## Title

Leveraging genetic and electronic health record data to understand complex traits and rare diseases

## Permalink

https://escholarship.org/uc/item/8c84j3q1

## Author

Johnson, Ruth Dolly

## Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

# Leveraging genetic and electronic health record data to understand complex traits and rare diseases

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

by

Ruth Dolly Johnson

2023

# ABSTRACT OF THE DISSERTATION

Leveraging genetic and electronic health record data to understand complex traits and rare diseases

by

Ruth Dolly Johnson
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2023
Professor Sriram Sankararaman, Chair

The biobank era of genomics has ushered in a multitude of opportunities for precision medicine research. In particular, biobanks connected to electronic health records (EHR) provide rich phenotype information used to study to clinical phenome. First, I describe two computational methods designed to infer the genetic architecture of complex traits using biobank-scale data. Both methods are based on Markov Chain Monte Carlo techniques. Next, I provide an overview of the UCLA ATLAS Community Health Initiative (ATLAS), an EHR-linked biobank embedded within UCLA Health. Using this data set, I explore the role of genetic ancestry in common disease risk across the UCLA patient population. Next, I include a review of how race, ethnicity, and genetic ancestry are utilized in the field of EHR-linked biobanks. Finally, I propose an EHR-based algorithm, called PheNet, which identifies undiagnosed patients with Common Variable Immunodeficiency Disorders and demonstrate its application across a total of 5 University of California Health systems.

The dissertation of Ruth Dolly Johnson is approved.

Bogdan Pasaniuc

Harold Joseph Pimentel

Eleazar Eskin

Sriram Sankararaman, Committee Chair

University of California, Los Angeles

2023

*To the Ruths Club.*

# Table of Contents

# List of Figures

# Curriculum Vitae

| | |
|---|---|
| 2013 – 2017 | B.S. in Mathematics of Computation, University of California, Los Angeles |
| 2017 – Present | Ph.D. candidate in Computer Science, University of California, Los Angeles |
| Spring 2017 | Honorable Mention NSF Graduate Research Fellowships Program (GRFP), National Science Foundation |
| 2017-2021 | Eugene V. Cota-Robles Fellowship, UCLA Graduate Division. |
| Fall 2018 | Honorable Mention Ford Foundation Fellowship, The National Academies of Sciences, Engineering, and Medicine |
| 2018-2019 | NRT-Modeling and Understanding Human Behavior Fellowship, UCLA, National Science Foundation |
| Winter 2020 | Teaching assistant, UCLA Department of Computer Science, COMP SCI CM146 - Introduction to Machine Learning |
| Fall 2022 | Semi-finalist Charles J. Epstein Trainee Award for Excellence in Human Genetics Research, American Society of Human Genetics |

# Publications

Selected publications (of 16)

Johnson R, Shi H, Pasaniuc B, Sankararaman S. A unifying framework for joint trait analysis under a non-infinitesimal model. Bioinformatics. 2018 Jul 1;34(13):i195-201.

Johnson R, Burch KS, Hou K, Paciuc M, Pasaniuc B, Sankararaman S. A scalable method for estimating the regional polygenicity of complex traits. International Conference on Research

in Computational Molecular Biology 2020 May 10 (pp. 253-254). Springer, Cham.

Johnson R, Burch KS, Hou K, Paciuc M, Pasaniuc B, Sankararaman S. Estimation of regional polygenicity from GWAS provides insights into the genetic architecture of complex traits. PLoS computational biology. 2021 Oct 21;17(10):e1009483.

Johnson R, Ding Y, Venkateswaran V, Bhattacharya A, Boulier K, Chiu A, Knyazev S, Schwarz T, Freund M, Zhan L, Burch KS. Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. Genome medicine. 2022 Dec;14(1):1-23.ß

Johnson RD, Ding Y, Bhattacharya A, Chiu A, Lajonchere C, Geschwind DH, Pasaniuc B. The UCLA ATLAS Community Health Initiative: promoting precision health research in a diverse biobank. medRxiv. 2022 Jan 1.

