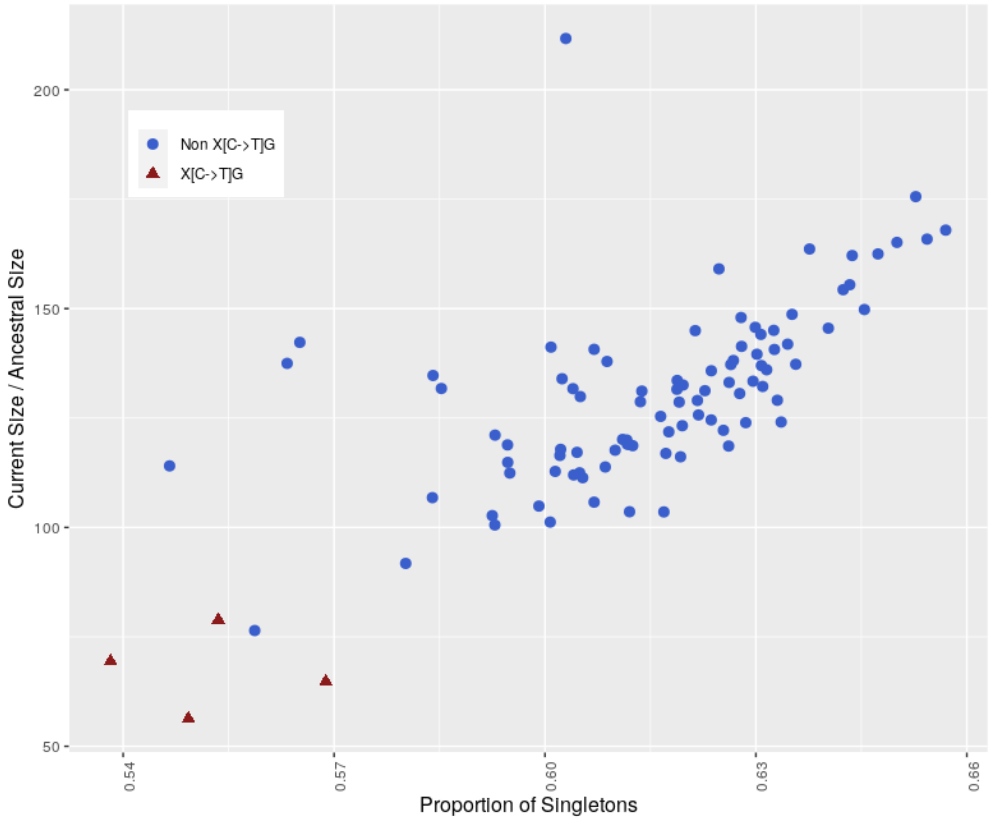

Figure 2.4 Scatterplot of inferred relative growth by the proportion of singletons for each of the 96 mutation subtype's AFS. Points in blue are non CpG TpG sites and points in red are CpG TpG sites with outlier higher mutation rates driving lower proportion of singletons.

For the constrained three-epoch model, inferred parameters across subtypes similarly varied drastically with the total time since the ancestral population first changed ranging from 7,530 to 824,292 generations, relative bottleneck depth ranging from 0.07 to 0.99, and relative recovery ranging from 3.77 to 90.43. Conclusions about the existence of a historical bottleneck varied across subtype inference, with four subtypes suggesting a nonexistent or very small bottleneck (less than a 15% decrease in effective population size). The remaining subtypes suggested a moderate to severe bottleneck with population contractions widely ranging from 40% to 93%. Excessively large times and recovery post bottleneck were likely driven by model constraints which forced a bottleneck to occur.

2.3.3 Effect of Heterogeneity in the Local Composition of Mutation Subtypes on the Regional AFS

In our regional analysis to assess whether local heterogeneity in subtype composition in a 100kb windows shaped the regional AFS, we observed a general non-independence in the distribution of variants and Tajima's D. Windows were first characterized as "abundant" in a given subtype for each of the 96 subtypes subtype if their proportion comprising the local AFS fell in the top 10% of windows genome wide (Figure 2.2). After separating windows into 5% quantiles based off Tajima's D, windows in the lowest and highest 5% of Tajima's D quantile had on average more subtypes out of 96 with extreme abundances (12.35 and 9.79 respectively vs 7.95) compared to the median D quantile average (Figure 2.5), suggesting a more extreme composition of mutation subtypes in windows falling in the tails of the genome wide Tajima's D distribution. When stratifying abundant subtypes by direction of gene conversion and low/high

mutation rate, we found no observable trends in proportions of each category across D quantiles (Figure 2.10).

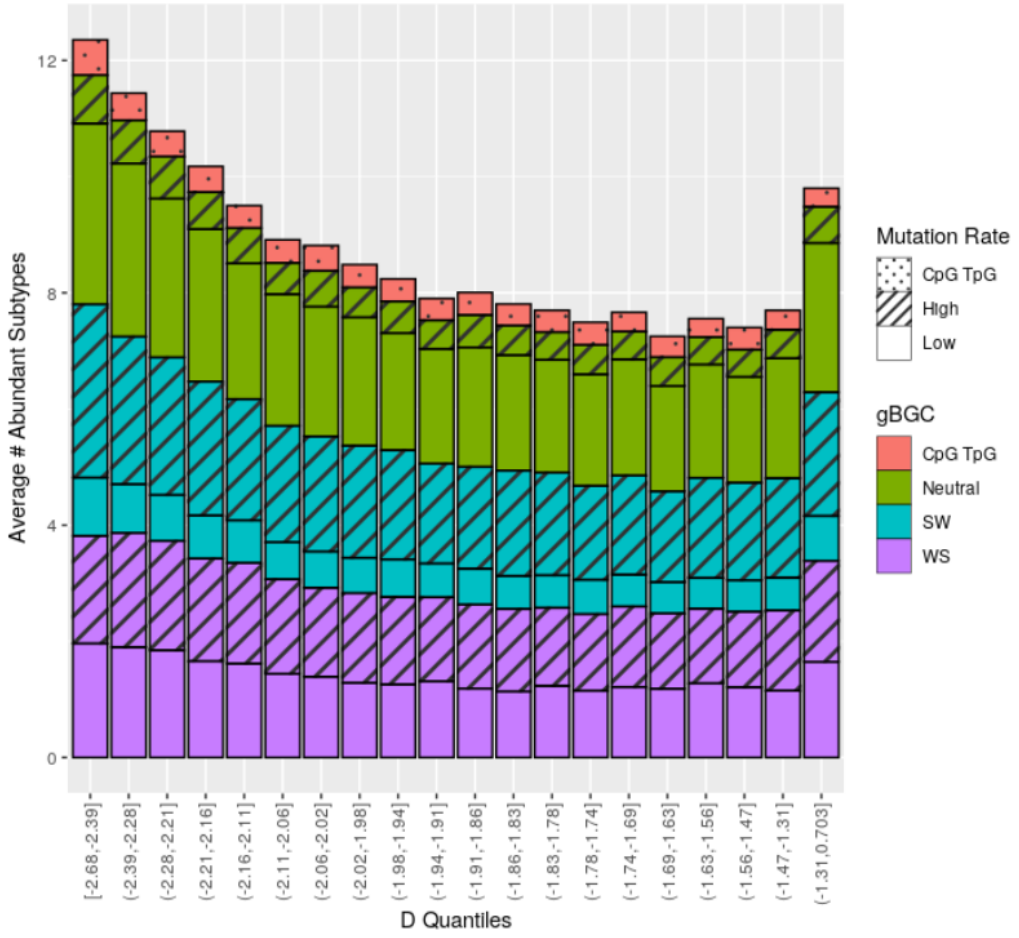

Figure 2.5 Average number of abundant subtypes in windows stratified by 5% Tajima's D quantiles. Number of abundant subtypes further broken down into direction of gene conversion (strong to weak vs weak to strong) and mutation rate (low vs high).

We compared four observed local AFS statistics (Tajima's D, % singletons, % doubletons, and % tripletons) to their expected value (see methods) to quantify the non-independence of local subtype composition and Tajima's D in shaping the regional AFS. The mean (standard deviation) for observed AFS statistic across windows for Tajima's D, % singletons, % doubletons, and % tripletons was -1.883 (0.335), 0.584 (0.037), 0.096 (0.015), and 0.039 (0.009). Similarly, expected local AFS statistics had means (standard deviation) of -2.036

(0.004), 0.603 (0.004), 0.099 (0.002), and 0.040 (0.001), with small standard deviations suggesting low variability in local subtype composition across 100Kb windows. When compared, the mean (standard deviation) of the difference between observed AFS statistic and expected AFS Statistic across windows for Tajima's D, % singletons, % doubletons, and % tripletons was 0.153 (0.335), -0.019 (0.038), -0.003 (0.015), and -0.001 (0.009) respectively (Figure 2.6). Comparable standard deviations in the differences to the observed values alone were consistent with small variability in the expected values, with mean differences half a standard deviation or less from zero suggesting a role of local subtype composition shaping the regional AFS. Fewer observed variants at the rarest frequencies than expected may be driven by local genomic factors such as late replicating regions that elevate mutation rates outside subtype composition[84,85].

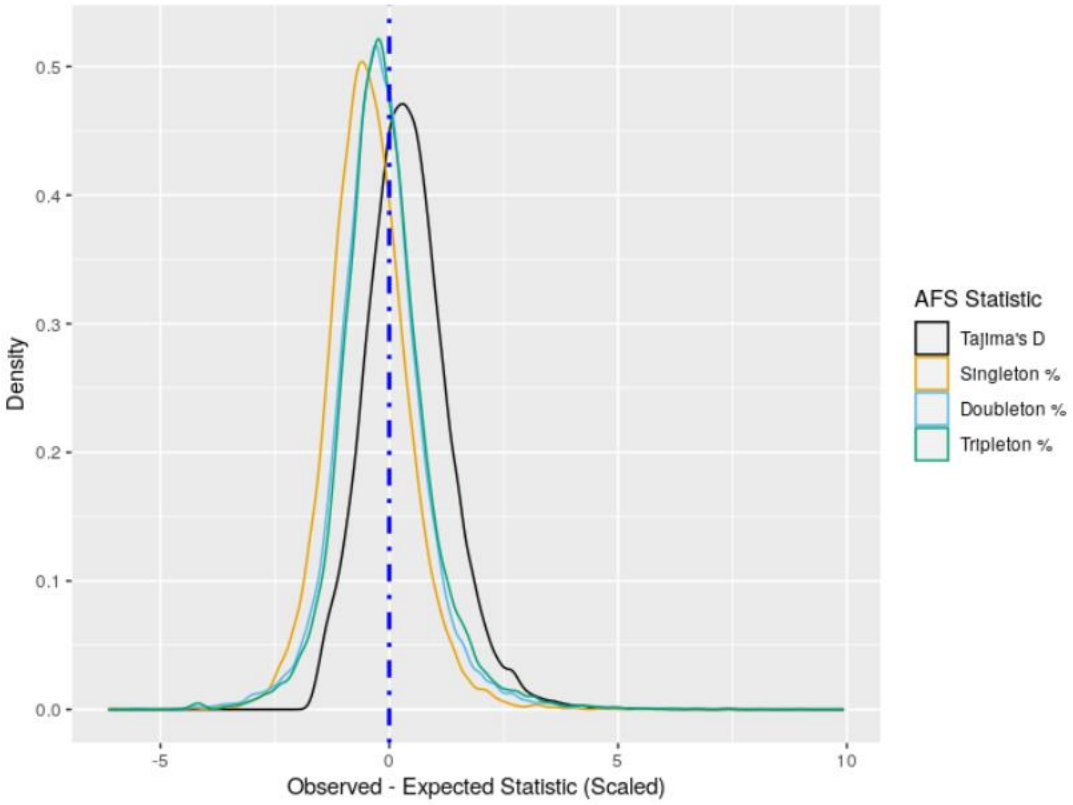

Figure 2.6 Difference in observed AFS statistics vs expected (MST genome wide statistic weighted by counts of sites in local window), standardized for comparison across statistics. Dotted blue line denotes zero, the difference if the local subtype composition perfectly determined the observed statistic.

From the GEE model directly regressing the expected local AFS statistic and local genomic factors on the observed local AFS statistic, we found each expected local AFS statistic was a significant predictor despite low variability in their values. A 0.10 increase in expected Tajima's D corresponded to a 0.285 (p=4.47e-03) increase in the observed Tajima's D. Similarly, a 0.10 increase in expected % singletons, % doubletons, and % tripletons corresponded to a -0.085 (p=1.85e-05), 0.105 (p=1.54e-15), and 0.100 (p=4.96e-04) change in the corresponding observed AFS statistic on average (Table 2.4). A negative coefficient for the expected % singletons was inconsistent, though could be explained by low variability in the expected values or negative correlations between CpG substitution rate and GC content (see supplementary for discussion). Furthermore, recombination rate was highly associated with the

observed Tajima's D (β=0.045, p=5.48e-49), % singletons (β=-0.006, p=1.24e-57), and % doubletons (β=0.000, p=1.92e-08), while GC percent was only associated in the observed Tajima's D model (β=-0.394, p=2.81e-04). Overall, significant associations for all expected local AFS statistics, recombination rate, and GC content on the observed local AFS statistic values suggest that local mutation subtype composition shapes the regional AFS in conjunction with local genomic factors.

2.4 Discussion

Our work uses large sample whole genome sequencing to assess how the AFS differs across variant subtypes and identify biological factors driving AFS heterogeneity. While previous work has studied AFS heterogeneity mainly among the six single base mutation types (A->C, A->G, etc), our results indicate increased heterogeneity when considering immediate flanking markers (especially for the rarest frequency singleton and doubleton variants). We further extend the effect of biological factors (hypermutability of certain motifs and biased gene conversion) on the AFS from 1-mer variants to 3-mers considering adjacent nucleotides. In particular, 3-mer motifs with higher mutation rates exhibited a lower ratio of singletons to doubletons while motifs, depending on direction of gene conversion, had an increase in either low or intermediate frequency variants as quantified by Tajima's D and our proposed D_{-2} statistic removing singletons and doubletons. While conclusions of signals for gene conversion were consistent with our D_{-2} statistic, we note attenuated differences across groups are likely explained by a loss of power as the singleton and doubleton counts are very informative in D-type statistics.

We first demonstrated the effect of AFS heterogeneity across subtypes on demographic inference through considering the case of a single subtype comprising the entire genome wide AFS. Under this scenario, inferred parameters differed drastically across subtypes and resulted in differing conclusions. Under the three-epoch population model, certain subtype specific AFS inferred a strong bottleneck or no bottleneck at all. Similarly, for an exponential growth model, growth rates varied drastically with a strong correlation between the singleton proportion and relative growth. CpG TpG sites inferred a lower relative growth due to having proportionally fewer singletons (driven by their higher mutation rates causing recurrent mutations). However, the subtypes A[A->G]G and T[A->T]A had a similarly smaller singleton proportion and also inferred lower growth. While CpG TpG sites are sometimes excluded from analysis as expected outliers[86], these are subtypes that would normally not be considered for exclusion.

Similarly, we found that regional AFS were also affected by AFS heterogeneity across mutation subtypes. Regions in the tails of the empirical distribution of Tajima's D tend to have more extreme composition of mutation subtypes than windows near the median, though we saw no evidence that this local composition of mutation subtypes across Tajima's D quantiles is driven by specific biological processes influencing the genome wide AFS (gBGC and mutation rate heterogeneity). Mean differences between expected and observed local AFS statistics within half a standard deviation or less of zero suggest local heterogeneity in subtype composition plays a role in shaping the regional AFS. Larger mean differences (relative to standard deviations) for Tajima's D and singleton proportion suggest the local subtype composition plays a relatively smaller role in shaping these statistics as compared to the doubleton and tripleton proportion. Furthermore, significant associations for expected statistics, recombination rate, and GC content

in our GEE model suggest a combination of local subtype composition and genomic factors jointly shape the regional AFS.

Several potential limitations need to be considered when interpreting these results. First, AFS heterogeneity is driven by multiple factors acting concurrently, causing the identification of mutation rate heterogeneity and biased gene conversion to potentially confound one another. While we aimed to mitigate this issue using our newly derived D_{-2} statistic that removed singletons and doubletons when assessing signals of gene conversion, the effect of recurrent mutations driven by higher mutation rates detectably extends to higher allele frequencies. Second, correlations between estimated relative mutation rates and statistics/estimates across subtypes may be confounded by the fact that mutation rates used were estimated from singletons even though estimated relative mutation rates are similar to rates estimated elsewhere. Lastly, our dataset used had a relatively lower average coverage (9.6x) given today's deep standards (>30x). While many extremely rare variants likely went undetected during variant calling, our stringent quality control procedure (see Carlson et al[47]) ensured analyzed variants across the distribution of allele frequencies were of high quality.

Despite potential confounding and sample limitations, our results thus demonstrate the challenges introduced by treating polymorphic sites as exchangeable in population genetic inference. From our genome-wide analysis, we can clearly see how model parameter estimates for demographic inference are sensitive to the subtype-specific AFS and suggest removing certain subtypes prior to analysis. While CpG TpG sites are sometimes already excluded due to their outlying nature[86], demographic models dependent on the singleton count should further consider excluding other subtypes with lower singleton proportions. A sensitivity analysis with and without subtypes removed can then reveal how much inference is being driven by said

subtypes. From our regional analysis, we can see that a combination of the local composition of subtypes and genomic factors play a role in shaping the regional AFS. Thus, we recommend approaches to identify regions under selection by identifying outlier local AFS statistics (i.e., Tajima's D) to perform a post hoc analysis and consider whether a local region with an outlier Tajima's D has a distribution of sites in the region comparable to the rest of the genome. Furthermore, we recommend considering whether the outlier D region is in a late replicating region (known to alter both the mutation rate and subtype composition [87]) or has an outlier recombination rate(as been suggested by [88]. If the region has a unique distribution of sites or is subject to one of the above specified local genomic factors, care may be needed in interpreting results of the analysis. Future work could potentially investigate adjusting either the overall AFS or the inference method itself to account for the composition of mutation subtypes in the sample prior to analysis.

Our findings about allele frequency heterogeneity imply that even non-AFS inference frameworks could bias inference by failing to differentiate between sites. For example, the coalescent-based methods PSMC[45] and the Singleton Density Score[89] treat sites as interchangeable. PSMC assumes a constant mutation rate in a window while the Singleton Density Score is reliant on distances to the nearest singleton, and thus both methods may be vulnerable to local regions being abundant in sites with outlying mutation rates or singleton proportions. As a result, we believe population genetics methods across multiple frameworks could benefit by carefully considering the unique evolution of mutation subtypes over time[90].

One benefit of the present era of population genetics is the availability of very large samples with deep genotyping. However, the same large sample sizes also amplify subtle population genetic effects that can mislead attempts at inference. As large samples provide an

abundance of variants available for population genetic inference, it is both feasible and advisable to assess the robustness of inference results to these effects and to adjust the inference accordingly.

2.5 Chapter 2 Appendix

2.5.1 Derivation of New D-2 Estimator

We sought to derive a Tajima's D type statistic that removed the contribution of singletons and doubletons. The traditional Tajima's D is formulated as the difference of two unbiased estimates of θ: Mean pairwise difference (θ_π) and Watterson's estimator (θ_W) based on the total number of segregating sites:

$$D = \frac{\theta_\pi - \theta_W}{\sqrt{var(\theta_\pi - \theta_W)}}$$

Where:

$$\theta_\pi = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \eta_i i(n-i)$$

$$\theta_W = \frac{S}{h_n}$$

Here S is the number of segregating sites, n is the haploid sample size, and $h_n = \sum_{i=1}^{n-1} \frac{1}{i}$.

For our novel D-2 statistic, the numerator takes the form of the traditional D estimator with the singleton and doubleton contributions subtracted from Mean Pairwise Difference and Watterson's Estimator. We then reweight the resulting estimators to both be unbiased estimators of θ.

1) MPD without singletons and doubletons:

Let $\pi^* = \binom{n}{2}^{-1} \{[\sum_{i=1}^{n-1} \eta_i i(n-i)] - \eta_1(n-1) - \eta_2 * 2(n-2)\}$

$$E[\pi^*] = E[\binom{n}{2}^{-1} \left\{\left[\sum_{i=1}^{n-1} \eta_i i(n-i)\right] - \eta_1(n-1) - \eta_2 * 2(n-2)\right\}]$$

$$= E\left[\pi - \binom{n}{2}^{-1}\eta_1(n-1) - \binom{n}{2}^{-1}\eta_2 * 2(n-2)\right]$$

$$= E[\pi] - \frac{n-1}{\binom{n}{2}}E[\eta_1] - \frac{2(n-2)}{\binom{n}{2}}E[\eta_2]$$

$$= \theta - \theta\frac{n-1}{\binom{n}{2}} - \frac{\theta(2(n-2))}{2 * \binom{n}{2}}$$

$$= \theta\left[1 - \frac{n-1}{\binom{n}{2}} - \frac{n-2}{\binom{n}{2}}\right]$$

$$= \theta\left[1 - \frac{2}{n} - \frac{2(n-2)}{n(n-1)}\right]$$

$$= \theta\left[\frac{n(n-1)}{n(n-1)} - \frac{2(n-1)}{n(n-1)} - \frac{2(n-2)}{n(n-1)}\right]$$

$$= \theta\left[\frac{n^2-5n+6}{n(n-1)}\right]$$

$$= \theta\left[\frac{(n-3)(n-2)}{n(n-1)}\right]$$

Thus

$$\theta_{\pi-2} = \left[\frac{n(n-1)}{(n-2)(n-3)}\right] * \binom{n}{2}^{-1}\left\{\left[\sum_{i=1}^{n-1}\eta_i i(n-i)\right] - \eta_1(n-1) - \eta_2 * 2(n-2)\right\}$$

Is an unbiased estimator of θ.

2) Watterson's estimator without singletons η_1 and doubletons η_2:

We know from coalescent theory that under a neutral population $E[\eta_1] = \theta, E[\eta_2] = \frac{\theta}{2}$.

Further $E[S] = h_n\theta = \sum_{i=1}^{n-1}\frac{1}{i}$.

Let $S^* = S - \eta_1 - \eta_2$

$$E[S^*] = E[S - \eta_1 - \eta_2]$$

$$= h_n\theta - \theta - \frac{\theta}{2}$$

$$= \theta(a_n - \frac{3}{2})$$

Thus

$$\theta_{W-2} = \frac{S - \eta_1 - \eta_2}{h_n - \frac{3}{2}}$$

Is an unbiased estimator of θ.

Our novel D-2 statistic then takes the form:

$$D = \frac{\theta_{\pi-2} - \theta_{W-2}}{\sqrt{var(\theta_{\pi-2} - \theta_{W-2})}}$$

For the denominator, we use results from Fu and Durett[69,70] to derive an analytical form. Let c_n^k be a weight vector of length n for estimator X_k. Test of neutrality statistics can be expressed in the general form as a weighted sum of the AFS:

$$X_k = \sum_i^{n-1} c_{n,i}^k \eta_i$$

Our modified Watterson's and MPD estimators retain the same functional form subtracting off singletons and doubletons. Thus, they can be expressed as a weighted sum where the first two weights are set to zero and the remaining are reweighted:

1. MPD expressed as weighted sum:

$$\theta_{\pi-2} = \left[\frac{n(n-1)}{(n-2)(n-3)}\right] * \binom{n}{2}^{-1} \left\{\left[\sum_{i=1}^{n-1} \eta_i i(n-i)\right] - \eta_1(n-1) - \eta_2 * 2(n-2)\right\}$$

$$\theta_{\pi-2} = \sum_i^{n-1} c_{n,i}^1 \eta_i$$

$$where\ c_n^1 = [0, 0, \frac{2i(n-i)}{(n-2)(n-3)}, \ldots\ldots,]$$

2. Watterson's expressed as weighted sum :

$$\theta_{W-2} = \frac{S - \eta_1 - \eta_2}{h_n - \frac{3}{2}}$$

$$\theta_{W-2} = \sum_{i}^{n-1} c_{n,i}^2\, \eta_i$$

$$where\ c_n^2 = [0, 0, \frac{1}{h_n - \frac{3}{2}}, \ldots\ldots,]$$

Fu derived the covariance for weighted sums of the AFS as:

$$cov(X_1, X_{2)} = a_n\theta + b_n\theta^2$$

$$a_n = \sum_{i=1}^{n-1} \frac{c_{n,i}^1 c_{n,i}^2}{i}$$

$$b_n = \sum_{i,j} c_{n,i}^1\, \sigma_{ij} c_{n,j}^2$$

$$\sigma_{ij} \text{ with } i > j = \begin{cases} B_n(i+1) - B_n(i) & i + j < n \\ \frac{h_n - h_i}{n-i} + \frac{h_n - h_j}{n-j} - \frac{B_n(i) + B_n(j+1)}{2} - \frac{1}{ij} & i + j = n \\ \frac{B_n(j) - B_n(j+1)}{2} - \frac{1}{ij} & i + j > n \end{cases}$$

$$where\ B_n = \frac{2n}{(n-i+1)(n-i)}(h_{n+1} - h_i)$$

$$where\ h_n = \sum_{i}^{n-1} \frac{1}{i}$$

Thus, we can derive the analytical form for the variance of our estimator:

$$var(\theta_{\pi-2} - \theta_{w-2}) = var(\theta_{\pi-2}) + var(\theta_{w-2}) - 2cov(\theta_{\pi-2}, \theta_{W-2})$$

$$= cov(\theta_{\pi-2}, \theta_{\pi-2}) + cov(\theta_{W-2}, \theta_{W-2}) - 2cov(\theta_{\pi-2}, \theta_{W-2})$$

Where each term can be expressed using Fu's closed form of the covariance described above and the corresponding weight vector c_n^k:

a. $cov(\theta_{\pi-2}, \theta_{\pi-2}) = a_n\theta + b_n\theta^2$ with $c_n^1 = c_n^2 = [0, 0, \frac{2i(n-i)}{(n-2)(n-3)}, \ldots,]$

b. $cov(\theta_{W-2}, \theta_{W-2}) = a_n\theta + b_n\theta^2$ with $c_n^1 = c_n^2 = [0, 0, \frac{1}{h_n - \frac{3}{2}}, \ldots,]$

c. $cov(\theta_{\pi-2}, \theta_{W-2}) = a_n\theta + b_n\theta^2$ with $c_n^1 = [0, 0, \frac{2i(n-i)}{(n-2)(n-3)}, \ldots,]$, $c_n^2 = [0, 0, \frac{1}{h_n - \frac{3}{2}}, \ldots,]$

Each of the above covariances requires knowledge of θ and θ^2 which are not known. Instead, we use estimates derived from Watterson's estimator, similar to the original Tajima's D statistic[70]:

$$\hat{\theta} = \frac{S}{h_n}$$

$$\hat{\theta}^2 = \frac{S^2 - S}{h_n^2 - g_n}$$

$$where\ g_n = \sum_i^{n-1} \frac{1}{i^2}$$

To assess the null distribution and type 1 error, we used fastsimcoal2 to simulate 2,000 neutral frequency spectra for two subtypes: A[A->C]A and A[C->T]G assuming no recombination. For each run, we simulated a 1Mb sequence passing as parameters the subtype's absolute mutation rates and effective population size (23,785 and 23,066 respectively, derived using Watterson's estimate of θ and $\theta = 4N_e\mu L$ where L is the number of subtype motifs across the genome). (Figure S2). Across simulations for both subtypes, the mean D_{-2} value was roughly zero (-0.103 and -0.048) with T1E rates of 0.031 and 0.0375 (roughly assuming a

standard normal under the null). We note the null distribution exhibited longer tail behavior as compared to the roughly standard normal that Tajima's D follows.

2.5.2 Negative Relationship Between Expected and Observed Singleton Proportions Across 100Kb Windows

In our 100Kb window analysis to determine whether the local subtype composition plays a role in shaping the regional AFS, we surprisingly observed a negative coefficient in our GEE regression model for the singleton proportion (Table S2). Plotting the expected singleton proportion vs observed singleton proportion confirms a negative trend (Figure S4). This may be caused by low variability in the expected singleton proportions [0.58, 0.61], which may obscure an actual positive relationship between values. Another possible explanation is GC content is known to be negatively correlated with CpG substitution rate correlations[91,92]. Regions of high GC content would have lower mutation rates than expected (based on local mutation subtype composition) and could counteract the effect of recurrent mutation and thus increase the singleton proportions. While GC content was included in our model, we may not have sufficiently captured the complexity between substitution rates, methylation, and GC content.

2.5.3 Code Availability

A GitHub repository containing all data and necessary code to reproduce results is available at: https://github.com/kliao12/AFS_subtype_analysis

2.5.4 Supplementary Tables and Figures

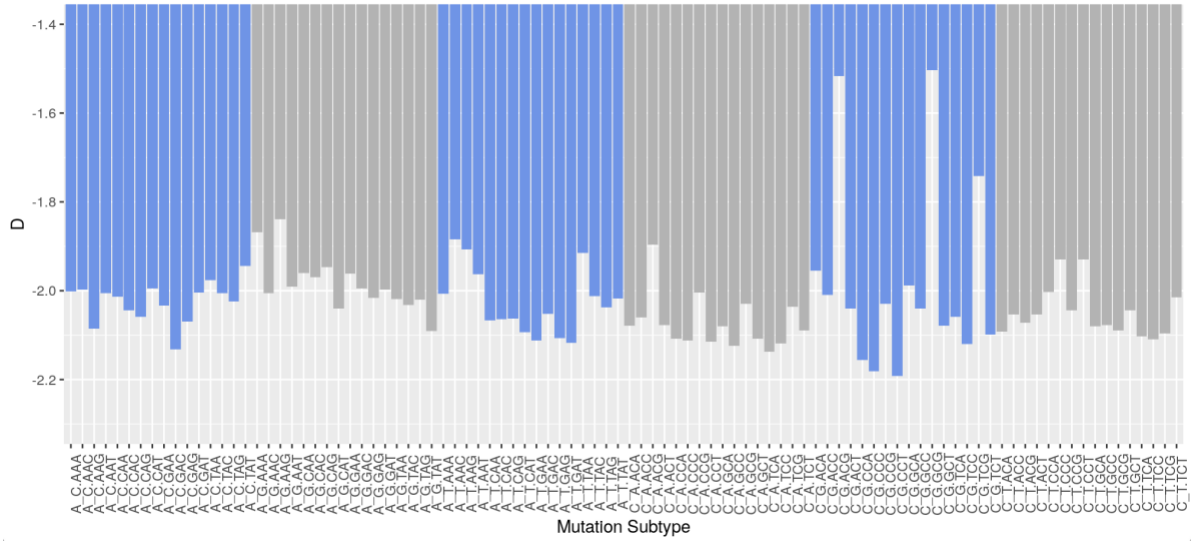

Figure 2.7 Bar plot showing Tajima's D computed for each of the 96 mutation subtypes' genome-wide allele frequency spectrum. Negative values across subtypes are consistent with recent explosive human population growth.

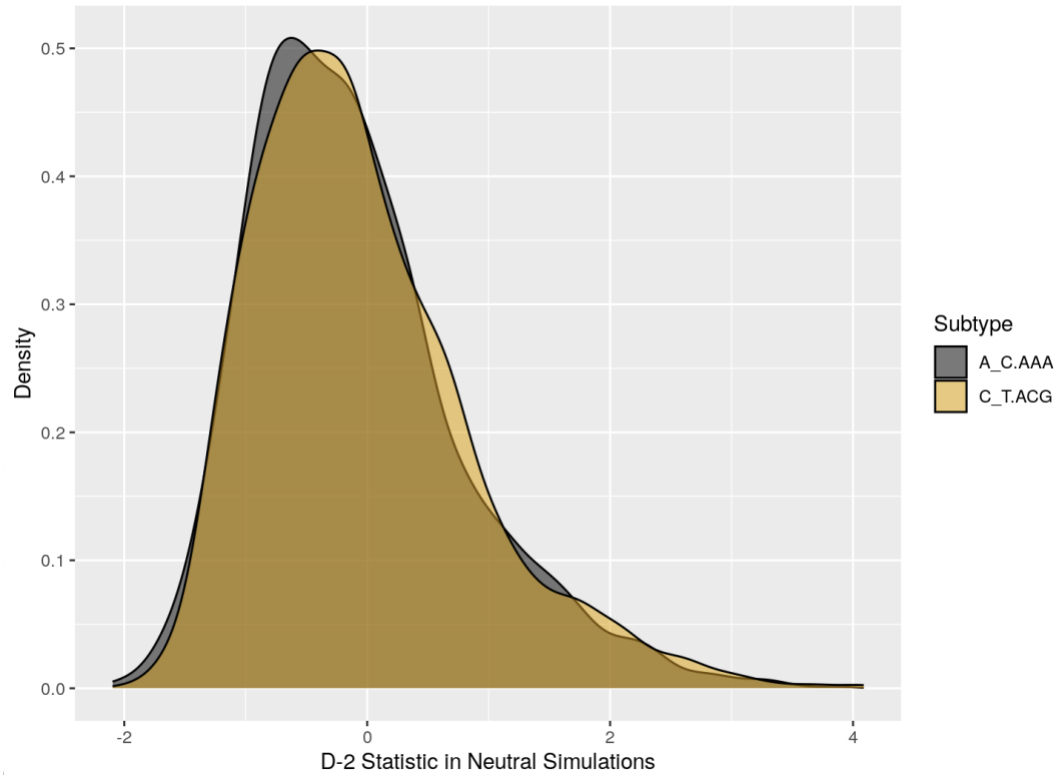

Figure 2.8 Null distribution for D-2 statistic across two subtypes: A[A->C]A and A[C->T]G. For each subtype, we simulated 2,000 neutral AFS using Fastsimcoal2 (see supplementary for details).

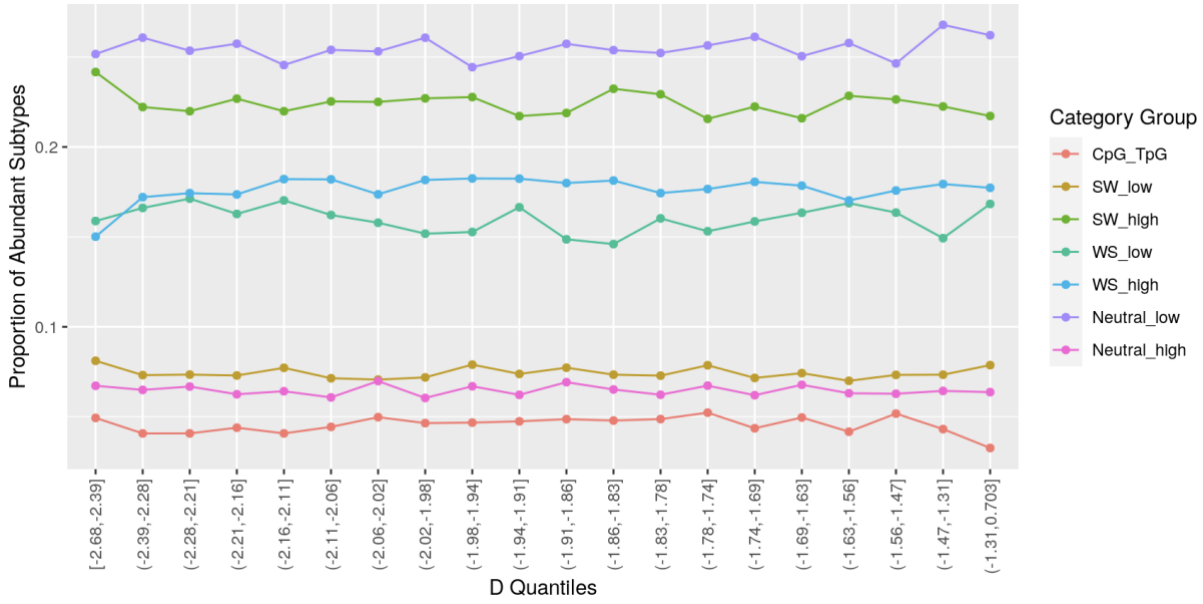

Figure 2.9 Line graph showing proportion of abundant subtypes in each Tajima's D quantile broken down by biased gene conversion x mutation rate heterogeneity category.

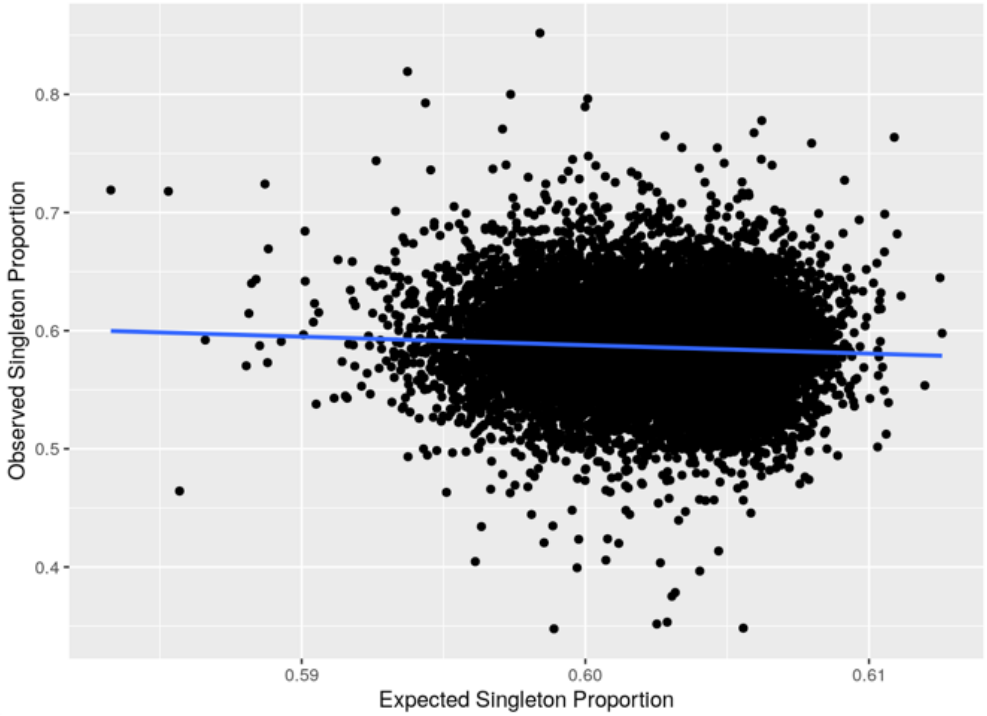

Figure 2.10 Scatter plot of negative relationship between expected and observed singleton proportion across 100Kb windows.

MT	Correlation	P-Value
A->C	-0.21	0.44
A->G	-0.61	0.01
A->T	-0.26	0.33
C->A	-0.66	0.01
C->G	-0.54	0.03
C->T	-0.97	4.2e-10

Table 2.3 Correlation and p-value between single-derived mutation rates and singleton to doubleton ratio. Each mutation type has 16 distinct 3-mer subtypes in which correlation was computed.

Covariates	Observed Statistic (Dependent variable)			
	Model 1: Observed D	Model 2: Observed Singles	Model 3: Observed Doubles	Model 4: Observed Triples
Expected Statistic Estimate (p value)	2.853 (**4.469e-03**)	-0.845 (**1.852e-05**)	1.045 (**1.536e-15**)	1.003 (**4.96e-04**)
Recombination Rate Estimate (p value)	0.045 (**5.479e-49**)	-0.006 (**1.237e-57**)	0.000 (**1.921e-08**)	0.000 (0.573)
GC Percent Estimate (p value)	-0.394 (**2.812e-04**)	0.019 (3.719e-01)	-0.008 (3.132e-02)	-0.004 (0.183)

Table 2.4 Regression output (β estimates and p values) from GEE analysis modeling observed local AFS statistics with expected statistics (defined as weighted mean of genome wide values, using counts of subtypes as weights) and adjusting for recombination rate and

Chapter 3 A Stacking Framework for Polygenic Risk Prediction in Admixed Individuals

3.1 Introduction

Since the first genome wide association study (GWAS) published in 2005, GWAS have successfully implicated thousands of risk variants across a variety of traits[93]. While a single risk variant may only explain a small percent of a trait's heritability, a sizable proportion of phenotypic variation can be explained by summarizing an individual's genetic risk for a given disease or trait in polygenic risk scores[17] (PRS). PRS are typically computed as a weighted sum of risk alleles using estimated effects from an external GWAS as weights. These PRS have been used[18,19] to identify individuals at high risk of disease, improve diagnostic accuracy, and allow for tailored personalized treatment for disease risk prediction in complex traits including coronary artery disease[20,21], type 1 and 2 diabetes[22,23], breast cancer[24,25], and more[26]. However, PRS fail to capture the full variability expected from heritability estimates while also being susceptible to environmental confounding and indirect genetic effects such as assortative mating[94–96].

Furthermore, performance of a PRS in predicting a phenotype for a target sample can be ancestry dependent. In particular, PRS prediction performance decays as genetic divergence increases between the target sample of interest and external GWAS.[97,98]. This performance decay can mainly be attributable to 1) differences in allele frequencies and 2) differences in both marginal and causal effect sizes of variants across populations[30]. Causal effect sizes themselves can differ across populations due to unique environments and demography, though recent work

in admixed individuals has suggested causal effect sizes are shared across populations[99]. However, even when causal effects are shared, marginal estimated GWAS effect sizes can still differ due to differences in linkage disequilibrium (LD) tagging the true causal variant. The extent in how LD differs across populations varies along the genome[100], prompting work in the transferability of PRS across diverse populations to often consider a local approach in combining genetic evidence[101,102]. Specifically, approaches often model local population-specific LD patterns in regions to better identify true local risk variants and increase effective sample size[102].

The ancestry dependence of PRS is further exacerbated in the context of admixed individuals. Historically, genetic studies group admixed individuals of varying ancestry proportions into a single ancestral label such as "African American" or "Hispanic". However, the genetic ancestry in an admixed sample varies across both individuals and regions prompting a recent push to consider ancestry on a continuum rather than as discrete ancestral groups[103]. In admixed individuals, Bitarello and Mathieson showed predictive accuracy of a PRS for height using European summary statistics increased linearly with global European ancestry proportion across various datasets[104]. Similarly, Cavazos and Witte showed in simulations a similar linear relationship with both European and African summary statistics performing better as the proportion of European and African ancestry respectively increased across admixed samples[105]. Such ancestry dependence of PRS in admixed individuals is problematic even if all ancestral groups have predictive PRS, as admixed individuals that have most of their genetic ancestry from one parental group will benefit more from potential downstream clinical utility of PRS than groups with equal contribution from both ancestries. Even developing PRS specifically

for the admixed group will not ameliorate this problem as such a PRS will only work well for admixed individuals with admixture proportions similar to the group mean. While the field of genetics has acknowledged and begun making strides in addressing inequity in genomic research[8,106], development of methods to construct well-performing PRS free of ancestry dependence in admixed samples is needed.

To overcome the ancestry dependence of PRS performance using a single population GWAS in admixed samples, recent work has proposed methods that leverage GWAS summary statistics from the multiple ancestral populations of an admixed sample. Incorporating GWAS effect sizes from multiple populations provides many benefits, including identifying population specific risk variants and boosting sample size if risk variants are shared. In admixed African Americans, methods have been proposed that 1) consider local ancestry by matching chosen risk variants with an individual's local ancestry at that position[104,107] and 2) ignore local ancestry and construct a joint PRS as a linear combination of global European and African PRS[108]. In simulations, Cavazos and Witte conducted a comprehensive review of both approaches[105]. While the first approach, deconvoluting ancestry and matching risk variants on population-specific GWAS effect sizes, was initially suggested to perform well[107], this result failed to consistently replicate as shown in Cavazos' simulations and Bitarello's real data application[105,107,108]. Surprisingly, the second approach using a linear combination of global European and African PRS was found to improve prediction across a range of European ancestry quantiles in admixed African American individuals. However, use of global population specific PRS ignores the unique local admixture present in any given region within a sample of admixed individuals, missing potential population specific risk variants in a region or local GxG interactions on a specific ancestral background. Thus, it is possible that performance of local population specific

PRS (i.e., a PRS using only risk variants in a genomic region and a specific population GWAS effect sizes) will vary across admixed individuals.

In this work we propose slaPRS (**s**tacking **l**ocal **a**ncestry **PRS**), a novel stacking framework to construct admixed PRS for quantitative traits that combines local population specific PRS constructed using population specific effect sizes in local genomic regions. Stacking is an ensemble machine learning method that aims to optimize prediction accuracy by combining separate prediction models[109,110]. In target samples of a single ancestry, Prive et al successfully used stacking to optimize the commonly used clumping and thresholding (C+T) PRS method by deriving a linear combination of PRS across all possible parameters, rather than learning a single set of optimal parameters[111]. Outside of PRS construction, stacking has been used in other genetic methods such as the recent REGENIE method for GWAS that improved computational efficiency by orders of magnitude through conditioning on the predicted individual trait values constructed from combining local polygenic risk predictors[112]. In our approach, we first divide the genome into windows of a predetermined size and in each local window compute population specific local PRS using the respective population specific GWAS effect sizes via C+T. In training data, we then fit a penalized regression model to combine local population specific PRS across the genome to determine unique weights that are used to predict the phenotype in testing data. We show in extensive simulations and analysis of admixed African Americans and African British that slaPRS removes the ancestry dependence of PRS performance present in traditional single-population GWAS PRS and outperforms or compares similarly to existing methods in an efficient data-driven process.

3.2 Methods

Consider a sample of N admixed individuals with ancestral contributions from population A and B (slaPRS is not restricted to two-way genetic admixture but is assumed here for notational simplicity). Let **X** be the NxM admixed genotype matrix (M is the total number of variants genome wide) and **Y** the $Nx1$ phenotype vector. Let L_{ij} be an NxM matrix denoting the haplotype-level local ancestry (l_{ij1}, l_{ij2}.) of individual i at marker j. We assume the phenotype can be expressed as:

$$Y_i = \sum_{j=1}^{m} X_{ij} f(\beta_{A_j}, \beta_{B_j}, L_{ij}) + \boldsymbol{\epsilon}_i$$

Where X_{ij} is the genotype dosage for individual i at marker j, and β_{A_j}, β_{B_j} are effects for marker j on the phenotype in populations A and B respectively. Here, $f(\beta_{A_j}, \beta_{B_j}, L_{ij})$ is a weighted average of population specific GWAS effect sizes and local ancestry (see supplementary for derivation):

$$f\left(\beta_{A_j}, \beta_{B_j}, L_{ij}\right) = \beta_{A_j}\left(w_{k,\beta_{A_j}} + w_{k,L_{ij}}^{(A)} L_{ij}\right) + \beta_{B_j}\left(w_{k,\beta_{B_j}} + w_{k,L_{ij}}^{(B)} L_{ij}\right)$$

Where $w_{k,\beta_{A_j}}$ and $w_{k,L_{ij}}^{(A)}$ (and similarly for population B) are weights for population A effect sizes β_{A_j} and local ancestry interaction in each genomic region k that are learned via ensemble learning (stacking) in the slaPRS framework (see details below).

3.2.1 slaPRS Framework

We developed slaPRS for constructing admixed PRS using three main features: 1) a local window approach 2) local population specific PRS and 3) an ensemble stacking framework to combine local population specific PRS. For slaPRS, we assume existence of GWAS effect size

estimates for each ancestral population in an admixed population and the admixed genotype matrix. We first partition the admixed genotype matrix into K non-overlapping genotype blocks $G = \{G_1, G_2, \dots, G_K\}$ with blocks predefined by physical distance. In our analysis we considered blocks spanning 1Mb and 5Mb of physical distance, each with m_k SNPs such that $\sum_{j=1}^{K} m_k = M$.

Level 0 Local Population-Specific PRS and Ancestry

In the training set of admixed individuals, in each block G_k across the genome (using the m_k SNPs in the block) we first separately computed vectors of local population A PRS (A_k) and local population B PRS (B_k) using clumping and thresholding (C+T). While C+T was used in slaPRS, any PRS construction method could be used in our framework. In this step, each block's C+T optimized ancestry PRS can be viewed as a level 0 model prediction to be stacked in our stacking framework (Figure 3.1). Clumping first removes variants in strong LD with others using in-sample LD for that region, while greedily retaining the most significant variants[14]. Varying p-value thresholds $p = \{5e-2, 5e-4, 5e-6, 5e-8\}$ were considered (cross validation in Level 1 stacking model used to select optimal p to use in testing set) to construct ancestry-specific local PRS in each block using the respective population's estimated effect sizes. In this step, we make no assumption on whether risk variants are shared across ancestral populations, and thus local PRS A_k and B_k can have varying risk variants.

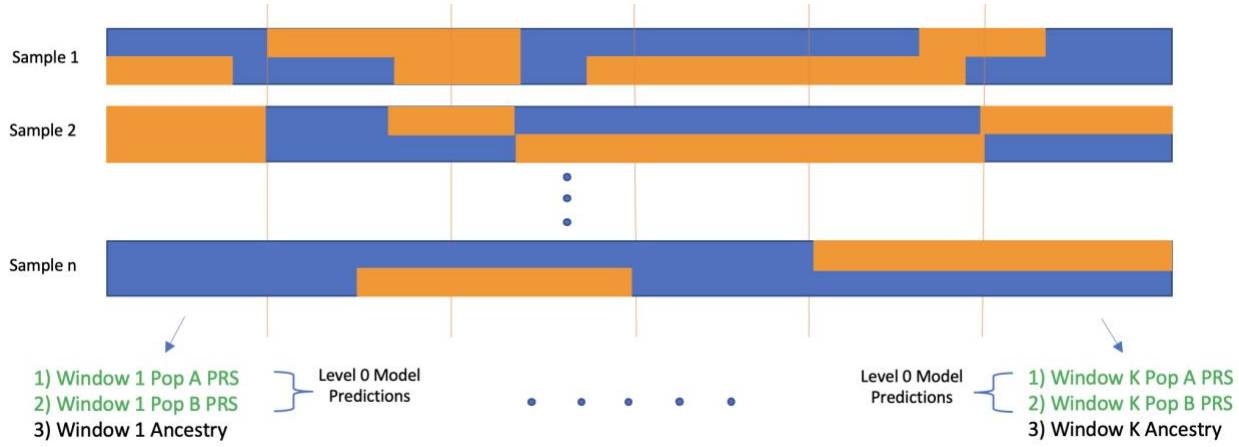

Figure 3.1 Diagram of local window and level 0 population specific PRS model predictions. Admixed genomes split into 5Mb windows and in each window a local population A and B PRS are computed using population-specific effect sizes. Local ancestry further computed to form covariate vector for level 1 stacking model.

For each sample, we computed the $Nx1$ vector of local ancestries Anc_k in block B_k as the % of population A ancestry: $Anc_k = \frac{\sum_{j=1}^{m_k} L_{ij}}{m_k}$. We constructed interaction terms $A_k * Anc_k$ and $B_k * Anc_k$ to allow for the effect of the local population PRS A_k and B_k to vary by a given ancestry. Following completion of level 0 in our framework, block k has the covariates (Figure 3.1):

$$C_{B_k} = [A_k, B_k, Anc_k, A_k * Anc_k, B_k * Anc_k]$$

After aggregating the B total local block covariates across the genome, let C be the N x $(k\ x\ 5)$ matrix:

$$C = [C_{B_1}, C_{B_2}, \dots, C_{B_k}]$$

Level 1 Elastic Net Stacking Model

We then trained an elastic net[113] penalized regression model to stack the local level 0 predictions (local population-specific PRS and ancestry) across the genome. The population's GWAS that optimizes the local PRS can vary across the genome (see introduction) in an

admixed sample, and stacking provides a data driven approach to inform which population's local PRS should be upweighted or shrunk. We used elastic net, which combines ridge regression[114] and LASSO[115], because the genetic architecture of a trait is unknown a priori (unknown which local blocks harbor causal risk variants and the distribution of local block heritability). When most local windows are weakly informative, ridge tends to have higher prediction accuracy while LASSO would likely outperform when only a small number of local windows are highly informative. Elastic net allows a data-adaptive approach to inform the amount of shrinkage and whether shrinkage patterns should favor ridge or LASSO to best accommodate a trait's genetic architecture.

To determine which aspects of our stacking framework drives increases in PRS performance, we considered three level 1 elastic net stacking models that vary in the covariates included from block B_k:

1) Local population A PRS only

$$C_{B_k} = \{A_k\}$$

2) Local population A and B PRS only

$$C_{B_k} = \{A_k, B_k\}$$

3) Local population A and B PRS, Ancestry and Interactions

$$C_{B_k} = \{A_k, B_k, Anc_k, A_k \times Anc_k, B_k \times Anc_k\}$$

Model 1 considered only local population A PRS A_k to investigate how stacking local PRS alone improves compared to a global population A PRS. Model 2 added local population B PRS B_k to assess the benefit of adding population B GWAS information, while Model 3 further included ancestry and interaction terms to allow for the effect of a local population specific PRS to vary

based on ancestral background. Total covariates in each proposed level 1 model aggregate covariates C_{B_k} across all blocks genome wide.

For each considered model, we fit a level 1 elastic net model[113] to combine the level 0 ancestry-specific PRS and additional covariates across the genome.

$$Y = w_0 + \boldsymbol{w_1} C_{B_1} + \boldsymbol{w_2} C_{B_2} + \cdots + \boldsymbol{w_k} C_{B_k}$$

Where $\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_k}$ are vectors of regression coefficients from the covariates in C_{B_k}. Estimates of $\widehat{\boldsymbol{w_k}}$ in the above model given the genome wide covariate matrix are obtained by minimizing the penalized objective function with respect to β:

$$\widehat{w(\lambda)} = argmin_w([\sum_{i=1}^{n}(y_i - C_i w)^2] + [\lambda(\alpha \sum_{j=1}^{k}|w_j| + (1-\alpha)\sum_{j=1}^{k} w_j^2])$$

The parameter λ determines the amount of shrinkage in model coefficients while $\alpha \in [0,1]$ balances the L1 and L2 penalty from ridge regression ($\alpha = 0$) and LASSO ($\alpha = 1$). To optimize all parameters including the p-value threshold $p = \{5e-4, 5e-6, 5e-8\}$ used in constructing level 0 local ancestry PRS via C+T, $\alpha = \{0, 0.1, 0.2, \ldots, 1\}$, and $\lambda = \{10^{-3}, \ldots, 10^{3}\}$, we employed K-fold cross validation with 10 folds and selected the set of p, α, and λ that produced the lowest adjusted R^2.

Estimates of $\boldsymbol{w_k}$ for each block across the genome can be used (see supplementary for derivation) to express the weight for each variant in PRS construction as a linear combination of population A (β_{A_j}) and B (β_{B_j}) GWAS effect sizes and learned block weights:

$$Y_i = \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} X_{ij} f(\beta_{A_j}, \beta_{B_j}, L_{ij}) + \boldsymbol{\epsilon}_i$$

$$f\left(\beta_{A_j}, \beta_{B_j}, L_{ij}\right) = \beta_{A_j}\left(w_{k,\beta_{A_j}} + w_{k,L_{ij}}^{(A)} L_{ij}\right) + \beta_{B_j}\left(w_{k,\beta_{B_j}} + w_{k,L_{ij}}^{(B)} L_{ij}\right)$$

Where $w_{k,\beta_{A_j}}$ and $w_{k,L_{ij}}^{(A)}$ (and similarly for population B) are weights for population A specific local PRS A_k and its local ancestry interaction term.

Once weights from the level 1 elastic net stacking models have been estimated from the training data, in testing data we then computed the same level 0 model predictions and covariates in each block and aggregated genome wide:

$$C = [C_{B_1}, C_{B_2}, \ldots, C_{B_k}]$$

Where C_{B_i} is defined as one of the three considered level 1 models. We then predicted trait values using estimated weights from the elastic net model:

$$\widehat{PRS} = C\hat{\beta}$$

The estimated PRS is then tested against simulated phenotypes or trait values in real data.

Genotype, Phenotype, and Population-Specific GWAS Simulation

For our simulations and real data applications we focused on admixed African Americans/British with European and African ancestral backgrounds. To simulate genotype and phenotype data for an African and European population with realistic allele frequencies and linkage disequilibrium patterns, we used the coalescent-based pipeline as described by Martin et al[98] and Cavazos et al[98,105]. Using msprime[116] with an out-of-Africa demographic mode modeling HapMap[117] chromosome 20 haplotypes, we simulated n=10,000 European samples and varying African sample sizes n={2000, 5000, 10,000}. Simulated population specific genotypes were then used to estimate marginal variant effect sizes.

We then simulated quantitative trait phenotypes using the simulated genotypes. We first assumed complete transethnic sharing of genetic architecture across African and European populations, in which true causal variants, causal effect sizes, and overall heritability are

consistent across populations. Under this scenario, performance of estimated PRS should vary only because of differences in allele frequency and LD across population. We subset variants with minor allele frequency > 5% in both populations and randomly sampled m={100, 500} shared causal variants. True causal effect sizes were drawn from a normal distribution $\beta \sim N(0, \frac{h^2}{m})$ where $h^2 = \{0.10, 0.30\}$ is the SNP-based heritability. In results, we focused on the most realistic simulation scenario consisting of $h^2 = 0.10$ and $m = 100$. We then considered the simulation scenario in which genetic architecture differs across ancestral populations by assuming true causal variant locations and overall heritability are shared, but now simulating causal effects $\boldsymbol{\beta} \sim MVN(\mathbf{0}, \begin{pmatrix} \frac{h^2}{m} & \frac{\rho h^2}{m} \\ \frac{\rho h^2}{m} & \frac{\rho h^2}{m} \end{pmatrix}$ varying transethnic genetic correlation $\rho = \{0.20, 0.50, 0.80\}$.

In both simulation scenarios, the true genetic score G was then defined as the product of sampled causal genotypes and their respective simulated effect sizes ($g = \sum_{j=1}^{m} X_j \beta_j$), standardized to ensure total heritability of h^2: $G = \frac{g - \mu_g}{\sigma_g} * h^2$. We then simulated the environmental effect from a normal distribution with variance comprising the remaining phenotype variance $\epsilon \sim N(0, 1 - h^2)$ and similarly standardized: $E = \frac{\epsilon - \mu_\epsilon}{\sigma_\epsilon} * (1 - h^2)$. We defined phenotype data Y for both populations as the sum of the standardized true genetic score and environmental effect $Y = G + E$. We then estimated effect sizes $\hat{\beta}$ for each variant genome wide using a linear model $Y = XB + \epsilon$, using each population's respective simulated phenotype and genotype data.

We additionally simulated n=1,000 European and n=1,000 African founder samples to simulate n=10,000 admixed African Americans genotypes via RFMix[118] with s=12 generations

of admixture for training and testing slaPRS. Simulated admixed genotypes had known phase and known local ancestry. We followed the same pipeline described above to generate the phenotype given the simulated genotypes. In the scenario where causal effects differed across populations, we considered haploid chromosomes H_{ij1} and H_{ij2}(corresponding haplotype 1 and 2 for individual i at variant j) and matched the population specific effect sizes on the local ancestry of a variant's haplotype background to derive the true genetic component: $X_i = \sum_{j=1}^{m} \beta_{j,AFR} \left[H_{ij1} I(l_{ij1} = AFR) + H_{ij2} I(l_{ij2} = AFR)\right] + \beta_{j,EUR} \left[H_{ij2} I\left(l_{ij2} = EUR\right) + H_{ij2}(l_{ij2} = EUR)\right]$. To prevent issues of overfitting, we split our sample into testing and training data using a 70:30 split, resulting in n=7000 and n=3000 admixed samples in the training and testing data splits. The outlined simulation procedure was repeated 150 times to evaluate slaPRS and perform method comparisons.

3.2.2 Comparison of Methods:

Clumping and Thresholding (C+T)

We first compared the proposed slaPRS method against global single population PRS, PRS_{EUR} and PRS_{AFR}, constructed using clumping and thresholding (C+T) with GWAS effect sizes from the respective population separately. In the C+T algorithm, we first clumped SNPs using each population's GWAS effect sizes with a window size of 250Kb and linkage threshold r^2 =0.10 and then optimized the threshold parameter in the 70% training set with $-\log_{10}(p)$ p value thresholds including $\{1, 2, \dots, 8\}$. The threshold that optimized PRS performance was then used in the 30% testing set to retain clumped risk variants to include in the PRS construction.

Linear Combination of Global Population Specific PRS

The second approach compared against was the method proposed by Marquez et al[108] which constructed a PRS as a linear combination of two global population-specific PRS:

$$PRS_{Marquez} = \alpha_{EUR} PRS_{EUR} + \alpha_{AFR} PRS_{AFR}$$

Here, PRS_{EUR} and PRS_{AFR} are the same global PRS constructed using C+T and the respective population GWAS as described above. To estimate the mixing weights (α_{EUR}, α_{AFR}) and global polygenic risk scores (PRS_{EUR}, PRS_{AFR}), we followed proposed guidelines and used cross validation. The 70% training set of admixed samples was first split in half, where the first half was used to estimate the thresholding parameter in the C+T algorithm. In the second half we constructed PRS_{EUR} and PRS_{AFR} using the optimal p-value threshold from the European GWAS (as is typically larger), as done by Marquez et al. In this same second half of the training set, we then estimated α_{EUR} and α_{AFR} by finding the least squares estimates to:

$$Y = \alpha_{EUR} PRS_{EUR} + \alpha_{AFR} PRS_{AFR}$$

With the optimal p-value threshold and mixing weights α_{EUR} and α_{AFR} derived from training data, we then constructed $PRS_{Marquez}$ as the weighted sum of PRS_{EUR} and PRS_{AFR}.

3.2.3 Quantifying Performance of Estimated PRS

To quantify and compare performance of each PRS across methods, we computed the proportion of variance explained (i.e. adjusted R^2) of the simulated quantitative phenotype with the estimated PRS adjusting for % European ancestry. Because one of our main objectives is to create a PRS with performance independent of the global ancestry of an admixed individual, we further stratified our adjusted R^2 performance metric by European ancestry quantiles [0-20%, 20-40%, 40-60% and 60-80%, 80-100%]. We also compared the mean simulated phenotype

value in the top 10% PRS quantile with the bottom 10% PRS quantile to assess the PRS' ability to identify high-risk and low-risk individuals.

3.2.4 Real Data Application

We evaluated slaPRS in real data applications using n=20,262 admixed African British individuals in the UK Biobank[6]. To choose samples, we selected admixed samples falling on the diagonal between the European and African corners of the PC plot (Figure 3.6). We used autosomal imputed genotypes in constructing polygenic risk scores. Phenotype data included the lipid biomarkers LDL, HDL, and total cholesterol. Lipid biomarker phenotypes were chosen because the Global Lipids Genetic Consortium[31] has collected large sample (excluding UK Biobank samples) ancestry specific GWAS data in Europeans (n=1.32 million) and Admixed African or Africans (N=99.4k). For all 20,262 samples we inferred local ancestry with genotypes first phased using BEAGLE 5.0[119]. We used RFMix[118] to infer local ancestry using phased haplotypes from European and African subpopulations from 1000 Genomes[3] individuals as references. From inferred local ancestry, we further computed global ancestry using tract lengths for sample stratification. We split the admixed dataset into 70% training and 30% testing for model training and method comparison.

Because the true PRS is unknown in real data, to quantify PRS performance across methods we computed the proportion of variance explained (adjusted R^2) between the estimated PRS and phenotypic value (instead of true genetic score) from the model including the first 4 principal components:

$$Y = \beta_0 + \beta_{PRS}PRS + \beta_{PC_1}PC_1 + \cdots + \beta_{PC_4}PC_4$$

Similar to simulations, we computed adjusted R^2 across the entire testing sample and then also stratified by European ancestry quantiles. We also compared the mean simulated phenotype value in the top 10% PRS quantile with the bottom 10% PRS quantile. Performance metrics were computed with the median reported over 50 folds.

3.3 Results

3.3.1 Comparison of PRS Performance Assuming Shared Genetic Architecture across Ancestral Populations

To evaluate the performance of slaPRS, we first conducted simulations with complete sharing of genetic architecture across ancestral populations (i.e., true effect sizes and risk variants are shared across European and African populations) for various disease architectures (see methods). Under this setup, differences in GWAS estimated effect sizes across ancestral populations are a function of solely LD. We constructed our stacked PRS using simulated European and African GWAS effect sizes for simulated admixed African Americans of varying ancestry proportions. The distribution of overall European ancestry in our simulated admixed African Americans was approximately normally distributed with a mean of around 50% (Figure 3.7).

We focus first on the full level 1 model with 5Mb windows using the local African and European PRS and local ancestry information in each block ($C_{B_i} = \{A_i, E_i, Anc_i, A_i \times Anc_i, E_i \times Anc_i\}$) with heritability $h^2 = 0.10$, number of causal variants $m = 100$, and equal size European and African GWAS sample size $n = 10{,}000$. Across simulations, our stacked PRS generally had an increased adjusted R^2 with the simulated phenotype compared to the existing approaches.

slaPRS had a 5.93% median adjusted R^2 for the true PRS across all admixed individuals in the testing set compared to C+T PRS_{EUR} (3.17%) and PRS_{AFR} (3.18) and $PRS_{Marquez}$ (3.39%) that globally combines PRS_{EUR} and PRS_{AFR}. Comparing individuals in the top vs bottom 10% of the PRS distribution, slaPRS had higher trait stratification ability with larger mean differences (0.84 vs 0.62, 0.64, 0.64 for PRS_{EUR}, PRS_{AFR}, and $PRS_{Marquez}$ respectively). We further stratified testing samples by quantiles of European ancestry and found our stacking approach using the full model explained more variance of the phenotype compared to both PRS_{EUR}, PRS_{AFR} and $PRS_{Marquez}$. Across all ancestry quantiles the percent increase in median adjusted R^2 for slaPRS compared to the other methods ranged from 38.46% to 120.61% (Figure 2). Most notably, slaPRS strongly reduced the ancestry dependence of PRS performance as compared to PRS_{EUR} and PRS_{AFR}. When quantified through a simple linear model, the adjusted R^2 for slaPRS increased by 0.0009 for every European ancestry quantile increase ranging from 5.69% (0-20% European ancestry) to 5.91% (80-100% European ancestry). On the other hand, single population PRS_{EUR} and PRS_{AFR} had larger changes in R^2 of 0.004 (2.60% to 4.22 %) and -0.001 (4.11%-3.60%) respectively for every quantile increase. $PRS_{Marquez}$ compared similarly to slaPRS with an R^2 increase of 0.0008 for every quantile increase, ranging from 3.46% to 3.91%.

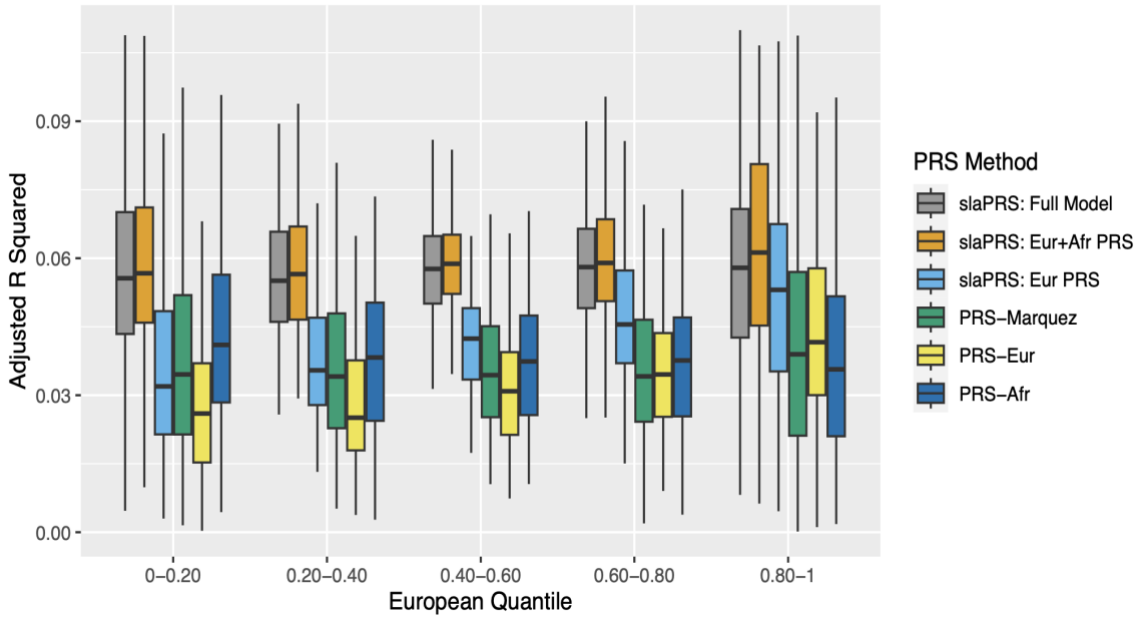

Figure 3.2 Boxplots comparing performance of slaPRS (differing in choice of level 0 predictors from each block), *PRS*-Marquez, and single population PRS: *PRS*-EUR & *PRS*-AFR (see methods) quantified through adjusted r-squared. Testing samples stratified by overall % of European ancestry.

While thus far we only considered the full slaPRS model ($C_{B_k} = \{E_k, A_k, Anc_k, A_k \times Anc_k, E_k \times Anc_k\}$, we then evaluated slaPRS under our alternative level 1 models that vary predictors from each local window. For the simplest case $C_{B_k} = \{E_k\}$ (i.e. only European GWAS considered and stacking local European PRS across blocks), slaPRS had adjusted R^2 ranging from 3.28% for 0-20% European ancestry to 5.45% for 80-100% European Ancestry and noticeably outperformed PRS_{EUR}. However, slaPRS under $C_{B_k} = \{E_k\}$ exhibited the strongest ancestry dependence (0.005 increase in adjusted R^2 across ancestry quantiles) across all methods. For $C_{B_k} = \{E_i, A_i\}$ (i.e. integrating European and African GWAS and stacking local European and African PRS across blocks), slaPRS further increased performance (compared to the single population case $C_{B_k} = \{E_k\}$) with adjusted R^2 ranging from 5.77% to 6.27% and had noticeably reduced ancestry dependence (0.001 increase in adjusted R^2 across ancestry

quantiles). The full level 1 model ($C_{B_k} = \{E_k, A_k, Anc_k, A_k \times Anc_k, E_k \times Anc_k\}$ further added local ancestry with interaction terms and performed comparably to the previous model ignoring ancestry $C_{B_i} = \{E_k, A_k\}$. Negligible differences in the full model and the model excluding local ancestry were present only in simulations of complete sharing of transethnic genetic effects.

Effect of Overall Heritability, Number of Causal Variants, Window Size, and African GWAS Sample Size

We quantified how slaPRS fared against other approaches across different simulation settings including: overall heritability $h^2 \in \{0.10, 0.30\}$, number of causal variants $m = \{5, 100, 500, 1000\}$, African GWAS sample size $n \in \{2000, 5000, 10000\}$, window sizes $\in \{1Mb, 5Mb\}$ (see Supplementary), and training data size $\in \{3000, 7000\}$ (see Supplementary) . Across all settings, slaPRS generally improved performance as compared to single ancestry PRS: PRS_{AFR} and PRS_{EUR} (Figure 3.8). Two factors had a sizable impact on the performance of slaPRS generally and its comparison to $PRS_{Marquez}$. The first major factor impacting PRS performance was the African GWAS sample size. As the African GWAS sample size decreased (while fixing $h^2 = 0.30$, $m = 100$) the C+T PRS_{AFR} performed increasingly worse compared to other methods (Figure 3.3). The performance of the full slaPRS model similarly decreased as the African GWAS sample size decreased, reflecting less informative contributions about the true risk variants from the African cohort. Furthermore, slaPRS exhibited a stronger ancestry dependence (converging towards the European only slaPRS model) as the African GWAS sample size decreased: For every increase in European ancestry quantile, slaPRS under the full model had an average change in average adjusted R^2 of 0.0009, 0.001 and 0.003 for African GWAS sample sizes of n=10000, n=5000, and n=2000 respectively. However, even for the

smallest African GWAS sample size scenario, slaPRS had the highest adjusted R^2 across ancestry quantiles.

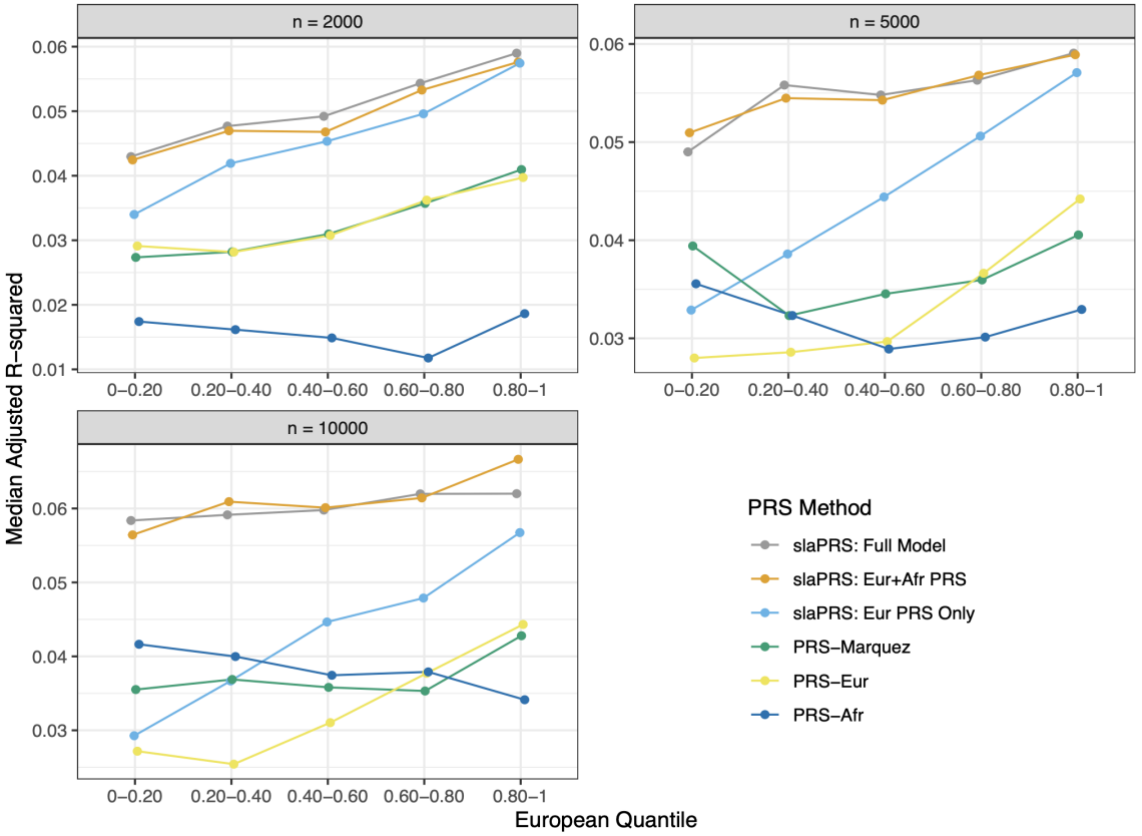

Figure 3.3 Line graph comparing PRS performance across methods (quantified by median adjusted r-squared between estimated PRS and phenotype value) as the African GWAS sample size changes (n=2000, 5000, 10,000). Testing admixed samples stratified by European ancestry quantile.

The second factor impacting slaPRS, especially compared to $PRS_{Marquez}$, was polygenicity and distribution of per variant effect sizes (Figures 3.8, 3.9). slaPRS generally had the greatest improvement in polygenic ($m = 100, 500$) simulations with moderate to large per variant effect sizes ($h^2 = 0.30, m = 100, 500$ and $h^2 = 0.10, m = 100$) driving clear genetic signals. Under these simulation parameters, the median adjusted R^2 of the full slaPRS model was 58.1% to 96.7% larger than the median adjusted R^2 of $PRS_{Marquez}$,. In such settings, a majority of window's local ancestry PRS contributing genetic signal to the stacking model. On the opposite

end, when polygenicity was lower ($m = 5$ causal variants, $h^2 = 0.10$) the median adjusted R^2 for slaPRS was more similar to $PRS_{Marquez}$ (23.4% increase), as a few large per variant effect sizes drive a small number of windows to dominate the genetic signal with remaining windows adding noise to the model. slaPRS similarly performed more similar to $PRS_{Marquez}$ (21.1% and 27.3% increase in adjusted R^2) in simulations of high polygenicity with low per variant effect sizes ($m = 500, 1000$ and $h^2 = 0.10$), as most windows are uninformative and those with very small genetic signal are likely overly penalized and shrunk.

3.3.2 Comparison of PRS Performance Assuming Differences in Genetic Architecture across Ancestral Populations

We also considered simulations in which the genetic architecture differed across ancestral populations (i.e., unique population-specific effect sizes), causing population-specific GWAS to vary from both differences in LD and true underlying effects across populations. We computed slaPRS using GWAS effect sizes varying the transethnic genetic correlation across risk variants $\rho = \{0.2, 0.5, 0.8\}$. We again focused on our base simulation parameters (heritability $h^2 = 0.10$, number of causal variants $m = 100$, and equal size European and African GWAS sample size $n = 10{,}000$). For the single population PRS_{EUR} and PRS_{AFR}, which do not consider a risk variant's local background, the adjusted R^2 from the PRS model was stable in their corresponding admixed groups (80-100% European and 0-20% European) across changing transethnic genetic correlation. However, when transethnic genetic correlation was low ($\rho = 0.2$), PRS_{EUR} and PRS_{AFR} notably had an increased decay in PRS performance as the admixed ancestry group diverged from the population GWAS (Figure 3.4): Comparing the shared transethnic genetic architecture case vs when ρ = 0.20, the change in adjusted R^2 was 0.005 vs

0.004 and -0.006 vs -0.001 across ancestry quantiles for PRS_{EUR} and PRS_{AFR} respectively. For slaPRS, notably the full level 1 stacking model ($C_{B_i} = \{E_i, A_i, Anc_i, A_i \times Anc_i, E_i \times Anc_i\}$ modeling local ancestry and interactions outperformed the model using only the local ancestry PRS ($C_{B_i} = \{E_i, A_i\}$ as the transethnic genetic correlation decreased. When genetic effects across ancestral populations were similar ($\rho = 0.8$), the percent increase in adjusted R^2 between the full model and model ignoring local ancestry ranged from 10.9% to 14.3% across ancestry quantiles, as compared to 23.4% to 50.5% when transethnic genetic effects are vastly different ($\rho = 0.2$) (Figure 3.4). Notably, the overall adjusted R-squared of the full level 1 model modeling ancestry specific effects dependent on a variant's ancestral background was stable across values of $\rho = \{0.2, 0.5, 0.8\}$: ($R^2 = 5.27\%, 5.18\%, \ 5.67\%$) as compared to the model ignoring local ancestry ($R^2 = 3.65\%, 4.09\%, 5.18\%$).

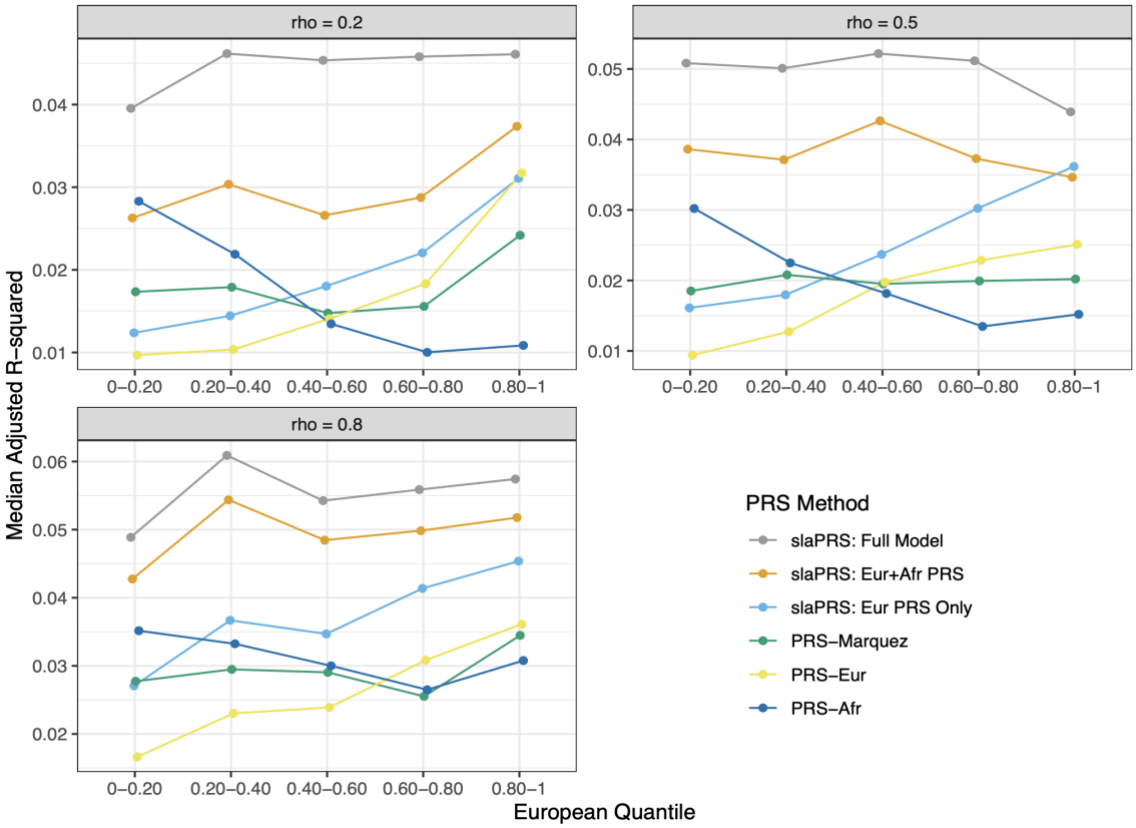

Figure 3.4 Line graph comparing PRS performance as quantified through median adjusted r-squared between the estimated PRS and phenotype value. Transethnic genetic correlation varies from $\rho = \{0.2, 0.5, 0.8\}$ and testing admixed samples stratified by European ancestry quantile.

3.3.3 Real Data Application

We conducted a real data application of our stacking method slaPRS using genotype and phenotype data from the UK Biobank. We considered three quantitative lipid traits: high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), and total cholesterol using estimated European and African American GWAS effect sizes from the Global Lipids Genetic Consortium (see methods for details). We first compared our approach to PRS_{EUR}, PRS_{AFR} (C+T using European and African GWAS effect sizes separately), and $PRS_{Marquez}$ (combining PRS_{EUR} and PRS_{AFR} globally) across all samples. For all three traits, slaPRS

improved the median adjusted R^2 values compared to PRS_{EUR} and PRS_{AFR} (Table 1). Similarly, slaPRS improved stratification ability as shown in larger mean phenotype values comparing individuals in the top and bottom 10% of the PRS distribution: HDL (0.373 vs 0.365, 0.324), LDL (1.019 vs 0.858, 0.905), TC (1.317 vs 1.028, 1.203). However, slaPRS performed similarly to $PRS_{Marquez}$ across all three traits with respect to both metrics, a pattern observed in simulation scenarios of lower polygenicity causing fewer windows to contribute to trait heritability (Table 3.1). Across the three traits, only 1.6% (HDL), 6.6% (LDL), and 2.1% (TC) of all level 0 local population PRS across the genome had an R^2 > 0.10 with the overall trait PRS. For LDL, which had the highest signal to noise ratio, we saw a minor improvement in both R^2 and top vs bottom 10% stratification ability for slaPRS. Furthermore, we found limited improvement in slaPRS using the full level 1 stacking model ($C_{B_i} = \{E_i, A_i, Anc_i, A_i x Anc_i, E_i x Anc_i\}$ compared to the reduced model ($C_{B_i} = \{E_i, A_i\}$)

a)

	Median Adjusted R^2				
Phenotype	slaPRS (Full Model	slaPRS (No Ancestry)	EUR C+T	AFR C+T	Global Stacked ($PRS_{Marquez}$)
HDL	0.081	0.084	0.069	0.054	0.083
LDL	0.112	0.112	0.088	0.097	0.110
TC	0.115	0.113	0.093	0.091	0.112

b)

	Mean Phenotype in Top vs Bottom 10% PRS Quantile				
Phenotype	slaPRS (Full Model	slaPRS (No Ancestry)	EUR C+T	AFR C+T	Global Stacked ($PRS_{Marquez}$)
HDL	0.377	0.390	0.369	0.321	0.401
LDL	1.025	1.024	0.853	0.903	1.008
TC	1.317	1.307	1.039	1.202	1.330

Table 3.1 Performance metrics for lipid phenotypes in UKB. a) Median adjusted r-squared from model PHENO ~ PRS + PC1 + PC2 + PC3 + PC4. b) Difference in mean phenotype for individuals in top 10% of PRS distribution vs bottom 10%.

We then stratified our testing samples by European ancestry quantile to 1) reassess overall PRS performance on admixed individuals in quantiles of 20%-80% European ancestry (removing primarily European or African admixed African British) and 2) quantify ancestry dependence of PRS performance across all five ancestry quantiles. In the bottom and top quantiles of predominantly homogenous African or European admixed African British, using single ancestry PRS_{EUR} and PRS_{AFR} tended to outperform. However, in the more heterogeneous admixed samples (20-80% European ancestry), slaPRS and $PRS_{Marquez}$ had the best median adjusted R^2 across all methods with comparable results for the three traits: HDL (0.066 and 0.070), LDL (0.103 and 0.098), TC (0.079 and 0.081) (Figure 3.5). Regarding ancestry dependence of PRS method, across traits PRS_{EUR} and PRS_{AFR} exhibited the strongest ancestry dependence, performing better as the proportion of European or African ancestry increased. On the other hand, methods using multiple ancestry GWAS had reduced ancestry dependence, with slaPRS having the smallest dependence followed by $PRS_{Marquez}$. For HDL, the average change in adjusted R^2 for each European quantile increase for slaPRS, PRS_{EUR}, PRS_{AFR}, and

$PRS_{Marquez}$ was 0.004, 0.019, -0.006, and 0.011 respectively. LDL (-0.003, 0.014, -0.016, and 0.003) and TC (-0.002, 0.012, -0.014, -0.005) had similar patterns across methods.

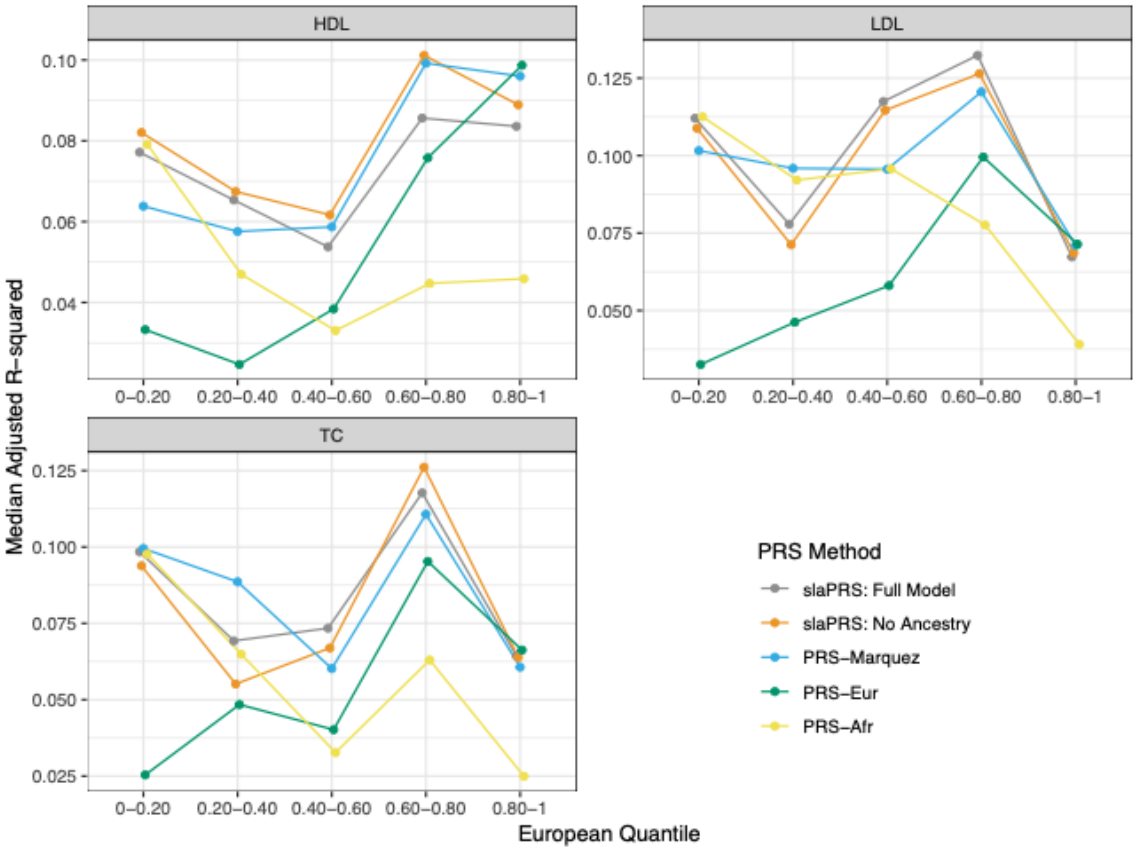

Figure 3.5 Line graph comparing PRS Performance for UKB lipid phenotypes. Performance quantified through median adjusted r-squared from model PHENO ~ PRS + PC1 + PC2 + PC3 + PC4. Testing admixed samples are stratified by European ancestry quantile.

3.4 Discussion

In this work we proposed a novel stacking framework to locally incorporate GWAS from multiple populations into construction of PRS for admixed individuals. Our method, slaPRS, segments admixed genomes into local regions of varying ancestry and optimizes a linear combination of local population specific PRS, local ancestry, and potential interactions to construct a PRS. In simulations, we first recapitulated previous findings that traditional PRS constructed using a single population GWAS in admixed samples are ancestry dependent. We then showed across a range of genetic architectures (varying heritability, number of causal variants, underrepresented GWAS sample size, and transethnic genetic correlation across ancestral populations) that slaPRS can outperform existing approaches (PRS_{EUR}, PRS_{AFR} and $PRS_{Marquez}$) and reduce the ancestry dependence compared to PRS_{EUR} and PRS_{AFR}. In admixed African British from the UK Biobank, we leveraged ancestry specific GWAS for lipid traits from the Global Lipids Genetic Consortium to compare slaPRS to existing PRS methods. We found in these lipid traits that incorporating multiple ancestry GWAS similarly improved performance and strongly reduced the ancestry dependence of PRS performance. However, for these data, there was inconclusive evidence combining information locally in slaPRS as opposed to globally ($PRS_{Marquez}$) was optimal across all traits.

From our simulations and UK Biobank applications, we conclude that slaPRS for PRS in admixed individuals is likely optimal (compared to existing approaches) for traits with high heritability and polygenicity. slaPRS extends $PRS_{Marquez}$ to combine information locally as opposed to globally and comparisons had interesting findings. In simulations, we found the smallest improvements were in trait architectures with low polygenicity (few windows meaningfully contribute to trait heritability with others add noise to the model) or in highly

polygenic settings where per-variant effect sizes are small (hard to distinguish signal from noise and genetic signals may be over shrunk). In real data applications, we found slaPRS and $PRS_{Marquez}$ performed similarly across the three lipid traits, likely driven by their trait genetic architecture. For the lipid traits studied, the former simulation scenario may be most prevalent as only 2-6% of all local PRS across windows contributed to the estimated PRS causing most regions to solely add noise to the model. As a result, noticeable improvements in slaPRS over $PRS_{Marquez}$ may be observed in more heritable and polygenic traits, such as height, in which more local windows across the genome will contribute genetic signal.

We surprisingly found explicitly modeling local ancestry in the slaPRS model (vs the model excluding local ancestry) provides the most improvement when there is at least moderate heterogeneity in true causal variant effect sizes across ancestral backgrounds. In simulations, the largest increase in PRS performance between slaPRS models occurs when transethnic genetic correlation is low ($\rho = 0.20$), with no improvements under scenarios of shared transethnic genetic architecture. In lipid traits from the UK Biobank, we observed similar findings regarding modeling local ancestry. In such traits, modeling local ancestry in the slaPRS model only provided marginal improvements, consistent with high estimated transethnic genetic correlations from Million Veteran Program participants for HDL (ρ =0.84) and moderate correlation for the other traits ($\rho \in [0.47, 0.69]$)[31]. High transethnic genetic correlations for the considered lipid traits are consistent with recent findings from Hou et al, that suggest a majority of common traits likely have similar causal effects across populations[99]. Such findings have immediate implications, as slaPRS and other approaches considering local ancestry background may find the most improvement in traits with significant differences in transethnic genetic architecture.

Historically in genetic studies, individuals are often discretized into ancestral populations and treated as homogenous within the group. Ding et al have recently challenged the historical paradigm by showing PRS accuracy varies between individuals even within a "homogenous" genetic ancestry cluster to ultimately push for treating genetic ancestry on a continuum[103]. Our method slaPRS is tailored to treat genetic ancestry on a continuum by taking a local approach to PRS prediction in admixed samples. As mentioned, $PRS_{Marquez}$ previously combined global population specific PRS successfully in admixed individuals, though in doing so uses a single weight for population specific effects. Potential heterogeneity in true population specific risk variants, estimated population specific GWAS effect sizes, and admixture proportions across loci and individuals would cause use of a single weight to be suboptimal. slaPRS extends $PRS_{Marquez}$ by combining population specific PRS at the local level instead to 1) allow for varying effects of local population specific PRS across the genome and 2) increase overall external GWAS sample sizes to improve effect size estimation and identify the true causal variants. The first benefit is accomplished through our level 1 elastic net stacking model that learns a linear combination of local population specific PRS (and local ancestry with interaction effects) to inform which population's local PRS should be upweighted or shrunk. In the case that the true causal effect differs due to ancestral background, slaPRS handles this scenario by modeling the local ancestry and interactions with the local population specific PRS, allowing for the effect of a local population specific PRS to differ based on its ancestral background. The second benefit is accomplished by increasing the overall effective GWAS sample size through incorporating information from each population's GWAS. In the case that the genetic architecture is shared across ancestral backgrounds, using information from both GWAS will boost power and improve effect size estimation of the shared risk variants and their locations.

However, when the genetic architecture differs across populations it is unclear whether using multiple population GWAS can be viewed in a similar manner.

slaPRS has desirable statistical and computational properties as well. First, similar to other machine learning-based PRS methods such as TL-PRS[120] in the context of cross population prediction incorporating multiple ancestry GWAS, slaPRS avoids the needs for any distributional assumptions on transethnic effect sizes as compared to the cross population PRS methods PRS-CSx[102] and PolyPred[121] (Utilizes BOLT-LMM[122] and PRS-CS[123] which treat SNP effects as random). As a result, our approach makes no assumption on whether a risk variant is shared across population, where each local population PRS in a genomic region can include its own set of risk variants. Second, slaPRS does not require an external LD reference panel or genotypes outside of the admixed genotypes. Third, slaPRS can accommodate any PRS algorithm to construct local population PRS (here we use the C+T algorithm for simplicity). For example, REGENIE[112] uses a ridge regression based approach to construct level 0 local PRS before stacking. Lastly, our approach is computationally very efficient, as discretizing the genome into local windows facilitates efficient parallel processing of level 0 predictions, with a final level 1 elastic net model that can be fit very fast with standard statistical packages.

While slaPRS provides a novel stacking approach to combine population specific GWAS information locally, it has a few limitations to consider. We assume existence of GWAS from each ancestry contributing to a genetic admixture, though high powered GWAS in understudied homogenous populations such as Africans are currently limited or non-existent. As a result, our real data application was limited to using African American GWAS as proxies for African GWAS, with only a handful of lipid traits from the Global Lipids Genetic Consortium having sufficiently large GWAS sample sizes. Recent efforts for genomic research in diverse

populations such as the African biobank[124] should help to resolve this issue. Furthermore, we describe our framework for continuous value phenotypes, owing to currently limited access to large sample GWAS for binary case/control traits in each ancestral population. Extending this framework to case/control traits using a logistic regression elastic net and liability threshold model should be straightforward. Lastly, while we push to treat admixed individuals on a genetic ancestry continuum, our approach assumes the super population groups such as "European" and "African" have homogenous genetic architecture with respect to a complex trait across their subpopulations. However, studies have shown a high degree of genetic diversity across the African continent[125,126] with unique demographic histories driving substantial cultural and ethnic differences that may cause treating all African subpopulations as homogenous to be problematic[103,127].

Despite the limitations, slaPRS provides an efficient data driven framework to constructing polygenic risk scores in admixed samples that leverage multiple population GWAS. In providing a method that not only performs well in admixed samples, but equally well across varying ancestry proportions we strive to improve on the current inequity in genetics research that is fast resolving in our community. Furthermore, as sample sizes increase in underrepresented populations for more traits, we expect slaPRS to have additional applications. Lastly, while our work thus far only considered two-way admixture, our approach can easily accommodate three or more ancestral populations and respective external GWAS. In coming years admixture will likely extend beyond the historically predominant African American and Latino admixed groups as people and cultures from various ancestral backgrounds are brought together geographically. As a result, we believe our method's flexibility to accommodate

increasingly complex admixture types using information from multiple GWAS will become even more relevant.

3.5 Chapter 3 Appendix

3.5.1 Derivation of Weighted Function Learned from slaPRS

We restate our model setup consisting of a sample of N admixed individuals with ancestral contributions from population A and B. Let **X** be the NxM admixed genotype matrix (M is the total number of variants genome wide) and **Y** the $Nx1$ phenotype vector. Let L_{ij} be an NxM matrix denoting the haplotype-level local ancestry (l_{ij1}, l_{ij2}.) of individual i at marker j. We assume the phenotype can be expressed as:

$$Y_i = \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} X_{ij} f(\beta_{A_j}, \beta_{B_j}, L_{ij}) + \boldsymbol{\epsilon}_i$$

Where X_{ij} is the genotype dosage for individual i at marker j, and β_{A_j}, β_{B_j} are effects for marker j on the phenotype in populations A and B respectively. Here, $f(\beta_{A_j}, \beta_{B_j}, L_{ij})$ is a weighted average of population specific GWAS effect sizes and local ancestry learned via our stacking approach.

Following construction of level 0 model predictions in each window C_{B_k}across the genome (includes local population A PRS A_k and local population B PRS B_k, local ancestry, and interaction terms) we fit the following stacking model:

$$Y = w_0 + \boldsymbol{w_1} C_{B_1} + \boldsymbol{w_2} C_{B_2} + \cdots + \boldsymbol{w_k} C_{B_k}$$

Expanding out terms for the k-th window:

$$= w_0 + \left[w_{k,A_k} A_k + w_{k,B_k} B_k + w_{k,anc} Anc + w_{k,anc:A_k} Anc\ x\ A_k + w_{k,anc:B_k} Anc\ x\ B_k\right] + \cdots$$

$$= w_0 + \left[w_{k,anc} Anc + A_k\left(w_{k,A_k} + w_{k,anc:A_k}\right) + B_k\left(w_{k,B_k} + w_{k,anc:B_k} Anc\right)\right] + \cdots$$

The stacking procedure learns a linear combination of level 0 model prediction in each window C_{B_k} across the genome through estimating the weights w_k. A_k and B_k are themselves weighted sum of risk variants using population specific GWAS reducing the form to:

$$= w_0 + \left[w_{k,anc} Anc + \left(\sum_{j=1}^{m_k} X_{ij} \beta_{A_j} \right) \left(w_{k,A_k} + w_{k,anc:A_k} \right) + \left(\sum_{j=1}^{m_k} X_{ij} \beta_{B_j} \right) \left(w_{k,B_k} + w_{k,anc:B_k} \right) \right]$$

$$+ \cdots$$

$$= w_0 + \left[w_{k,anc} Anc + \sum_{j=1}^{m_k} X_{ij} \left[\beta_{A_j} \left(w_{k,A_k} + w_{k,anc:A_k} \right) + \beta_{B_j} \left(w_{k,B_k} + w_{k,anc:B_k} \right) \right] \right] + \cdots$$

Where w_{k,A_k} and $w_{k,anc:A_k}$ are weights for population A specific local PRS A_k and its local ancestry interaction term. Because A_k (and likewise for B_k) is a function of population A GWAS effect sizes that is shared across all variants in the window k, we replace the notation w_{k,A_k} with $w_{k,\beta_{A_j}}$ and similarly anc_k is a function of L_{ij} so we replace $w_{k,anc:A_k}$ with $w_{k,L_{ij}}^{(A)}$.

$$f\left(\beta_{A_j}, \beta_{B_j}, L_{ij}\right) = \beta_{A_j} \left(w_{k,\beta_{A_j}} + w_{k,L_{ij}}^{(A)} L_{ij} \right) + \beta_{B_j} \left(w_{k,\beta_{B_j}} + w_{k,L_{ij}}^{(B)} L_{ij} \right)$$

3.5.2 Effect of window size and training dataset size

slaPRS takes a sliding local window approach to construct local population-specific polygenic risk scores and thus may be sensitive to the size of the window. In simulations under our base scenario ($h^2 = 0.10, m = 100$) we considered both 1Mb and 5Mb windows. PRS performance quantified by adjusted R^2 with the phenotype were highly consistent across window sizes suggesting slaPRS is robust to window size (Figure 3.10). We further quantified the effect of varying the training dataset size of admixed individuals (n = 3000, n = 7000). As compared to

PRS_{EUR} and PRS_{AFR}, slaPRS uses the training data to weight local population specific PRS (and the variants effects themselves) and increased performance should be dependent on the training dataset size. In general, slaPRS for training sizes n=3000 and n=7000 generally had increased adjusted R^2 when the training size was larger compared to PRS_{EUR} (77.3%, 84.9%), $PRS_{AFR}(35.3\%, 66.1\%)$ and $PRS_{Marquez}$ $(64.4\%, 66.4\%)$.

3.5.3 Code Availability

An R package for slaPRS has been developed with code and example workflow available at: https://github.com/kliao12/slaPRS

3.5.4 Supplementary Tables and Figures

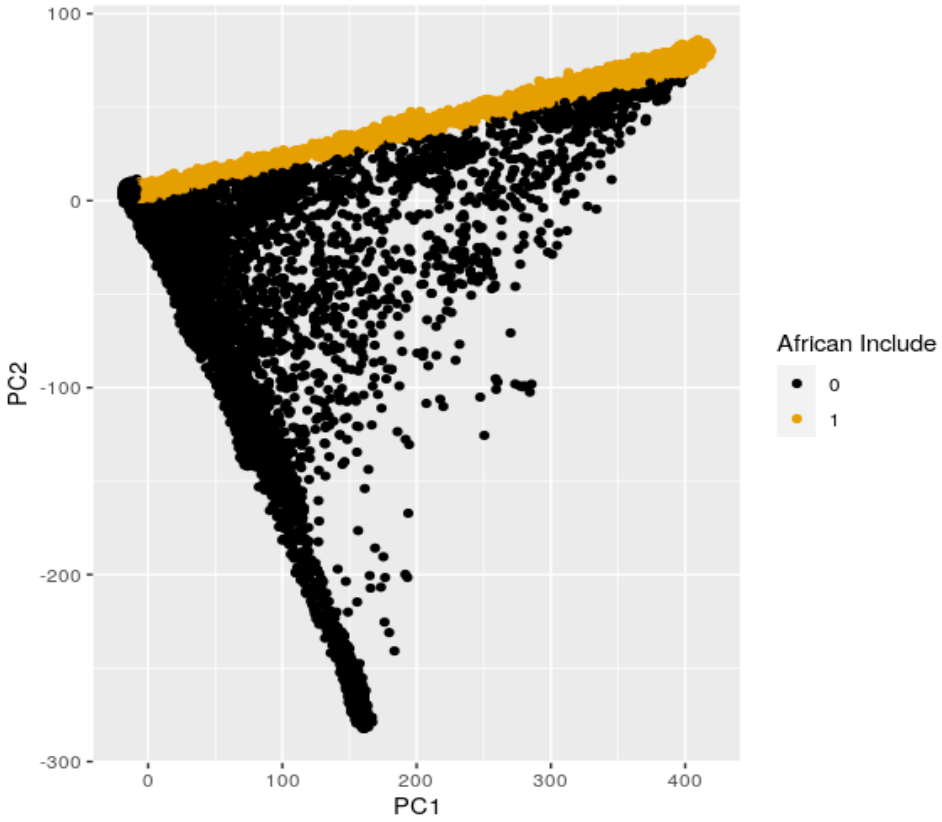

Figure 3.6 Scatterplot of n=20,262 UKB samples containing African ancestry along diagonal of PC1.

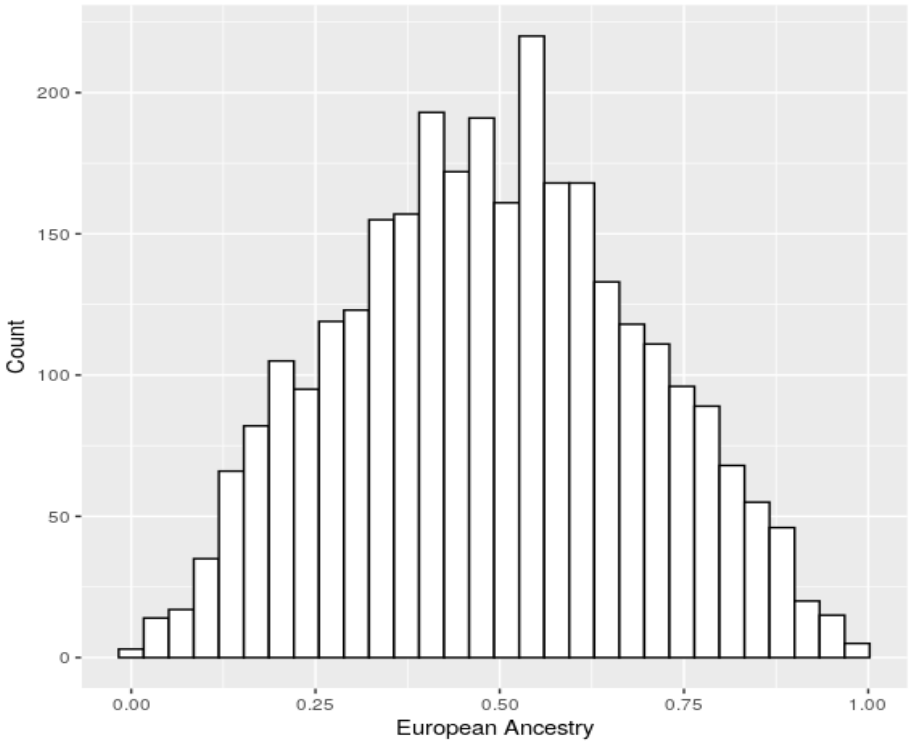

Figure 3.7 Histogram of the distribution of overall European ancestry across n=10,000 simulated admixed African Americans (for a single simulation).

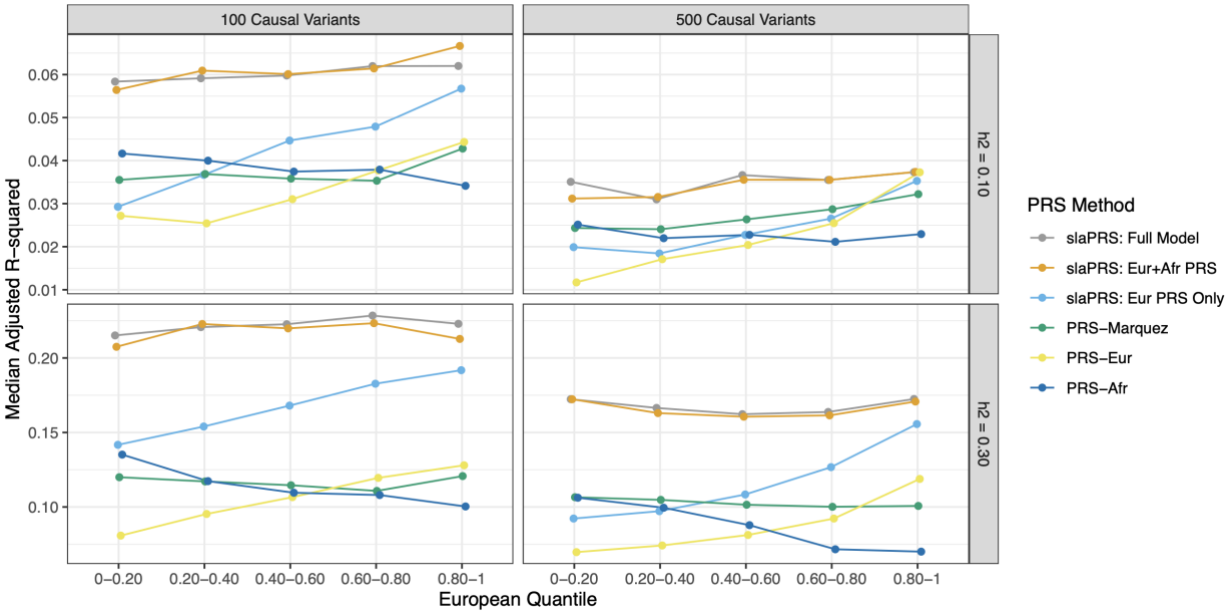

Figure 3.8 Line graph comparing PRS performance across PRS methods for different simulation settings using adjusted r-squared between estimated PRS and simulated phenotype. Simulation parameters: heritability (0.1,0.3) and number of causal variants (100,500).

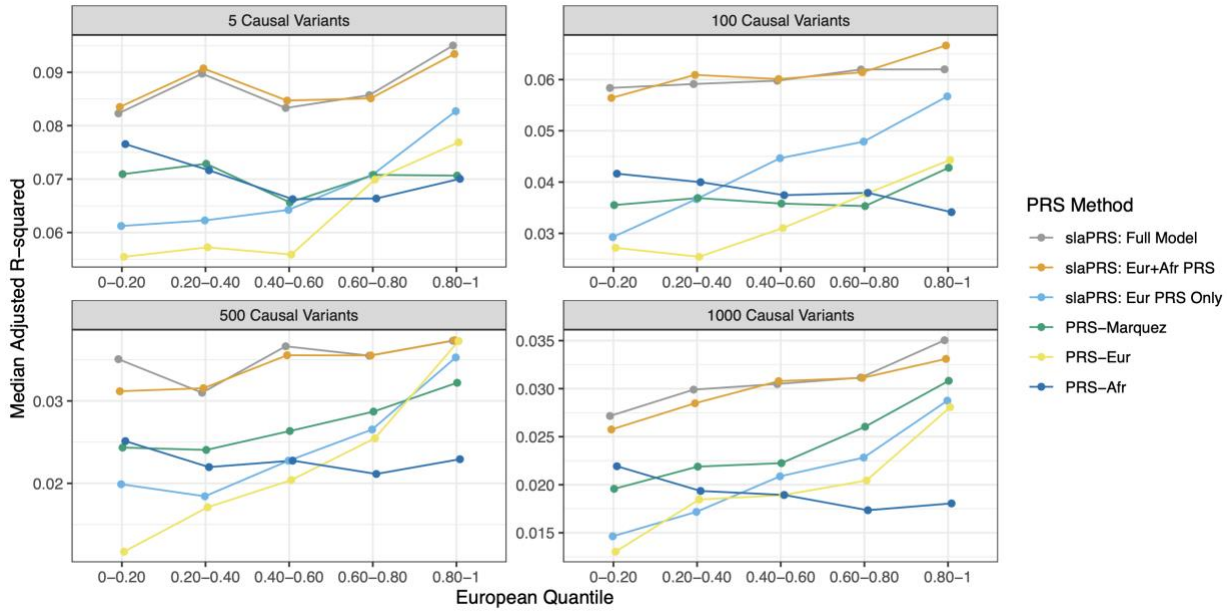

Figure 3.9 Line graph comparing PRS performance across PRS methods for different simulation settings using adjusted r-squared between estimated PRS and simulated phenotype. Simulation parameters: heritability (0.1) and number of causal variants (5,100,500,1000).

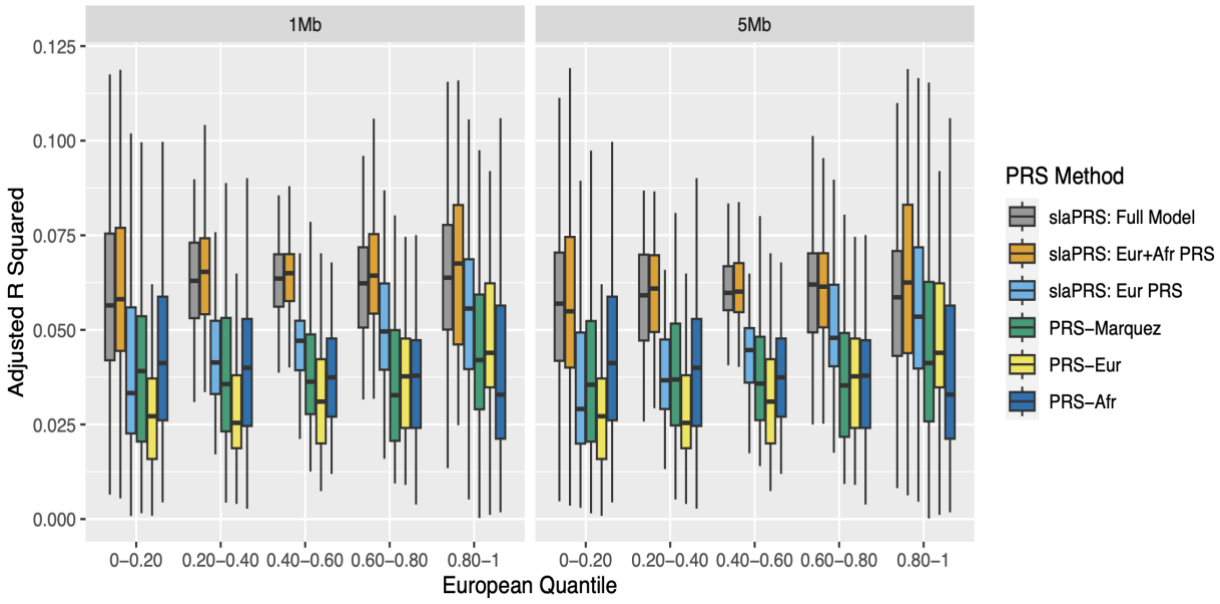

Figure 3.10 Comparison of PRS performance across methods (quantified by adjusted r-squared between estimated PRS and phenotype value) as the window size in slaPRS varies (1Mb, 5Mb).

Chapter 4 Mixture of Cross Trait LD Score Regressions Identifies Variant Sets and Genes Driving Signals of Local Genetic Correlation

4.1 Introduction

Genetic association testing over the past decade has successfully implicated numerous risk variants and genes across thousands of traits[93,128]. In doing so, studies have gleaned insight into the genetic architecture of traits, defined as the characterization of the genetic variation responsible for broad-sense phenotypic heritability[129]. As the number of GWAS studies has increased, widespread pleiotropy across traits has further been revealed in which a variant or gene is associated with more than one phenotype[28,130]. A well-known example is phenylketonuria (PKU), in which a single gene that codes for the enzyme phenylalanine hydroxylase effects phenotypes including eczema, intellectual disability, and skin pigmentation[131].

Using GWAS summary statistics across traits, pleiotropy can be studied at three broad levels: SNP-level, region level, or genome wide. Two popular approaches for identifying pleiotropy at the SNP-level are colocalization[40] and pheWAS[132,133]. pheWAS identifies SNP-level pleiotropy through performing hypothesis-free GWAS across multiple traits (typically in biobanks collecting phenotype data on hundreds of traits) and discovering shared GWAS hits. On the other hand, colocalization utilizes GWAS summary statistics across traits in a Bayesian probabilistic approach to consider scenarios of whether a region shares a causal variant and the most likely specific variant among many SNPs[40,134,135]. Extensions of colocalization have further allowed for the presence of multiple shared causal variants in a region and consideration of more than two traits[134,136,137].

Genetic correlation is another technique to characterize pleiotropy, done through quantifying the correlation of genetic effects across two phenotypes[138,139]. In quantifying a correlation, genetic correlation identifies a stronger version of pleiotropy as it requires genetic signals to not only have similar magnitude of effects, but now also the same direction of effects. Methods for computing genetic correlation can be categorized into those using individual level data[36,140] (estimated from variance component of linear mixed models) or summary statistics[141,142], with summary statistic-based approaches being more popular due to computation and limitations in individual level data sharing. Genetic correlation was initially considered genome wide and thus considered pleiotropy at the genome wide scale. However, methods for estimating local genetic correlation have recently been proposed[37–39] and become popular to study pleiotropy at the region-level, as there can exist heterogeneity in magnitude and direction of genetic correlations across regions of the genome[38].

Local genetic correlation (LGC) and colocalization have become two popular methods in the field to study cross-trait genetic architectures and conceptually answer a similar question in "how does a local region share genetics across two traits (or more)?". However, as contrasted by Werme et al[37], the underlying model and specific aims differ drastically. Colocalization aims to identify a shared causal variant (or multiple) across two traits in a region, while LGC aims to quantify the overall strength of the local genetic sharing in a region. Thus, colocalization is the preferred approach when SNP-level pleiotropy is of interest, as it allows for identifying specific shared variants. However, colocalization typically relies on having GWAS hits and thus a regional colocalization analysis could fail in the presence of shared genetics if studies are underpowered or a region has weaker effect shared causal variants. In such settings LGC may still be able to quantify the extent of the sharing of weaker genetic effects and reveal insights into

the pleiotropic effects for a genomic region. Thus, LGC and colocalization provide complementary approaches to interrogating local patterns of shared genetic signal.

While LGC provides insights into the magnitude and presence of shared genetic signals in a region, existing frameworks are not capable of identifying pleiotropy at the variant or gene level. One major reason is LGC analysis in local regions is often complicated by pervasive LD causing marginally estimated GWAS z scores to be highly correlated. To account for this extensive LD and potentially noisy LD estimates from external reference panels, methods such as Rho-hess[38], SUPERGNOVA[39], and LAVA[37] all project GWAS effect sizes or z scores onto eigenvectors of the regional LD matrix. Furthermore, only the top eigenvalues are retained in the analysis, as eigenvalues explaining a low proportion of variability are pruned away. While projecting z scores onto the LD matrix is an effective approach to handle the extensive LD in local regions, in doing so these frameworks lose the ability to efficiently draw conclusions at the original SNP-level as observations are now represented as linear combinations of the original SNPs in a reduced dimensionality. Thus, LGC frameworks can identify regions sharing genetic signals but are unable to identify which specific SNP sets or genes drive that signal of LGC. While approaches such as colocalization can be applied in LGC regions to identify shared causal SNPs, as mentioned such methods implicitly rely on GWAS hits and would likely fail if the LGC signal is driven by weaker shared causal variants.

In this work we propose LDSC-MIX, a novel mixture of regressions method to identify SNP and gene-level pleiotropy in a region using the framework of genetic correlations (while not losing SNP-level interpretability through projecting z scores onto LD matrices as done in existing LGC methods). LDSC-MIX extends the cross trait bivariate LD score regression (LDSC) framework used to estimate genome level genetic correlation, and thus takes as input

GWAS summary statistics across two traits and LD scores using a matched-ancestry reference panel. Our method LDSC-MIX similarly regresses the product of GWAS z scores across two traits on variant LD scores, but uniquely assumes two latent groups of SNPs in a local region: 1) SNPs with correlated genetic effects 2) SNPs with uncorrelated genetic effects. Using an empirical Bayes mixture of regressions model, LDSC-MIX infers group membership for each SNP to identify SNP-level pleiotropy (where inferred group labels can then be used to test for enrichment in SNPs at the gene level). In simulations of varying shared genetic architecture, we compared LDSC-MIX to colocalization in their ability to identify a candidate set of shared causal variants containing true shared risk variants and recover the true LGC. With respect to these criteria, we found LDSC-MIX is generally preferable in the presence of multiple shared weak causal genetic variants. In such settings, colocalization with a multiple variant extension (coloc-SuSiE) often failed (or performed poorly) while LDSC-MIX was still able to leverage the linear relationship between LD scores and products of z scores. On the other hand, in the presence of a single shared causal variant or multiple strong genetic signals, colocalization was the preferred approach. Using GWAS summary statistics for two trait pairs from the UK Biobank: asthma-basal cell carcinoma and asthma-HDL, we applied LDSC-MIX on LGC regions and identified candidate sets of shared causal variants that recaptured the estimated LGC and highlighted specific genes in the region that may drive the LGC signal.

4.2 Methods

4.2.1 Original Cross-trait LD Score Regression Framework

We follow the framework established in cross trait LD score regression defined by Bulik-Sullivan et al[141]. For a given pair of traits assume there are two studies of size N_1 and N_2 with standardized phenotype vectors Y_1, Y_2 of size $N_1\ x\ 1$ and $N_2\ x\ 1$. Each study has the standardized

(mean zero and variance 1) genotype matrix X_1, X_2 of size $N_1 x\ M$, $N_2\ x\ M$ where M is the number of shared SNPs genome wide. Consider the following linear models

$$Y_1 = \beta_1 X_1 + \epsilon$$

$$Y_2 = \beta_2 X_2 + \delta$$

Where β_1 and β_2 are vectors of genetic effects and δ and ϵ are residual vectors capturing environmental effects. $\beta_1, \beta_2, X_1, X_2, \epsilon, \delta$ are treated as random. Suppose the genetic effects (β_1, β_2) has mean 0 and covariance matrix

$$var(\beta_1, \beta_2) = \frac{1}{M}\begin{pmatrix} h_1^2 I & \rho_g I \\ \rho_g I & h_2^2 I \end{pmatrix}$$

And similarly, the environmental effects (ϵ, δ) have mean 0 and covariance matrix

$$var(\epsilon, \delta) = \frac{1}{M}\begin{pmatrix} (1 - h_1^2) I & \rho_e I \\ \rho_e I & (1 - h_2^2) I \end{pmatrix}$$

Where ρ_g, ρ_e are genetic and environmental correlations. Let z_{1i}, z_{2i} be the corresponding z-scores for snp i in each study. From this model, Bulik-Sullivan et al propose the cross-trait LD Score regression equation:

$$E[z_{1i} z_{2i}] = \frac{\sqrt{N_1 N_2}\rho_g}{M} L_i + \frac{\rho N_s}{\sqrt{N_1 N_2}}$$

Where z_{1i}, z_{2i} are per-variant z-scores, $L_i = \sum r_{ij}^2$ is the LD score[143] of variant i measuring the tagging of genetic variation in a predefined window, N_s is the number of overlapping samples between the two studies and ρ is the phenotypic correlation among the N_s samples. Estimates of the slope from this regression can be used to estimate the genetic correlation ρ_g.

Unweighted regression estimates from this model are suboptimal due to violations in regression assumptions including 1) correlated outcomes: SNP z-scores in a region can be highly correlated

due to LD and 2) heteroskedasticity (unequal variance): SNPs with high LD scores typically have higher variance in z-scores. Thus, the LD score regression adopts a weighted least squares approach with the following weights:

$$W_{1,i} = \frac{1}{LD_i}$$

$$W_{2,i} = var(z_{1i}z_{2i}|LD_i)^{-1} = \left[\left(\frac{N_1 LD_i}{M} + 1\right)\left(\frac{N_2 LD_i}{M} + 1\right) + \left(\frac{\sqrt{N_1 N_2}\rho_g}{M} LD_i + \frac{\rho N_s}{\sqrt{N_1 N_2}}\right)^2\right]^{-1}$$

W_{1i} handles correlation among SNPs by downweighing SNPs with high LD to avoid over counting. W_{2i} handles heteroskedasticity by weighting by the inverse of the conditional variance where parameters are defined as above.

4.2.2 Mixture of Cross-Trait LD Score Regression Framework

LDSC-MIX extends the original cross-trait LD score regression framework to be applied in local regions to identify SNPs and genes driving local genetic correlation. We note the original cross-trait LDSC was originally proposed to estimate genetic correlation at the genome level. Previous work[38,39] has suggested their method of moments estimation to be suboptimal in estimating genetic correlation locally due to an insufficient number of SNPs and pervasive LD. While estimates of LGC from cross-trait LDSC are less precise compared to explicit LGC methods[38], they were shown to remain accurate and unbiased. Thus, while suboptimal, we adopted the cross-trait LDSC framework for LDSC-MIX to allow for a latent mixture model-based approach while retaining information at the SNP-level, rather than projecting SNP z scores onto the local LD matrix.

Mixture of Regressions Model: LDSC-MIX

For LDSC-MIX, we subset the genotype matrices X_1, X_2 to a local region with Q variants with z-scores z_{1i}, z_{2i} $(i = 1, \ldots., Q)$. For each variant i, we compute LD scores $L_i = \sum_k r_{jk}^2$ using all SNPs in the local region.

Let C be a latent class indicator variable with $P(C_i = j) = \pi_i$ for $j = 1,2$. Here, C corresponds to two possible groups: 1) a set of genetically correlated SNPs contributing to shared genetic architecture across two traits in a region and 2) a set of uncorrelated SNPs in a region. Given the latent group C we assume a linear relationship between product of z-scores and the LD score as per cross trait LD score regression:

$$(z_{1i}z_{2i}|LD_i, C_i = j) = \gamma_{0j} + LD_i\gamma_{1j} + \epsilon_j$$

$$\epsilon_j \sim N(0, \sigma^2)$$

Following linear regression theory and assumed normally distributed residuals:

$$(z_{1i}z_{2i}|LD_i, C_i = j) \sim N(\gamma_{0j} + LD_i * \gamma_{1j}, \sigma_j^2)$$

Thus, the joint probability distribution for a given SNP i's pair of z scores and latent class variable C can be expressed as:

$$p(z_{1i}z_{2i}, C_i|LD_iC_i, \boldsymbol{\theta})$$

$$= \left\{\frac{\pi}{\sqrt{2\pi}\sigma_1}\exp\left(\frac{[z_{1i}z_{2i} - (\gamma_{01} + LD_i\gamma_{11})]^2}{2\sigma_1^2}\right)\right\}^{C_i}$$

$$* \left\{\frac{1-\pi}{\sqrt{2\pi}\sigma_2}\exp\left(\frac{[z_{1i}z_{2i} - (\gamma_{02} + LD_i\gamma_{12})]^2}{2\sigma_2^2}\right)\right\}^{1-C_i}$$

Where the vector of model parameters $\boldsymbol{\theta} = \{\pi_1, \pi_2, \sigma_1^2, \sigma_2^2, \gamma_{01}, \gamma_{11}, \gamma_{02}, \gamma_{12}\}$. π_j is the group membership probabilities that a variant belongs to the j-th latent group ($\pi_1 + \pi_2 = 1$), γ_{0j} and

γ_{1j} are linear model intercept and slopes for the j-th latent group, and σ_j^2 are group specific variances.

Similar to the original LD score regression and its cross-trait extension framework, our mixture of regressions approach is made inefficient due to the violations in regression assumptions 1) correlation among SNPs and 2) heteroskedasticity. Thus, we adopt a weighted least squares framework for our mixture of regressions model. To accommodate heteroskedasticity and correlation among SNPs, we similarly introduce weighting by the inverse of the conditional variance and inverse of the LD score:

$$W_{1i} = var(z_{1i}z_{2i}|LD_i)^{-1} = \left[\left(\frac{N_1 LD_i}{M} + 1\right)\left(\frac{N_2 LD_i}{M} + 1\right) + \left(\frac{\sqrt{N_1 N_2}\rho_g}{M} LD_i + \frac{\rho N_s}{\sqrt{N_1 N_2}}\right)^2\right]^{-1}$$

$$W_{2i} = \frac{1}{LD_i}$$

$$W_i = W_{1i} * W_{2i}$$

The weight in the weighted least squares probability density given group C is incorporated in the variance and can be expressed as:

$$(z_{1i}z_{2i}|LD_i, W_i,\ C_i = j) \sim N(\gamma_{0j} + LD_i * \gamma_{1j},\ \sigma_j^2 * W_i)$$

The total weighted least squares likelihood across the j latent groups is then defined as

$$L(\boldsymbol{\theta}|z_1 z_2,\ LD,\ C, W)$$

$$= \prod_{i=1}^{Q}\left\{\frac{\pi}{\sqrt{2\pi}\sigma}\exp\left(\frac{[z_{1i}z_{2i} - (\gamma_{01} + LD_i\gamma_{11})]^2}{2W_i\sigma_1^2}\right)\right\}^{C_i}$$

$$* \left\{\frac{1-\pi}{\sqrt{2\pi}\sigma}\exp\left(\frac{[z_{1i}z_{2i} - (\gamma_{02} + LD_i\gamma_{12})]^2}{2W_i\sigma_2^2}\right)\right\}^{1-C_i}$$

Priors:

LDSC-MIX takes a Bayesian approach and thus incorporates the following prior information. We start by assuming a Dirichlet prior for the group membership probability

$$\pi \sim Dirichlet(\alpha)$$

Where $\alpha = \{1,9\}$ is a vector of length two (corresponding to the two groups) that down weights the correlated group (group 1), reflecting belief that the minority of variants share genetic signal. For the group j means (parameterized by the regression intercepts and slopes) we assume conjugate normal priors and take an empirical bayes approach to define the mean and variance:

$$Group\ j\ Regression\ Intercept: \gamma_{01}, \gamma_{02} \sim N(\widehat{\gamma_{0,overall}}, \sqrt{N_1 N_2} SE(\widehat{\gamma_{0,overall}}))$$

$$Group\ 1\ (Correlated)\ Regression\ Slope: \gamma_{11} \sim N(\widehat{\gamma_{1,Final}}, \sqrt{N_1 N_2} SE(\widehat{\gamma_{1,overall}}))$$

$$Group\ 2\ (Non\ Correlated)\ Regression\ Slope: \gamma_{12} \sim N(0, 0.001)$$

For group 1 (correlated group), the prior mean for the slope γ_{11} is defined running the original cross trait LDSC twice using 1) all variants in the region and 2) clumped "hits" in the region. In the first run using all variants in the region, we estimate the slope and intercept $\widehat{\gamma_{1,overall}}$ and $SE(\widehat{\gamma_{1,overall}})$. However, the estimated slope $\widehat{\gamma_{1,overall}}$ is likely an underestimate for the correlated group of causal variants we are interested in identifying. We again run cross trait LDSC using clumped variants (clumped at $r^2 = 0.10$) to estimate $\widehat{\gamma_{1,clumped}}$, which is likely an overestimate of the slope as we use top hits in each proximal group of SNPs. The final prior mean for the correlated group is defined as the average:

$$\widehat{\gamma_{1,Final}} = \frac{\widehat{\gamma_{1,overall}} + \widehat{\gamma_{1,clumped}}}{2}$$

To define the prior variance for group 1 (correlated group), we use the corresponding standard error of the original slope estimate scaled by the sample sizes N_1 and N_2: $\sqrt{N_1 N_2}\, SE(\widehat{\gamma_{1,overall}})$,.

For group 2 (uncorrelated group), the prior mean is set to 0 (as would be expected if there is no shared genetic signal) and the prior variance set to 0.0001 to enforce strong prior knowledge of "no correlation" on the model. For each group's intercept γ_{01}, γ_{02}, we follow a similar approach using point estimates of the intercept and standard error running the original cross trait LD score regression using all variants in the region.

For group *j* variances, we use the non-informative Jeffrey's prior for a normal distribution with known mean:

$$\sigma_1^2 = \frac{1}{\sigma_1^2}$$

$$\sigma_2^2 = \frac{1}{\sigma_2^2}$$

Lastly, Bayesian mixture models commonly suffer from label-switching[144], causing model parameters to become unidentifiable. We alleviate this issue by ordering group specific slopes $\gamma_{11} > \gamma_{12}$ (if the original LDSC estimated slope using all variants $\widehat{\gamma_{1,overall}} > 0$) and vice versa if $\widehat{\gamma_{1,overall}} < 0$, forcing group 1 to be the SNP set of genetically correlated effect sizes. The full joint prior of $\theta = \{\pi, \sigma_1^2, \sigma_2^2, \gamma_{01}, \gamma_{11}, \gamma_{02}, \gamma_{12}\}$. is then:

$$p(\theta) = p(\pi) * p(\sigma_1^2) * p(\sigma_2^2) * p(\gamma_{01}) * p(\gamma_{02}) * p(\gamma_{11}) * p(\gamma_{12})$$

4.2.3 Model Estimation

Parameters to be estimated from the model include $\boldsymbol{\theta} = \{\pi_1, \pi_2, \sigma_1^2, \sigma_2^2, \gamma_{01}, \gamma_{11}, \gamma_{02}, \gamma_{12}\}$. To estimate model parameters, we use Gibbs Sampling[145] to perform MCMC simulations to estimate posterior distributions. The full conditionals (see supplementary for derivations) are defined as:

$$\pi|\boldsymbol{\theta}_{-\pi}, z_1 z_2, LD \sim Dirichlet(\alpha + Q_k)$$

Where $Q_k = \sum_{i=1}^{Q} I(C_i = k)$

$$C_i|\boldsymbol{\theta}_{-C}, z_1 z_2, LD \sim Categorical(w_i)$$

Where $w_i = \frac{\pi N(z_1 z_2|\ldots)}{\sum \pi_j N(z_1 z_2|\ldots)}$

$$\gamma_{01}|\boldsymbol{\theta}_{-\gamma_{01}}, z_1 z_2, LD \sim N\left(\frac{1}{\frac{1}{s^2}+\frac{Q_1}{\sigma_1^2}}\left(\frac{m}{s^2}+\frac{\sum_{I:C_i=1} z_{1i}z_{2i}}{\sigma_1^2}\right), \left(\frac{1}{s^2}+\frac{Q_1}{\sigma_1^2}\right)^{-1}\right)$$

Where m=$\widehat{\gamma_{0,overall}}$, $s = \sqrt{N_1 N_2}SE(\widehat{\gamma_{0,overall}})$

$$\gamma_{02}|\boldsymbol{\theta}_{-\gamma_{02}}, z_1 z_2, LD \sim N\left(\frac{1}{\frac{1}{s^2}+\frac{Q_1}{\sigma_2^2}}\left(\frac{m}{s^2}+\frac{\sum_{I:C_i=2} z_{1i}z_{2i}}{\sigma_2^2}\right), \left(\frac{1}{s^2}+\frac{Q_1}{\sigma_2^2}\right)^{-1}\right)$$

Where m=$\widehat{\gamma_{0,overall}}$, $s = \sqrt{N_1 N_2}SE(\widehat{\gamma_{0,overall}})$

$$\gamma_{11}|\boldsymbol{\theta}_{-\gamma_{11}}, z_1 z_2, LD \sim N\left(\frac{1}{\frac{1}{s^2}+\frac{Q_1}{\sigma_1^2}}\left(\frac{m}{s^2}+\frac{\sum_{I:C_i=1} z_{1i}z_{2i}}{\sigma_1^2}\right), \left(\frac{1}{s^2}+\frac{Q_1}{\sigma_1^2}\right)^{-1}\right)$$

Where m=$\widehat{\gamma_{1,Final}}$, $s = \sqrt{N_1 N_2}SE(\widehat{\gamma_{1,overall}})$

$$\gamma_{12}|\boldsymbol{\theta}_{-\gamma_{12}}, z_1 z_2, LD \sim N\left(\frac{1}{\frac{1}{s^2}+\frac{Q_1}{\sigma_2^2}}\left(\frac{m}{s^2}+\frac{\sum_{I:C_i=2} z_{1i}z_{2i}}{\sigma_2^2}\right), \left(\frac{1}{s^2}+\frac{Q_1}{\sigma_2^2}\right)^{-1}\right)$$

Where m=0, $s = 100$

$$\sigma_1^2|\boldsymbol{\theta}_{-\sigma_1^2}, z_1 z_2, LD \sim IG\left(\frac{Q_1}{2}, \sum_{i\in C_i=1}\frac{[z_{1i}z_{2i}-(\gamma_{01}+\gamma_{11}LD_i)]^2}{2}\right)$$

$$\sigma_2^2|\boldsymbol{\theta}_{-\sigma_2^2}, z_1 z_2, LD \sim IG\left(\frac{Q_1}{2}, \sum_{i\in C_i=2}\frac{[z_{1i}z_{2i}-(\gamma_{01}+\gamma_{11}LD_i)]^2}{2}\right)$$

The stationary distribution obtained for Gibbs Sampling algorithm uses the following steps to approximate the joint posterior distribution:

$$p(\pi, C, \sigma_1^2, \sigma_2^2, \gamma_{01}, \gamma_{11}, \gamma_{02}, \gamma_{12} | z_1 z_2, LD) \propto$$

$$p(z_1 z_2 |\, \pi, C, \sigma_1^2, \sigma_2^2, \gamma_{01}, \gamma_{11}, \gamma_{02}, \gamma_{12})\; p(C|\pi) p(\pi) *$$

$$p(\gamma_{01}) p(\gamma_{02}) p(\gamma_{11}) p(\gamma_{12}) p(\gamma_{12}) *$$

$$p(\sigma_1^2)\; p(\sigma_2^2)$$

1. Initialize values: $\pi^{(0)}, \gamma_{01}^{(0)}, \gamma_{02}^{(0)}, \gamma_{11}^{(0)}, \gamma_{12}^{(0)}, {\sigma^2}_1^{(0)}, {\sigma^2}_2^{(0)}$.
2. For each SNP i, update sampling C_i from its full conditional: $C^{(m+1)} \sim C|\boldsymbol{\theta}_{-C}^{(m)}, z_1 z_2, LD$
3. Update sampling $\pi^{(m+1)} \sim\ \pi^{(m)}|\boldsymbol{\theta}_{-\pi}, z_1 z_2, LD$
4. Update sampling σ_1^2 and σ_2^2
5. Update sampling $\gamma_{01}^{(m+1)}, \gamma_{02}^{(m+1)} \sim \gamma_{01}^{(m)}|\boldsymbol{\theta}_{-\gamma_{01}}^{(m)}, z_1 z_2, LD, \gamma_{02}^{(m)}|\boldsymbol{\theta}_{-\gamma_{01}}^{(m)}, z_1 z_2, LD$
6. Update sampling $\gamma_{11}^{(m+1)}, \gamma_{12}^{(m+1)} \sim \gamma_{11}^{(m)}|\boldsymbol{\theta}_{-\gamma_{11}}^{(m)}, z_1 z_2, LD, \gamma_{12}^{(m)}|\boldsymbol{\theta}_{-\gamma_{12}}^{(m)}, z_1 z_2, LD$
7. Order $\gamma_{11}^{(m+1)}, \gamma_{12}^{(m+1)}$ and arrange each parameter with that order
8. Repeat

In the Gibbs sampling algorithm, the first 1000 iterations are discarded as burn-in and the chain is repeated for 10,000 iterations.

4.2.4 Parameters of Interest

The primary parameter of interest is the probability of correlated group membership which is approximated from directly sampling from the stationary distribution:

$$P(C_i = j \mid z_{1i}z_{2i}, LD_i, \boldsymbol{\theta}) = \frac{\pi * \phi(z_{1i}z_{2i}|LD_i, \theta_1)}{\pi * \phi(z_{1i}z_{2i}|LD_i, \theta_1) + (1-\pi) * \phi(z_{1i}z_{2i}|LD_i, \theta_2)}$$

$$\approx \sum_{m}^{M} I(C_i^{(m)} = j)$$

We label a SNP as being in the genetically correlated group if $P(C_i = 1 \mid z_{1i}z_{2i}, LD_i, \boldsymbol{\theta}) = 1$. The high probability required was chosen to avoid false positives (at the risk of missing potential correlated SNPs). In real data applications, to identify pleiotropy at the gene level in a region we first annotate each SNP to a gene in the region using the *snpsettest* R package with human gene locations extracted from GENCODE release 19 (build GRCh37). Using a frequentist post-analysis step, with inferred group labels we fit a GEE logistic regression model to test for an enrichment of correlated SNPs in each gene:

$$C = \beta_0 + \beta_1 Gene_1 + \cdots + \beta_k Gene_k$$

Where C is the inferred group label and $Gene_k$ are the k genes in the region. We use the GEE framework with an exchangeable correlation structure to produce robust standard errors accounting for correlation among SNPs.

4.2.5 Simulation Settings

We followed similar simulation settings first defined by Shi et al[38]. We first simulated 50,000 European genotypes using HAPGEN2[146] with chromosome 1 haplotypes from n=503 phased Europeans from the 1000 Genomes Project[3]. Variants with minor allele frequency < 1% were removed. The sample of simulated genotypes was divided in half, such that 25,000 genotypes were used to simulate trait effect sizes and 25,000 were used as an external LD reference panel. Chromosome 1 genotypes were then partitioned into 133 approximately independent LD blocks defined in European populations[147].

For each LD block, we simulated GWAS effect sizes across two quantitative traits. To simulate phenotype data, we assumed the linear model $Y = XB + \epsilon$. True causal effect sizes were drawn from a multivariate normal $\boldsymbol{\beta} \sim MVN_2(\mathbf{0}, \begin{pmatrix} \frac{h_1^2}{m} & \frac{\rho h^2}{m} \\ \frac{\rho h^2}{m} & \frac{\rho h_2^2}{m} \end{pmatrix})$ where $h_1^2 = h_2^2 = \{0.001, 0.005, 0.01\}$ was the local SNP-based heritability, $m = \{1, 5\}$ is the set of causal SNPs, and $\rho = \{0.5, 0.8\}$ is the genetic correlation (only moderate to large values considered because LDSC-MIX is a follow-up method for regions of LGC). All other SNP effect sizes were set to zero. The true genetic score G for each trait was then defined as the product of sampled causal genotypes and their respective simulated effect sizes ($X = \sum_{i=1}^{m} G_i \beta_i$), standardized to ensure total heritability of h^2: $G = \frac{X - \mu_x}{\sigma_x} * h^2$. Environmental noise was separately generated for each trait i from a normal distribution $\epsilon \sim N(0, 1 - h_i^2)$,), standardized to ensure variance $1 - h^2$: $E = \frac{\epsilon - \mu_\epsilon}{\sigma_\epsilon} * (1 - h^2)$. We then estimated effect sizes $\hat{\beta}$ for each variant genome wide using a linear model $Y = XB + \epsilon$, using each population's respective simulated phenotype and genotype data.

4.2.6 Running Colocalization

We compared the performance of LDSC-MIX to colocalization of GWAS summary statistics across two simulated traits. To run colocalization, we used the R package *coloc*[136] with the extension to allow for multiple shared causal variants using the fine mapping SuSiE[148] framework. Under this multiple causal variant framework, SuSiE is first used to form credible sets of causal variants for each trait separately, with each pair of credible sets across traits then tested for colocalization under the single shared causal variant hypothesis. The default coverage probability for SuSiE is 0.95, though we considered values of {0.10, 0.50, 0.95) to allow for detection of weaker effect colocalizing variants (at the cost of introducing more false positives).

4.2.7 Method Evaluation and Comparison

The objective of LDSC-MIX is to identify the set of SNPs or genes in a locus that drive a signal of local genetic correlation. The first metric we used to assess the performance of LDSC-MIX was the number of simulations LDSC-MIX can successfully identify a candidate set. A candidate set is defined as a set of SNPs that contain potentially true shared causal variants. LDSC-MIX will fail to identify a candidate set when no SNPs have a posterior probability of one for correlated group membership, while coloc-SuSiE will fail when SuSiE fine-mapping fails to identify credible sets for either trait or colocalization fails to run. The second metric was the number of true shared causal variants in the candidate set of SNPs, $X_{c,\mathrm{LDSC-MIX}}$, with SNP-level posterior probability of genetically correlated group membership equal to 1. To compare against colocalization, we rank SNPs by their colocalization posterior probability and define $X_{c,coloc}$ as the top R SNPs (where R is defined to be the cardinality of set $X_{c,\mathrm{LDSC-MIX}}$ to allow equal comparison). Using $X_{c,coloc}$ we then assess the number of true shared causal variants contained.

The last metric used to assess the performance of LDSC-MIX was the similarity in estimated genetic correlation using SNPS in $X_{c,\mathrm{LDSC-MIX}}$ to the true genetic correlation. Because the candidate set of potential shared causal variants $X_{c,\mathrm{LDSC-MIX}}$ contains variants in LD, we first LD clumped variants at a pairwise threshold $R^2 = 0.10$ to produce a set of approximately independent SNPs $X^*_{c,\mathrm{LDSC-MIX}}$. Using the final set of proposed shared causal variant set $X^*_{c,\mathrm{LDSC-MIX}}$, we computed the genetic correlation as $cor(X^*_{c,\mathrm{LDSC-MIX}}\widehat{\beta_1}, X^*_{c,\mathrm{LDSC-MIX}}\widehat{\beta_2})$ and defined our performance metric as the ratio of genetic correlations:

$$RGC = \frac{cor(X^*_{c,LDSC-MIX}\widehat{\beta_1}, X^*_{c,LDSC-MIX}\widehat{\beta_2})}{cor(X_c\beta_1, X_c\beta_2)}$$

RGC is defined as the ratio of the estimated genetic correlation vs the true correlation (using true causal variants X_c and respective effect sizes β). To compare against colocalization, we similarly define R^* as the cardinality of LDSC-MIX pruned set $X^*_{c,LDSC-MIX}$ and subset $X_{c,coloc}$ to $X^*_{c,coloc}$, the R^* top ranked SNPs by colocalization posterior probability. With $X^*_{c,coloc}$ we then similarly compute RGC.

4.2.8 Real Data Application

We evaluated LDSC-MIX using European GWAS summary statistics of varying sample sizes for autoimmune, cancer, and lipid phenotypes from the UK Biobank[6]. Phenotypes trait pairs included asthma-basal cell carcinoma and asthma-HDL. Trait pairs were chosen to consider trait pairs with (asthma-bcc) and without (asthma-HDL) likely signals of shared genetics. We first partitioned the genome into 2045 roughly independent regions using previously estimated LD blocks based on data from the 1000 Genomes Project[147]. For a given trait pairing, to identify candidate regions for follow up analysis with LDSC-MIX, we ran LAVA[37] to estimate local SNP-based heritability for each trait and bivariate local genetic correlation. The top genomic regions with significant heritability for both traits ($p < \frac{0.05}{2045}$) and bivariate genetic correlation ($p < \frac{0.05}{\#\ of\ regions\ with\ significant\ heritability\ for\ both\ traits}$) were included for follow up analysis. For each top genomic region, we used genotype data from 503 European individuals from the 1000 Genomes Project to derive LD scores. LD scores were computed using the *bigsnpr* package from defined PLINK[149] files. LDSC-MIX was used to first infer correlated group membership for each SNP. Correlated group membership labels were then used to test the effect of each gene in the region for an enrichment of correlated group SNPs (see above).

4.3 Results

4.3.1 Comparison of LDSC-MIX and coloc-SuSiE Across Various Disease Architectures

We evaluated the performance of LDSC-MIX using simulated GWAS effect sizes in approximately independent LD blocks under a variety of genetic architectures, varying values of local SNP-based heritability $h^2 = \{0.001, 0.005, 0.01\}$, number of causal variants $m = \{1,5\}$, and true genetic correlation $\rho = \{0.5, 0.8\}$. For comparison, we included colocalization allowing for multiple shared causal variants using the SuSiE framework. We broadly dichotomized simulations into two scenarios 1) strong shared genetic effects corresponding to high local-SNP based heritability ($h^2 = 0.005, 0.01, m = 1,5, \rho = 0.5,0.8$) and 2) moderate or weak shared genetic effects ($h^2 = 0.001\ or\ m = \ 1,5, \rho = 0.5, 0.8$). In simulations of strong shared genetic effects, 91.4% of simulations had a GWAS hit ($p < 5e$-8) in both traits as compared to 28.6% in simulations of weaker shared genetic effects.

Our first performance metric comparing LDSC-MIX and colocalization was the number of instances each method could be run out of n=100 simulated LD blocks across a variety of disease architectures. In scenarios of strong shared genetic effects, LDSC-MIX can be used to estimate a candidate set of shared causal SNPs set in more scenarios (85.6%) when compared to coloc-SuSiE (50.1%). For LDSC-MIX, identified candidate sets were generally large (average 338.8 SNPs) as our framework does not fine-map the signal among a set of correlated variants. On the other hand, coloc-SuSiE performs SuSiE fine-mapping prior to colocalization analysis to remove correlated SNPs (though we use equal number of SNPs for method comparison metrics (see methods)). In simulations of weaker shared genetic effects corresponding to fewer shared GWAS hits across both traits, both LDSC-MIX and coloc-SuSie succeeded in fewer cases (64.4% and 28.7%). As failure of coloc-SuSiE to run is dependent on successful fine-mapping,

we considered three values of coverage probability (0.10, 0.50, and 0.95) for formation of a credible set in SuSiE (Figure 4.1). In simulations of weak genetic effects, at the default 0.95 coverage probability coloc-SuSiE ran successfully for only 20.6% of cases. As the coverage probability was lowered, coloc-SuSiE successfully identified a candidate set more often and ran in 20.6%, 32.5%, and 32.8% of simulations for coverage probabilities of 0.95 (default SuSiE value), 0.50, and 0.10.

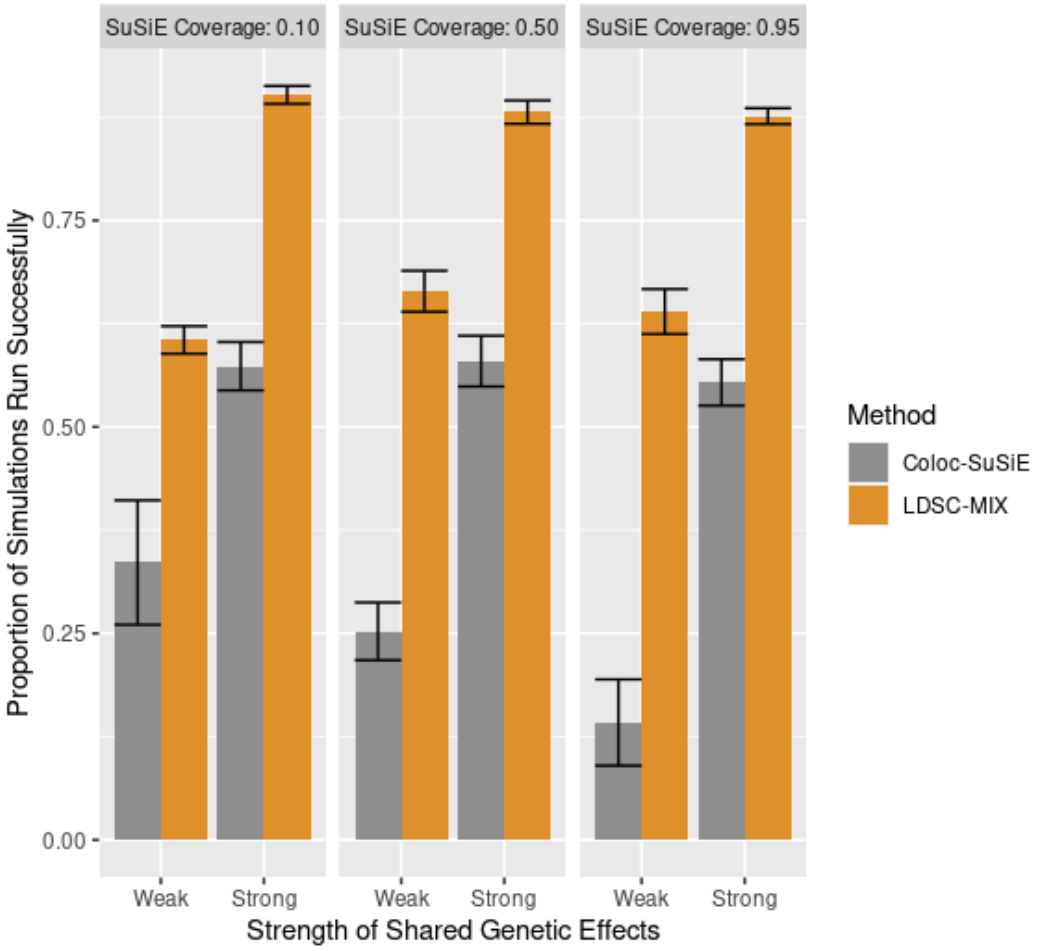

Figure 4.1 Barchart comparing proportion of simulations (strong vs weak genetic effects) each method (Coloc-Susie vs LDSC-MIX) can successfully identify a candidate set. Columns correspond to coverage probability for SuSiE fine mapping in Coloc-SuSiE.

We quantified how well the candidate set of potential shared causal SNPs identified by both methods recovered the true genetic correlation and contained the true number of shared

causal variants. Here, simulations were filtered to scenarios where both methods could be run successfully for direct method comparison. Furthermore, we did not dichotomize simulation scenarios into “weak” and “strong” to interrogate specific disease architectures (number of shared causal variants, local genetic correlation, and local SNP-based heritability) for method comparison. For the ratio of genetic correlations (RGC, see methods), when only a single causal variant was present LDSC-MIX performed poorly compared to coloc-SuSiE (Figure 4.2). In such single causal variant settings, the mean RGC (with a RGC = 1 indicating perfect capturing of shared causal SNPs) for LDSC-MIX and coloc-SuSiE was 0.83 and 0.99 across varying local SNP-based heritability and genetic correlations. LDSC-MIX had the worst mean RGC (0.55) when the single shared genetic signal was weakest ($h^2 = 0.001$, ρ=0.50), but comparable to the mean RGC of coloc-SuSie (0.90 vs 0.99) when the single shared genetic signal was strong ($h^2 = 0.01$). On the other hand, when multiple causal variants were present in a region estimated RGC were closer to 1 for LDSC-MIX (1.10) compared to coloc-SuSiE (1.14) across both weak and strong scenarios of shared signals. Improvement was most notable in scenarios of weaker genetic effects ($h^2 = 0.001$ or $h^2 = 0.005$ & ρ=0.50), in which mean RGC for coloc-SuSiE (1.31) were larger than LDSC-MIX (1.07) indicating coloc-SuSiE tended to overestimate the true genetic correlation. However, when multiple shared strong genetic effect variants were present coloc-SuSiE and LDSC-MIX both performed well.

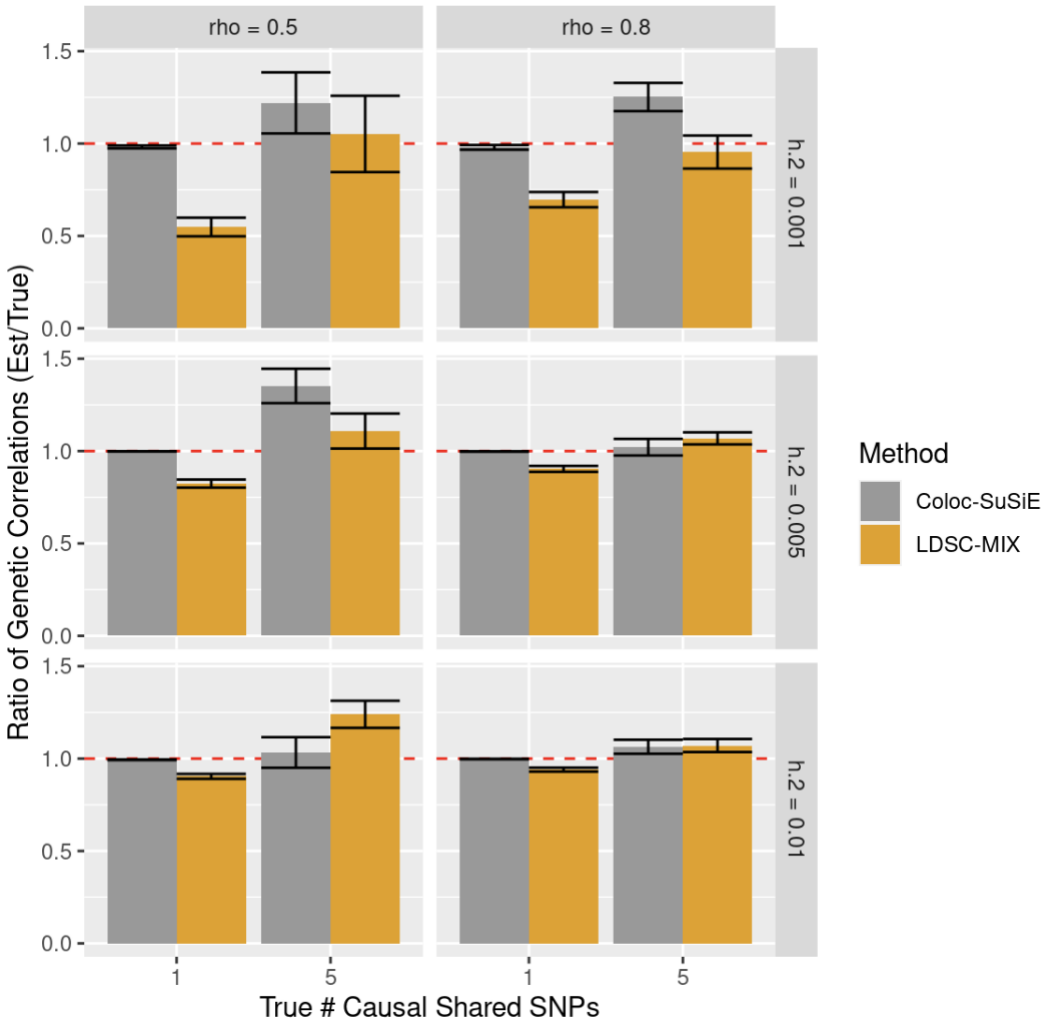

Figure 4.2 Bar chart comparing ratio of genetic correlations by method (Coloc-Susie vs LDSC-MIX). Columns correspond to local genetic correlation in region and rows correspond to local SNP-based heritability.

When comparing the number of true causal variants contained in candidate sets, across simulations LDSC-MIX (following pruning of candidate sets) tended to contain more true shared risk variants compared to coloc-SuSiE, though both tended to under contain the true number of causal variants (Figure 4.3). When there was a single shared causal variant, on average LDSC-MIX contained 0.75 variants compared to 0.51 for coloc-SuSiE. When there were multiple shared causal variants (five), LDSC-MIX similarly contained more true causal variants (2.09) compared to coloc-SuSiE (1.53). However, our metric included scenarios where either of the

methods could not be run (resulting in zero true risk variants contained) and thus coloc-SuSiE had more density of zero contained variants due to more simulation scenarios failing (Figure 4.3). When we consider only cases in which both LDSC-MIX and coloc-SuSiE could be run (i.e., scenarios of clear shared genetic signals), coloc-SuSiE tended to contain more true shared causal variants. In such restricted successful settings, for both the single and multiple shared causal risk variant scenarios, the average number of contained true variants for coloc-SuSiE (0.99 and 3.76) compared to LDSC-MIX (0.94, 2.61) was closer to the true number (one and five). When assessing the effect of disease architecture parameters across all simulations (both successful and unsuccessful), increasing local SNP-based heritability generally had the largest improvement on containment for both LDSC-MIX and coloc-SuSiE. For a single shared causal variant, increasing values of local SNP-based heritability $h^2 = \{0.001, 0.005, 0.01\}$ had increasing containment for LDSC-MIX (0.54, 0.85, 0.88) and coloc-SuSiE (0.33, 0.57, 0.64) as shared genetic signals were larger and easier to identify. A similar pattern was observed for multiple shared causal variants (five) across both methods LDSC-MIX (1.28, 2.29, 2.71) and coloc-SuSiE (0.39, 2.23, 1.97).

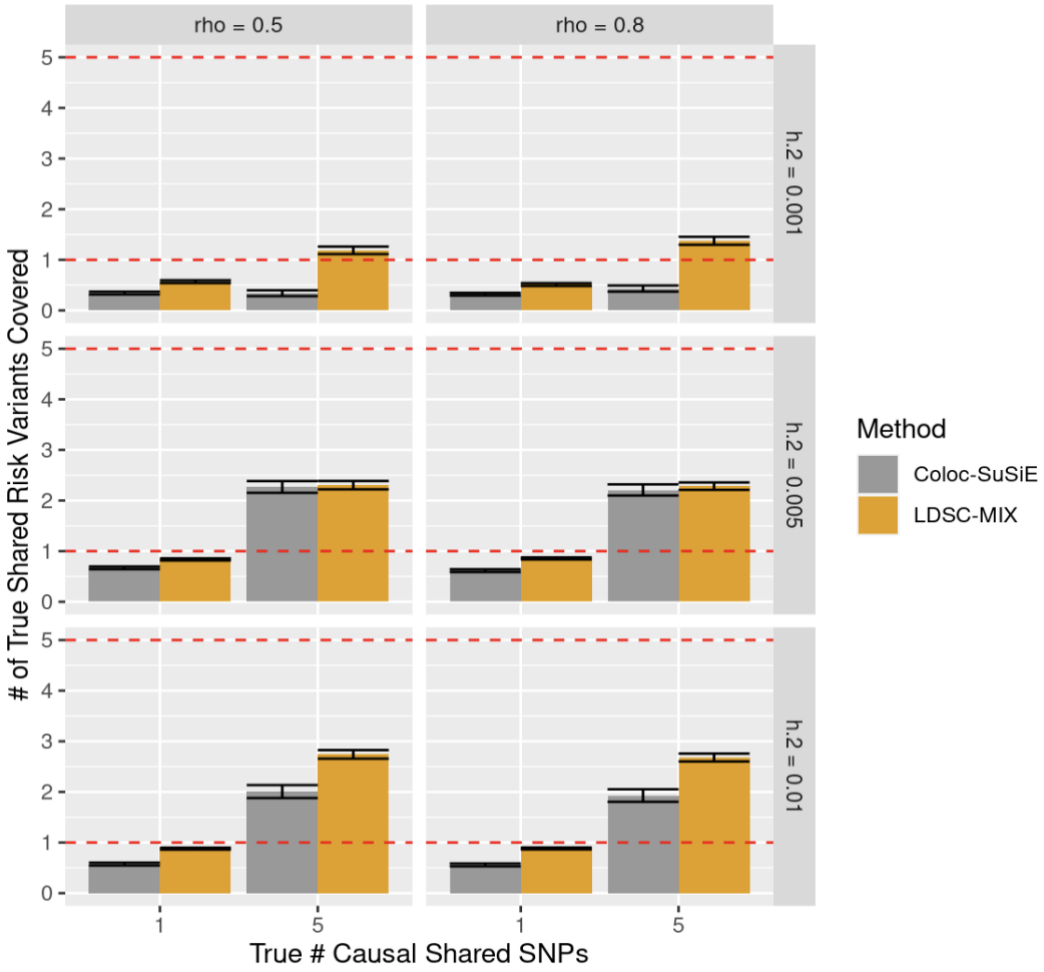

Figure 4.3 Bar chart comparing true # of causal shared SNPs contained by Coloc-SuSiE and LDSC-MIX by local genetic correlation and local SNP-based heritability. Red horizontal dotted lines correspond to true number of risk variants simulated (1, 5).

4.3.2 Applications of LDSC-MIX in Traits from the UK Biobank

We demonstrated the application of LDSC-MIX in empirical data using GWAS summary statistics from the UK Biobank for two trait pairings across three traits: asthma-basal cell carcinoma and asthma-HDL. Meta-data for each GWAS study are presented in Table 4.1. We first estimated bivariate genetic correlations using LAVA across 2045 independent LD blocks, with loci having significant SNP-based heritability and local genetic correlation identified as candidate regions for LDSC-MIX. For each trait pairing, we focus on the most significant LGC

regions with significant local SNP-based heritability (see methods). Because in real data we do not know the true causal shared risk variants and local genetic correlation, we evaluated LDSC-MIX's ability to recover the LAVA estimated local genetic correlation and identify genes enriched in SNPs with high posterior probability of correlated group membership.

Phenotype	Sample Size (case/control)	Phenotype Pair	# Significant LGC Regions
Asthma	80070/11717	-	-
Basal Cell Carcinoma	3441/357700	Asthma-BCC	4
High Density Lipoprotein	41234	Asthma-HDL	2

Table 4.1 Phenotypes analyzed from the UK Biobank with sample size denoted. # of significant LGC regions for phenotype pairs correspond to regions with significant individual local trait SNP-based heritability and bivariate correlation.

For the asthma-basal cell carcinoma pairing, the top LAVA region (excluding the MHC region due to complex LD patterns) was chr14:24906057-25585041 containing 2852 SNPs with local SNP-based heritability $h^2_{asthma} = 0.002$, $h^2_{bcc} = 0.002$, estimated LGC $\hat{\rho} = -1.0$, and LGC p-value $p = 1.60 \times 10^{-6}$. The top GWAS hits for both traits had p values of $1.9 * 10^{-3}$ and $2.1 * 10^{-4}$ for asthma and BCC respectively. Estimated local heritability and lack of GWAS hits in the region correspond to the "weak shared genetic effect" scenario for our simulations. We fit LDSC-MIX in the region and identified 39 variants with high posterior probability of belonging to the correlated SNP set, producing an estimated LGC of -1.0 recovering the LAVA estimated value. In this region, coloc-SuSiE was unable to be run owing to lack of GWAS hits in the region preventing formation of credible sets for asthma and bcc respectively. Relaxing to the single shared causal variant assumption, colocalization suggested only a 0.016% posterior

probability of a shared variant. The region contained nine, potentially overlapping, genes which we annotated each SNP with if it fell in a gene region and tested the annotation matrix on inferred correlated vs uncorrelated group. SNPs belonging to the CBLN3 gene were significantly associated with belonging in the inferred correlated group (p = 0.0096). CBLN3 is a protein coding gene that is known to be associated with certain types of breast cancer and Ganglioneuroblastoma, though its effect on asthma and basal cell carcinoma is unknown.

We repeated the analysis for the asthma-HDL trait pairing and found the top LAVA region (again excluding the MHC regions) was chr2:29627933-30575619 containing 3549 SNPs with local $h^2_{asthma} = 0.0005$, $h^2_{HDL} = 0.0004$, and $\rho = -0.89, p = 9.5 * 10^{-4}$. The top GWAS hits for both traits in the region correspond to p-values of $8.47 * 10^{-5}$ and $6.51 * 10^{-6}$, while the product of GWAS z scores for the top shared SNP chr2:30478185 was -8.58. From LDSC-MIX, we identified 1,226 variants with high posterior probability of correlated group membership, producing an estimated LGC of -0.59. Coloc-SuSiE under the single variant assumption had a colocalization posterior probability of 0.83%. Allowing for multiple causal shared variants at a SuSiE coverage of 0.50 was unable to run. The region contained three potentially overlapping genes: ALK, YPEL5, LBH. SNPs belonging to the ALK ($OR = 2.37, p = 4.23 * 10^{-17}$) and YPEL5 ($OR = 1.81, p = 5.42 * 10^{-3}$) had a significant association with belonging in the inferred correlated group.

4.4 Discussion

In this work we proposed a novel genetic correlation-based framework to identify SNP-level pleiotropy at a locus using GWAS summary statistics from two phenotypes. Our method, LDSC-MIX, extends the bivariate LD score regression by assuming two underlying latent groups

in a region: a set of genetically correlated variants and a set of uncorrelated variants. LDSC-MIX then fits an empirical Bayesian mixture of regressions to estimate posterior probabilities of group membership. Inferred group labels can then be used to test for enrichment of correlated SNPs in annotated genes to test for gene-level pleiotropy driving signals of LGC. In extensive simulations, we compared LDSC-MIX to a SNP-level colocalization method, coloc-SuSiE, with respect to number of simulations each method could be run successfully, true number of shared risk variants contained by identified candidate sets, and how well the true LGC could be recovered by identified candidate sets. We found LDSC-MIX generally outperformed coloc-SuSiE in cross-trait genetic architectures with multiple shared weak genetic effects, though colocalization was the preferred approach in single shared causal variant scenarios or multiple strong genetic effect variants. In real data applications from two trait pairings from the UKBB: asthma-basal cell carcinoma and asthma-HDL we found regions of significant bivariate genetic correlation using LAVA. In the top LGC regions, LDSC-MIX was able to identify SNP sets that recover LAVA estimated LGC and implicated potential shared genes while coloc-SuSiE failed to run.

LDSC-MIX uses the framework of local genetic correlations to identify SNP-level pleiotropy as compared to colocalization, which does the same through identifying shared GWAS hits after accounting for LD (when multiple shared causal variants are assumed). However, their implementation differs as colocalization relies on GWAS hits (and LD information) to identify shared causal variants while LDSC-MIX uses the linear relationship between the product of z scores with LD scores. Local genetic correlation can be quantified even in the absence of GWAS hits and thus LDSC-MIX was preferable in the presence of multiple weak genetic signals. In such situations without strong GWAS hits, coloc-SuSiE was hampered

due to frequent failed fine-mapping via SuSiE across a variety of coverage probabilities. Variant sets identified by LDSC-MIX tended to contain more true shared risk variants across simulations (as coloc-SuSiE had a heavy density of zero contain variants due to failed scenarios) though estimated LGC were similar. Conversely, when a single shared causal variant was present colocalization was the better approach compared to LDSC-MIX. In such cases, LDSC-MIX likely struggles as either the single shared causal variant does not tag enough variants (low LD score) or its genetic signal is too weak to inflate tagged variant effect sizes and fails to distinguish separate latent correlated and uncorrelated groups. The SNP set identified by LDSC-MIX, especially when the single shared causal variant had weak genetic effects, often underestimated the true local genetic correlation. In such cases, the two latent groups are likely not separated well and thus LDSC-MIX incorrectly contains non-causal variants. Lastly, when underlying shared genetic effects are strong, as expected colocalization (when able to run) outperformed LDSC-MIX with respect to containing true shared risk variants and capturing the true LGC due to strong distinguishable GWAS hits.

In our real data applications, for the asthma-bcc trait pairing we identified CBLN3 as a potential shared gene in the top significantly correlated region chr14:24906057-25585041. For asthma, CBLN3 was found to be differentially expressed in a study[150] of BAL cells from horses with and without evidence of respiratory disease. While no direct evidence for CBLN3 exists for bcc specifically, CBLN3 has been suggested to play a role in cancer proliferation through involvement in synaptic functions[151,152]. For the asthma-HDL trait pairing we found genes ALK and YPEL5 to be potential shared genes in the top region chr2:29627933-30575619. ALK is a protein coding gene involved in cell growth, with known implications in cancers including non-small cell lung cancer[153], though its effect on asthma and HDL are unknown. YPEL5 is a gene in

the YPEL family, a highly conserved protein coding genes involved in zinc-finger-like metal-binding domains[154]. The family has been suggested to play a role in both immune/pulmonary response and HDL[154,155]. We note that while such highlighted genes may play a role in the shared etiology of the considered trait pairings, future studies are needed to validate a shared biological mechanism.

LDSC-MIX is the first approach to bridge the gap between local genetic correlation, which quantifies the strength of genetic sharing, and methods to identify SNP and gene-level pleiotropy such as colocalization. In doing so, LDSC-MIX provides interpretability of which specific set of SNPs or genes is driving a signal of local genetic correlation. LDSC-MIX is unique in that no assumptions on the existence of GWAS hits or the number of causal variants is made, and we show in our simulations that LDSC-MIX is preferable in situations of multiple weaker genetic signals. This specific weaker genetic signal scenario where local genetic correlations have improved benefits colocalization has been suggested by Werme et al[37]. We reaffirm their findings and provide a statistical framework to leverage this specific scenario. We suggest that future studies investigating pleiotropy across two traits in highly polygenic regions with weaker genetic effects to use our approach. Thus, an overall analysis of pleiotropic effects in a local region with identified local genetic correlation may consist of 1) A first pass using colocalization if strong genetic signals are present and driving LGC and 2) A second pass using LDSC-MIX if colocalization failed in the absence of strong shared signals. Overall, our proposed approach and findings prompt consideration of identifying pleiotropy across traits at increasingly polygenic architectures. If we consider the omnigenic[156] model in the context of pleiotropy, approaches such as colocalization may be well suited for identifying strong shared core genes

while LDSC-MIX may be beneficial to interrogate peripheral genes with weaker effects (i.e. genes effect tissues that contribute to both disease's risk).

Our method LDSC-MIX has limitations to consider in applications. The first is while LDSC-MIX identifies variant sets contributing to shared genetic signal, suggestive variants can be in high LD making identification of the true shared genetic variants difficult (though identification of shared genes is less problematic). While we performed post-group identification pruning, it's entirely possible the top hit among SNPs in proximity is not the true causal variant. A secondary post fine mapping analysis carefully modeling the LD patterns in the region could be done to further narrow down the signal. Second, LDSC-MIX does not distinguish scenarios where a high productive of z-scores is observed due to both trait z-scores being large vs a single trait z-score is large while the other is small (not true pleiotropy). We restrict LDSC-MIX use to regions with significant LGC making this scenario likely less common. An additional correlated group membership criteria of z-scores > minimum cutoff could be implemented, though in doing so may miss weaker effects. Third, to aid in computational efficiency, LDSC-MIX uses conjugate priors to efficiently derive full conditionals which allow for Gibbs Sampling. More informative priors such as a skewed prior placing higher weight on larger values for the slope of the correlated group (reflecting belief that the empirical bayes approach using all variants to get a starting slope is underestimated) could be used instead of our averaging approach. Lastly, in simulations variant effects were drawn from a polygenic model that by chance can result in scenarios of null effects. Given parameter settings, this scenario is likely to be infrequent though an additional LGC estimation step (e.g., LAVA) could be used to focus on simulations with detectable true shared genetic effects.

LDSC-MIX provides a novel framework to bridge the gap between local genetic correlation analysis and SNP-level pleiotropy analysis. While colocalization can be used to highlight specific single or multiple strong shared genetic in LGC regions, LDSC-MIX provides improved shared causal variant detection in the presence of multiple weak genetic effects. Through identifying potentially weaker shared genetic signals, LDSC-MIX allows for the biology across (polygenic) traits to be studied at a resolution previously inaccessible. Furthermore, as sample sizes continue to increase for GWAS studies, more weakly pleiotropic local regions should become identifiable resulting in additional scenarios where LDSC-MIX is the preferred approach.

4.5 Chapter 4 Appendix

4.5.1 Derivation of Full Conditionals

For derivations of full conditionals, we generally express the joint posterior distribution as the product of the likelihood times the joint prior. Assumed independence of priors allows us to easily remove terms not involving the parameter of interest. Furthermore, in general our choice of prior distributions allows easy recognition of the form for many posterior parameters.

Recall the likelihood can be expressed as:

$$p(z_{1i}z_{2i}|\boldsymbol{\theta}) = \prod_{i=1}^{Q}\left\{\frac{\pi}{\sqrt{2\pi}\sigma}\exp\left(\frac{[z_{1i}z_{2i} - (\gamma_{01} + LD_i\gamma_{11})]^2}{2W_i\sigma_1^2}\right)\right\}^{C_i}$$

$$* \left\{\frac{1-\pi}{\sqrt{2\pi}\sigma}\exp\left(\frac{[z_{1i}z_{2i} - (\gamma_{02} + LD_i\gamma_{12})]^2}{2W_i\sigma_2^2}\right)\right\}^{1-C_i}$$

1. Full conditional for $C_i \sim Categorical(\pi)$

$$p(C_i|\dots) \propto p(\pi, C_i, \sigma_1^2, \sigma_2^2, \gamma_{01}, \gamma_{11}, \gamma_{02}, \gamma_{12}, z_1z_2, LD)$$

$$\propto_{C_i} p(z_1z_2|\dots)p(C_i|\dots)$$

$$= N(z_{1i}z_{2i}|\dots)\pi$$

$$\propto_{C_i} Categorical(w_i)$$

Where $w_i = \frac{\pi N(z_1z_2|\dots)}{\sum \pi_j N(z_1z_2|\dots)}$

2. Full conditional for $\pi \sim Dirichlet(\alpha)$:

$$p(\pi|\dots) \propto_{\pi} p(\pi, C_i, \sigma_1^2, \sigma_2^2, \gamma_{01}, \gamma_{11}, \gamma_{02}, \gamma_{12}, z_1z_2, LD)$$

$$\propto_{\pi} p(C|\pi, \dots)p(\pi)$$

$$\propto_{\pi} (\prod_{i=1}^{n} \pi)(\prod_{k=1}^{K} \pi_k^{\alpha})$$

$$\propto Dirichlet(\alpha + n_k)$$

Where $n_k = \sum_{i=1}^{n} I(C_i = k)$

3. Full conditionals for $\gamma_{01}, \gamma_{02} \sim N(\gamma_{\widehat{0,overall}}, \sqrt{N_1 N_2} SE(\widehat{\gamma_{0,overall}}))$

Let prior hyperparameters for the prior mean and variance be denoted: $m = \gamma_{\widehat{0,overall}}$, $s = \sqrt{N_1 N_2} SE(\widehat{\gamma_{0,overall}})$

$$p(\gamma_{01} | \dots) \propto p(z_1 z_2 | \dots) p(\gamma_{11})$$

$$\propto \exp\left[- \sum_{i \in C_i = 1} \frac{[z_{1i} z_{2i} - (\gamma_{01} + \gamma_{11} LD_i)]^2}{2\sigma_1^2} \right] * \exp\left[- \frac{(\gamma_{01} - m)^2}{2s^2} \right]$$

Using conjugacy of normal prior for the normal distribution we know this has a normal distribution with the form:

$$\propto N\left(\frac{1}{\frac{1}{s^2} + \frac{n_1}{\sigma_1^2}} \left(\frac{m}{s^2} + \frac{\sum_{I:C_i=1} z_{1i} z_{2i}}{\sigma_1^2} \right), \left(\frac{1}{s^2} + \frac{n_1}{\sigma_1^2} \right)^{-1} \right)$$

4. Full conditionals for $\gamma_{11} \sim N(\gamma_{\widehat{1,Final}}, \sqrt{N_1 N_2} SE(\widehat{\gamma_{1,overall}}))$ and $\gamma_{12} \sim N(0,100)$

Let prior hyperparameters for the prior mean and variance be denoted: $m = \gamma_{\widehat{1,Final}}$, $s = \sqrt{N_1 N_2} SE(\widehat{\gamma_{1,overall}})$

$$p(\gamma_{11} | \dots) \propto p(z_1 z_2 | \dots) p(\gamma_{11})$$

$$\propto \exp\left[- \sum_{i \in C_i = 1} \frac{[z_{1i} z_{2i} - (\gamma_{01} + \gamma_{11} LD_i)]^2}{2\sigma_1^2} \right] * \exp\left[- \frac{(\gamma_{11} - m)^2}{2s^2} \right]$$

Using conjugacy of normal prior for the normal distribution we know this has a normal distribution with the form:

$$\propto N\left(\frac{1}{\frac{1}{s^2}+\frac{n_1}{\sigma_1^2}}\left(\frac{m}{s^2}+\frac{\sum_{I:C_i=1} z_{1i}z_{2i}}{\sigma_1^2}\right),\left(\frac{1}{s^2}+\frac{n_1}{\sigma_1^2}\right)^{-1}\right)$$

The derivation for γ_{12} follows the same procedure but instead let prior hyperparameters be denoted: m=0, $s = 100$

$$p(\gamma_{11}|\dots) \propto p(z_1 z_2|\dots)p(\gamma_{11})$$

$$\propto \exp\left[-\sum_{i\in C_i=2}\frac{[z_{1i}z_{2i}-(\gamma_{02}+\gamma_{12}LD_i)]^2}{2\sigma_2^2}\right] * \exp\left[-\frac{(\gamma_{11}-m)^2}{2s^2}\right]$$

Using conjugacy of normal prior for the normal distribution we know this has a normal distribution with the form:

$$\propto N\left(\frac{1}{\frac{1}{s^2}+\frac{n_1}{\sigma_2^2}}\left(\frac{\sum_{I:C_i=1} z_{1i}z_{2i}}{\sigma_2^2}\right),\left(\frac{1}{s^2}+\frac{n_1}{\sigma_2^2}\right)^{-1}\right)$$

5. Full conditionals for Jeffrey's priors σ_1^2, σ_2^2

$$p(\sigma_1^2|\dots) \propto p(z_1 z_2|\dots)p(\sigma_1^2)$$

$$\propto \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n_1} \exp\left[-\sum_{i\in C_i=1}\frac{[z_{1i}z_{2i}-(\gamma_{01}+\gamma_{11}LD_i)]^2}{2\sigma_1^2}\right] * \frac{1}{\sigma_1^2}$$

$$(\sigma_1^2)^{-(\frac{n1}{2}+1)} * \exp\left[-\sum_{i\in C_i=1}\frac{[z_{1i}z_{2i}-(\gamma_{01}+\gamma_{11}LD_i)]^2}{2\sigma_1^2}\right]$$

$$\sim InverseGamma(\frac{n_1}{2}, \sum_{i\in C_i=1}\frac{[z_{1i}z_{2i}-(\gamma_{01}+\gamma_{11}LD_i)]^2}{2})$$

And similarly, for σ_2^2:

$$p(\sigma_2^2|\ldots) \propto p(z_1 z_2|\ldots)p(\sigma_2^2)$$

$$\propto \left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{n_1} \exp\left[-\sum_{i\in C_i=1}\frac{[z_{1i}z_{2i}-(\gamma_{01}+\gamma_{11}LD_i)]^2}{2\sigma_2^2}\right] * \frac{1}{\sigma_2^2}$$

$$(\sigma_2^2)^{-(\frac{n1}{2}+1)} * \exp\left[-\sum_{i\in C_i=1}\frac{[z_{1i}z_{2i}-(\gamma_{01}+\gamma_{11}LD_i)]^2}{2\sigma_2^2}\right]$$

$$\sim InverseGamma(\frac{n_1}{2}, \sum_{i\in C_i=2}\frac{[z_{1i}z_{2i}-(\gamma_{01}+\gamma_{11}LD_i)]^2}{2})$$

4.5.2 Code Availability

An R package for LDSC-MIX has been developed with code and example workflow available at:

https://github.com/kliao12/LDSC-MIX

Chapter 5 Discussion

This dissertation has tackled several current problems in the field of population and statistical genetics. While the three chapters presented were diverse, they had the unifying theme of analyzing commonly used summary statistics of genetic variation or genetic-phenotypic associations. Summary statistic-based approaches for population and statistical genetics are attractive due to their computational tractability and ease of data sharing, though require careful analysis. In this section, we review the key findings of each of the three projects and discuss the broader implications and questions raised to the field.

The first project dealt with investigating allele frequency spectrum (AFS) heterogeneity across 3-mer mutation subtypes and potential biases in AFS-based methods for population genetics inference that assume an interchangeability of sites. While previous studies have researched AFS heterogeneity across sites at a broader resolution, this study presented the first detailed investigation on how violations in the assumed interchangeability of sites can impact population genetics inference. We found AFS-heterogeneity across 3-mer subtypes driven by forces such as mutation rate heterogeneity and biased gene conversion was sufficient to impact demographic inference and shape the local AFS in a region. Our findings have immediate relevance on the broader field of population genetics regarding whether consideration of mutation subtypes can improve inference for both AFS-based and non-AFS based frameworks. For AFS-based frameworks, a natural question is whether an expected AFS can be estimated, devoid of confounding due to parallel mutations or biased gene conversion, based on the

observed composition of subtypes. For example, as we highlighted in the study, parallel mutations are a function of mutation rate. If the expected number of parallel mutations could be estimated conditional on the observed site-specific mutation rate, distortions in the extremely rare variant counts could be adjusted. For non-AFS based frameworks, popular methods such as PSMC assume a constant mutation rate and the singleton density score is reliant on distance between singletons. Such frameworks could be adapted to use region-specific mutation rates estimated using local sequence context and genomic factors[47], as well as consider post-hoc whether there exists an increased density of high singleton mutation subtypes.

Of methodological note, we introduced a novel D-type test of neutrality statistic D_{-2} that removes the contribution of singletons and doubletons. D_{-2} derives unbiased estimators of the population genetics parameter θ (Watterson's estimator and mean pairwise difference) to summarize the high dimensional AFS without distortions in the extremely rare variant counts due to parallel mutations. Our derived framework is important because the issue of parallel mutations is becoming increasingly prevalent as genetic datasets become increasingly large. While we focused on the singleton and doubleton counts being distorted, ever increasing sample sizes can cause distortions in higher allele counts such as tripletons and onwards. Lastly, while in our analysis we used D_{-2} primarily for summarizing the high dimensional AFS for comparison across subtypes, D_{-2} could potentially be used for AFS-based tests of neutrality without confounding in the singleton and doubleton counts. Though we warn most of the genetic variation is contained in the count of extremely rare variants causing our statistic to likely be underpowered.

In the second project we introduced a method slaPRS (stacking local ancestry PRS), a stacking-based framework to construct polygenic risk scores in admixed individuals. Construction in polygenic risk scores for admixed individuals is a challenging problem due to uncertainty in what population GWAS to use, how risk variants are selected and weighted, how to accommodate differing ancestral backgrounds for a risk variant, and how to handle an existing ancestry dependence of PRS performance. An existing approach by Marquez et al integrated multiple population GWAS to construct PRS in admixed individuals through stacking global ancestry specific PRS. slaPRS extends their approach by providing a more flexible model that stacks ancestry specific PRS locally rather than globally and models local ancestry. Using population-specific GWAS for variants in a region, local ancestry-specific PRS are constructed and then stacked across the genome in a penalized regression model. We found in simulations for admixed African Americans, slaPRS outperformed single ancestry PRS and the global stacked approach while also reducing the ancestry dependence across ancestry quantiles, though improvements were most noticeable in simulated highly polygenic and heritable traits. When predicting lipid traits for African British in the UK biobank, slaPRS similarly performed well though was equitable compared to the global stacking approach. Similar performance across methods was likely driven by the genetic architecture of the lipid traits, as only a small proportion of windows genome wide across traits contributed meaningfully to their heritability.

The development of slaPRS has immediate broader implications on the field of genetic epidemiology and polygenic risk scores in admixed individuals. Methods for polygenic risk prediction in admixed individuals are lacking and traditional single population can be ancestry dependent. Recently proposed methods for admixed PRS differ in approach, but all share incorporation of ancestry specific GWAS and local ancestry at a given variant[107,157]. Our results

confirm that use of multiple ancestry GWAS is effective in improving PRS performance while also reducing ancestry dependence of constructed PRS in admixed samples. However, an important finding from our study is the degree of PRS improvement from explicitly modeling local ancestry in admixed genomes may ultimately depend on the trait's transethnic genetic architecture. Specifically, for the slaPRS framework we found explicit consideration of local ancestry may be less important when transethnic genetic correlation is high for a trait across ancestral populations. However, more research and method development are needed to validate such a claim. Notably, recent work using admixed genomes has actually suggested that most traits do have high transethnic genetic correlation[99]. Considering such findings, this raises the question of whether explicit PRS method development for admixed genomes should focus and be tailored towards traits with low transethnic genetic architecture.

Well-performing PRS across a range of admixture is important for equitable benefits of genomic research both across diverse populations and within an admixed population. As mentioned, admixed populations categorized by discrete groups are already historically understudied in genomic studies. However, further inequities exist even within a discrete admixture group for individuals who deviate from the group's mean ancestry composition. Recent work has acknowledged heterogeneity within single ancestral groups and pushed to consider all populations on an ancestry continuum rather than a discrete group[103]. Our proposed method slaPRS is an important step to consider admixture on a continuum, as our local approach considers individual genomes at a finer resolution to integrate multiple population GWAS and consider local ancestry for PRS construction. As PRS are increasingly incorporated into prediction models (with other health factors), precision medicine, and even clinical trial

requirement, it is imperative that an admixed PRS can benefit all individuals within a discrete admixture "group".

Lastly, increasing globalization and immigration of human populations has and will continue to result in gene flow between historically diverged groups. As a result, genetic admixture will likely become increasingly complex and expand upon currently defined populations. For example, the 1000 Genomes Project admixed American super populations currently include African Americans, African-Carribean, Mexican-American, Puerto-Rican, among others. However, admixture is rapidly increasing between other defined 1000 Genomes super populations (e.g. East Asian and Europeans, South Asian and African) outside of traditionally defined admixture groups. As genetic data for such individuals become increasingly available, methods such as slaPRS are needed to provide a flexible framework to incorporate increasingly complex scenarios of admixture and provide a tool for inclusion of such individuals in genomic research.

In the third project, we introduced a novel method LDSC-MIX to identify SNP sets and genes driving signals of local genetic correlation (LGC) through a Bayesian mixture of cross trait LD score regressions model. As the genetic basis for many complex traits and diseases have now been elucidated, such data has allowed for researchers to better study how the genetic basis is shared across two or more traits. LGC is an increasingly popular approach for both quantifying and identifying pleiotropic regions with shared genetic signals across two (or more) complex traits. In just the past year, LGC studies have been conducted on a number of trait pairs including: mental disorders[158], obesity and PCOS[159], and thyroid and psychiatric disorders[160] to identify regions harboring shared genetic signals. While useful to pinpoint regions with shared

genetics, current LGC current frameworks are limited in being unable to identify the specific SNP sets or genes driving the signal of LGC in such regions.

Current approaches such as pheWAS (variant is associated with both traits) and colocalization (Bayesian modeling across scenarios of shared association signals) can be used in LGC regions to identify shared genetic signals, though they typically rely on the existence of GWAS hits. Our approach differs in that we use the linear relationship of LD scores and product of z scores to identify shared signals, and thus does not explicitly require GWAS hits. As expected, in simulations we generally found that LDSC-MIX outperforms in scenarios of multiple weaker shared causal variants with respect to number of applicable instances a candidate set could be formed, identifying true shared causal variants, and recovering the true LGC. On the other hand, when there are clear single shared genetic signals colocalization was the preferred approach. In real data applications we applied LDSC-MIX to summary statistics from the UK Biobank for trait pairs asthma-basal cell carcinoma and asthma-HDL. We focus on the most significant LGC regions, where LDSC-MIX highlighted specific genes that are enriched for SNPs identified to belong in the genetically corelated group. While circumstantial evidence existence for potential cross-trait functionality for identified genes, further follow up studies are necessary to make definitive conclusions.

Our approach provides a method to identify gene-level pleiotropy through allowing for the presence of multiple weaker shared genetic effects, which has implications on how the field studies pleiotropic effects. While current genetic studies for many complex traits are ever increasing in sample sizes, rare disease or complex traits with low prevalence are still difficult to collect adequately powered sample sizes. In such cases, detection of pleiotropic signals remains difficult, as genetic signals are typically not statistically significant. LDSC-MIX and other

approaches are necessary to allow researchers to study such trait pairings with absences of strong shared genetic signals. The issue of underpowered studies is further exacerbated in understudied populations (non-European), as large consortiums and meta-analysis are similarly difficult to collect large enough studies. On the other hand, even for adequately powered studies LDSC-MIX may have immediate impacts. For a single phenotype the genetic architecture of a complex trait can range from fully mendelian (single gene affects trait) to the omnigenic model (core genes with peripheral genes having effects on core genes through interconnected regulatory networks). Under the omnigenic model[156], core genes which play a direct role on the trait typically have large effect sizes. Such genes are suggested to be common if not ubiquitous[161], and their large effects lend to easier studying of pleiotropic effects across traits using traditional approaches, with the occurrence. On the other hand, even in genetic studies with large sample sizes, peripheral genes under the omnigenic model are inherently pleiotropic from potentially all genes having an effect through various networks. However, the effect size of such peripheral genes effects may become undetectably small causing pleiotropic detection to be difficult[162]. Thus, methods such as our proposed LDSC-MIX are important to study and validate the existence of proposed weakly pleiotropic peripheral genes across networks.

The work in this dissertation has highlighted how the current wealth of genetic data has resulted in broad opportunities for studying population histories, predicting phenotypes using genetic data, and studying how genetics vary across diverse populations and complex traits. In particular, the latter two projects proposing methods for polygenic risk prediction in admixed individuals and interpreting local genetic correlation are especially active areas of research that are becoming increasingly possible as sample sizes of genetic studies routinely grow. Our novel

contributions in these areas are just a small step that we anticipate will be further actively developed by others in the field.

Bibliography

1. Mardis, E. R. DNA sequencing technologies: 2006-2016. *Nat. Protoc.* **12**, 213–218 (2017).
2. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).
3. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
4. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
5. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
6. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
7. Zawistowski, M. *et al.* The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. *Cell Genom.* **3**, 100257 (2023).
8. All of Us Research Program Investigators *et al.* The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
9. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genom.* **2**, 100192 (2022).
10. Craig, D. W. *et al.* Assessing and managing risk when sharing aggregate genetic variant data. *Nat. Rev. Genet.* **12**, 730–736 (2011).

11. Fisher, R. A. XVII.—the distribution of gene ratios for rare mutations. *Proc. R. Soc. Edinb.* **50**, 204–219 (1931).

12. Ewens, W. J. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112 (1972).

13. Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: From association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).

14. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

15. Fu, W., O'Connor, T. D. & Akey, J. M. Genetic architecture of quantitative traits and complex diseases. *Curr. Opin. Genet. Dev.* **23**, 678–683 (2013).

16. Vicente, C. T. *et al.* Long-range modulation of PAG1 expression by 8q21 allergy risk variants. *Am. J. Hum. Genet.* **97**, 329–336 (2015).

17. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics* vol. 9 e1003348 Preprint at https://doi.org/10.1371/journal.pgen.1003348 (2013).

18. Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat. Med.* **27**, 1876–1884 (2021).

19. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

20. Inouye, M. *et al.* Genomic risk prediction of coronary artery disease in 480,000 adults: Implications for primary prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).

21. Ganna, A. *et al.* Multilocus genetic risk scores for coronary heart disease prediction. *Arterioscler. Thromb. Vasc. Biol.* **33**, 2267–2272 (2013).

22. Oram, R. A. *et al.* A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults. *Diabetes Care* **39**, 337–344 (2016).
23. Udler, M. S., McCarthy, M. I., Florez, J. C. & Mahajan, A. Genetic risk scores for diabetes diagnosis and precision medicine. *Endocr. Rev.* **40**, 1500–1520 (2019).
24. Mars, N. *et al.* The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat. Commun.* **11**, 6383 (2020).
25. Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).
26. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
27. Stearns, F. W. One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767–773 (2010).
28. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).
29. Gratten, J. & Visscher, P. M. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med.* **8**, 78 (2016).
30. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
31. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
32. Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* **7**, 170125 (2017).

33. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).

34. Lareau, C. A. *et al.* Polygenic risk assessment reveals pleiotropy between sarcoidosis and inflammatory disorders in the context of genetic ancestry. *Genes Immun.* **18**, 88–94 (2017).

35. van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* **20**, 567–581 (2019).

36. Zhang, Y. *et al.* Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. *Brief. Bioinform.* **22**, (2021).

37. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C. A. An integrated framework for local genetic correlation analysis. *Nat. Genet.* **54**, 274–282 (2022).

38. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* **101**, 737–751 (2017).

39. Zhang, Y. *et al.* SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome Biol.* **22**, 262 (2021).

40. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

41. Wong, L.-P. *et al.* Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.* **92**, 52–66 (2013).

42. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).

43. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

44. Pool, J. E., Hellmann, I., Jensen, J. D. & Nielsen, R. Population genetic inference from genomic sequence variation. *Genome Res.* **20**, 291–300 (2010).

45. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).

46. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).

47. Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat. Commun.* **9**, 3753 (2018).

48. Aikens, R. C., Johnson, K. E. & Voight, B. F. Signals of variation in human mutation rate at multiple levels of sequence context. *Mol. Biol. Evol.* **36**, 955–965 (2019).

49. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).

50. Ronen, R., Udpa, N., Halperin, E. & Bafna, V. Learning natural selection from the site frequency spectrum. *Genetics* **195**, 181–193 (2013).

51. Kamm, J., Terhorst, J., Durbin, R. & Song, Y. S. Efficiently inferring the demographic history of many populations with allele count data. *J. Am. Stat. Assoc.* **115**, 1472–1487 (2020).

52. Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942 (2000).

53. Williamson, S. H. *et al.* Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7882–7887 (2005).

54. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).

55. Marth, G. T., Czabarka, E., Murvai, J. & Sherry, S. T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372 (2004).

56. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).

57. Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413 (2000).

58. Kamm, J. A., Terhorst, J. & Song, Y. S. Efficient computation of the joint sample frequency spectra for multiple populations. *J. Comput. Graph. Stat.* **26**, 182–194 (2017).

59. Zhang, W., Bouffard, G. G., Wallace, S. S., Bond, J. P. & NISC Comparative Sequencing Program. Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. *J. Mol. Evol.* **65**, 207–214 (2007).

60. Seplyarskiy, V. B. *et al.* Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science* **373**, 1030–1035 (2021).

61. Harpak, A., Bhaskar, A. & Pritchard, J. K. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. *PLoS Genet.* **12**, e1006489 (2016).

62. Wakeley, J., Fan, W. T. (louis), Koch, E. & Sunyaev, S. Recurrent mutation in the ancestry of a rare variant. *bioRxiv* (2022) doi:10.1101/2022.08.18.504427.

63. Jenkins, P. A. & Song, Y. S. The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theor. Popul. Biol.* **80**, 158–173 (2011).

64. Lachance, J. & Tishkoff, S. A. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *Am. J. Hum. Genet.* **95**, 408–420 (2014).

65. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10**, 285–311 (2009).

66. Kostka, D., Hubisz, M. J., Siepel, A. & Pollard, K. S. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol. Biol. Evol.* **29**, 1047–1057 (2012).

67. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

68. Korneliussen, T. S., Moltke, I., Albrechtsen, A. & Nielsen, R. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics* **14**, 289 (2013).

69. Fu, Y. X. Statistical properties of segregating sites. *Theor. Popul. Biol.* **48**, 172–197 (1995).

70. Durrett, R. *Probability models for DNA sequence evolution. Probability and its applications*. (Springer, 2008).

71. Excofffier, L. *et al.* Fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics* **37**, 4882–4885 (2021).

72. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).

73. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).

74. Ponting, C. P. & Lunter, G. Signatures of adaptive evolution within human non-coding sequence. *Hum. Mol. Genet.* **15 Spec No 2**, R170-5 (2006).

75. Asthana, S. *et al.* Widely distributed noncoding purifying selection in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 12410–12415 (2007).

76. Kang, H. M. EPACTS: Efficient and Parallelizable Association Container Toolbox. *http://genome.sph.umich.edu/wiki/EPACTS*.

77. O'Reilly, P. F., Birney, E. & Balding, D. J. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res.* **18**, 1304–1313 (2008).

78. Glémin, S. *et al.* Quantification of GC-biased gene conversion in the human genome. *Genome Res.* **25**, 1215–1228 (2015).

79. Smith, N. G. C., Webster, M. T. & Ellegren, H. Deterministic mutation rate variation in the human genome. *Genome Res.* **12**, 1350–1356 (2002).

80. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* **1**, 131 (2010).

81. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).

82. Reppell, M., Boehnke, M. & Zöllner, S. The impact of accelerating faster than exponential population growth on genetic variation. *Genetics* **196**, 819–828 (2014).

83. Gao, F. & Keinan, A. High burden of private mutations due to explosive human population growth and purifying selection. *BMC Genomics* **15 Suppl 4**, S3 (2014).

84. Gaboriaud, J. & Wu, P.-Y. J. Insights into the link between the organization of DNA replication and the mutational landscape. *Genes (Basel)* **10**, 252 (2019).

85. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).

86. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).

87. Kenigsberg, E. *et al.* The mutation spectrum in genomic late replication domains shapes mammalian GC content. *Nucleic Acids Res.* **44**, 4222–4232 (2016).

88. Pouyet, F., Aeschbacher, S., Thiéry, A. & Excoffier, L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *Elife* **7**, (2018).

89. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).

90. Barroso, G. V. & Dutheil, J. Y. Mutation rate variation shapes genome-wide diversity in*Drosophila melanogaster*. *bioRxiv* (2021) doi:10.1101/2021.09.16.460667.

91. Mugal, C. F. & Ellegren, H. Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. *Genome Biol.* **12**, R58 (2011).

92. Fryxell, K. J. & Moon, W.-J. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* **22**, 650–658 (2005).

93. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).

94. Lewis, A. C. F., Green, R. C. & Vassy, J. L. Polygenic risk scores in the clinic: Translating risk into action. *HGG Adv.* **2**, 100047 (2021).

95. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–428 (2018).

96. Plomin, R. & von Stumm, S. Polygenic scores: prediction versus explanation. *Mol. Psychiatry* **27**, 49–52 (2022).

97. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications* vol. 10 Preprint at https://doi.org/10.1038/s41467-019-11112-0 (2019).

98. Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across diverse populations. Preprint at https://doi.org/10.1101/070797.

99. Hou, K. *et al.* Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat. Genet.* **55**, 549–558 (2023).

100. *The impact of Linkage Disequilibrium on differences in predictive ability of polygenic risk score across populations*.

101. Miao, J. *et al.* Quantifying portable genetic effects and improving cross-ancestry genetic prediction with GWAS summary statistics. *Nat. Commun.* **14**, 832 (2023).

102. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).

103. Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., ... & Pasaniuc, B. Polygenic scoring accuracy varies across the genetic ancestry continuum in all human populations. *bioRxiv* (2022).

104. Bitarello, B. D. & Mathieson, I. Polygenic Scores for Height in Admixed Populations. *G3* **10**, 4027–4036 (2020).

105. Cavazos, T. B. & Witte, J. S. Inclusion of Variants Discovered from Diverse Populations Improves Polygenic Risk Score Transferability. Preprint at https://doi.org/10.1101/2020.05.21.108845.

106. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 1080 (2019).

107. Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat. Commun.* **11**, 1628 (2020).

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

109. Breiman, L. Stacked regressions. *Machine Learning* vol. 24 49–64 Preprint at https://doi.org/10.1007/bf00117832 (1996).

110. Džeroski, S. & Ženko, B. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning* vol. 54 255–273 Preprint at https://doi.org/10.1023/b:mach.0000015881.36452.6e (2004).

111. Privé, F., Vilhjálmsson, B. J., Aschard, H. & Blum, M. G. B. Making the Most of Clumping and Thresholding for Polygenic Scores. *Am. J. Hum. Genet.* **105**, 1213–1221 (2019).

112. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

113. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).

114. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55 (1970).

115. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* **58**, 267–288 (1996).

116. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology* vol. 12 e1004842 Preprint at https://doi.org/10.1371/journal.pcbi.1004842 (2016).

117. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).

118. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).

119. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).

120. Zhao, Z., Fritsche, L. G., Smith, J. A., Mukherjee, B. & Lee, S. The construction of cross-population polygenic risk scores using transfer learning. *Am. J. Hum. Genet.* **109**, 1998–2008 (2022).

121. Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).

122. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

123. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).

124. Maxmen, A. The next chapter for African genomics. *Nature* **578**, 350–354 (2020).

125. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).

126. Majara, L. *et al.* Low and differential polygenic score generalizability among African populations due largely to genetic diversity. *HGG Adv.* **4**, 100184 (2023).

127. Fatumo, S. *et al.* Promoting the genomic revolution in Africa through the Nigerian 100K Genome Project. *Nat. Genet.* **54**, 531–536 (2022).

128. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).

129. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110–124 (2018).

130. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).

131. Paul, D. A double-edged sword. *Nature* **405**, 515 (2000).

132. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

133. Pendergrass, S. A. *et al.* The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.* **35**, 410–422 (2011).

134. Foley, C. N. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* **12**, 764 (2021).

135. Thom, C. S. & Voight, B. F. Genetic colocalization atlas points to common regulatory sites and genes for hematopoietic traits and hematopoietic contributions to disease phenotypes. *BMC Med. Genomics* **13**, 89 (2020).

136. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* **17**, e1009440 (2021).

137. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

138. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).

139. The Brainstorm Consortium *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).

140. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).

141. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

142. Lu, Q. *et al.* A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *Am. J. Hum. Genet.* **101**, 939–964 (2017).

143. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

144. Stephens, M. Dealing with label switching in mixture models. *J. R. Stat. Soc. Series B Stat. Methodol.* **62**, 795–809 (2000).

145. Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-6**, 721–741 (1984).

146. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).

147. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

148. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).

149. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

150. Davis, K. U. & Sheats, M. K. Differential gene expression and Ingenuity Pathway Analysis of bronchoalveolar lavage cells from horses with mild/moderate neutrophilic or mastocytic inflammation on BAL cytology. *Vet. Immunol. Immunopathol.* **234**, 110195 (2021).

151. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

152. Sehovic, E., Hadrovic, A. & Dogan, S. Detection and analysis of stable and flexible genes towards a genome signature framework in cancer. *Bioinformation* **15**, 772–779 (2019).

153. Deng, L., Sharma, J., Ravera, E., Halmos, B. & Cheng, H. Hypersensitivity in ALK-positive lung cancers exposed to ALK inhibitors: a case of successful switch to an alternative ALK inhibitor and systematic review of the literature. *Lung Cancer (Auckl.)* **9**, 73–77 (2018).

154. Truong, L., Zheng, Y.-M., Song, T., Tang, Y. & Wang, Y.-X. Potential important roles and signaling mechanisms of YPEL4 in pulmonary diseases. *Clin. Transl. Med.* **7**, 16 (2018).

155. Yao, C. *et al.* Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation* **131**, 536–549 (2015).

156. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

157. Sun, Q. *et al.* Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-specific effects via GAUDI. *bioRxiv* (2022) doi:10.1101/2022.10.06.511219.

158. Hindley, G. *et al.* Charting the landscape of genetic overlap between mental disorders and related traits beyond genetic correlation. *Am. J. Psychiatry* **179**, 833–843 (2022).

159. Liu, Q. *et al.* Genomic correlation, shared loci, and causal relationship between obesity and polycystic ovary syndrome: a large-scale genome-wide cross-trait analysis. *BMC Med.* **20**, 66 (2022).

160. Soheili-Nezhad, S., Sprooten, E., Tendolkar, I. & Medici, M. Exploring the genetic link between thyroid dysfunction and common psychiatric disorders: A specific hormonal or a general autoimmune comorbidity. *Thyroid* **33**, 159–168 (2023).

161. Chesmore, K., Bartlett, J. & Williams, S. M. The ubiquity of pleiotropy in human disease. *Hum. Genet.* **137**, 39–44 (2018).

162. Brown, S. D. M. & Lad, H. V. The dark genome and pleiotropy: challenges for precision medicine. *Mamm. Genome* **30**, 212–216 (2019).