Johnson RD, Stephens AV, Knyazev S, Kohn LA, Freund MK, Bondhus L, Hill BL, Schwarz T, Zaitlen N, Arboleda V, Butte MJ. Electronic health record signatures identify undiagnosed patients with Common Variable Immunodeficiency Disease. medRxiv. 2022 Jan 1.

# CHAPTER 1

# Introduction

The biobank era of genomics has ushered in a multitude of opportunities for precision medicine research [55, 34, 78, 326]. These large-scale genetic datasets comprising data from hundreds of thousands of individuals provide unprecedented sample sizes, enabling the statistical power needed to test hypotheses regarding the role of genetics in disease development. More recently, these biobanks have been linked with electronic health record (EHR) information, providing a vast catalog of phenotype information for a wide range of diseases. These EHR-linked biobanks provide unprecedented opportunities for studying the genetic basis for both rare and common diseases [9]. This catalog also allows for a variety of hypotheses to be tested without the need for costly recruitment efforts. Furthermore, access to longitudinal information enables the ability to test hypotheses about disease progression and future diagnoses.

However, there are numerous challenges associated with working with EHR-linked biobanks. First, performing statistical analyses with biobank-scale data has the potential of increased power, but also considerably more computational constraints. Thus designing computational methods that efficiently run at this scale is necessary to fully utilize the potential of this data. Second, numerous types of genetic and epidemiological studies, especially those analyzing common genetic variation, require genetic ancestry information. Disease risk is heavily intertwined with genetic ancestry, but the amount of interplay between these two factors has largely been characterized. In particular, genetic ancestry is also correlated with race and ethnicity, but are distinct concepts. A critical complexity of EHR-linked biobanks is investigating disease risk due to variation in genetic ancestry in a clinical landscape shaped by race and ethnicity. Next, a logical application of EHR is utilizing the information as predictive features in disease models. Leveraging this type of data could be especially useful when studying diseases that have a largely undiscovered genetic basis. However, the scale and heterogeneity of EHR make it difficult to derive clinically informative features for prediction algorithms.

To address these knowledge gaps, my thesis focuses on utilizing both genetic and EHR data to further understand complex traits and rare diseases. First, I describe novel scalable statistical models for studying the genetic architecture of complex traits and the application of these methods in the UK Biobank. Second, I provide a technical overview of the UCLA ATLAS Community Health initiative Biobank. Using this resource, I study the broad characterization of the role of genetic ancestry in common disease risk. I also provide a review of how race, ethnicity, and genetic ancestry are currently used in EHR-linked biobanks as well as the associated considerations and challenges. Next, I aim to develop models that identify phenotypic patterns within the EHR that can be used to predict specific diseases or so-called "EHR-signatures". Specifically, I develop a prediction algorithm for common variable immunodeficiency (CVID) in collaboration with the Department of Pediatrics.

The projects described above are organized into the following thesis chapters:

1. A unifying framework for joint trait analysis under a non-infinitesimal model

2. Estimation of regional polygenicity from GWAS provides insights into the genetic architecture of complex traits

3. The UCLA ATLAS Community Health Initiative: promoting precision health research in a diverse biobank

4. Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative

5. Race, ethnicity, and genetic ancestry in the age of EHR-linked biobanks

6. Electronic health record signatures identify undiagnosed patients with Common Variable Immunodeficiency Disease

# CHAPTER 2

# A unifying framework for joint trait analysis under a non-infinitesimal model

## 2.1 Introduction

Genome-wide association studies (GWAS) have identified thousands of regions in the genome that contain variants that contribute to the risk for many diseases. Many of these risk regions are known to be implicated in multiple phenotypes such as autism and schizophrenia [24], multiple autoimmune diseases [77, 242, 246], Crohn's disease and psoriasis[92], and many others. Understanding which causal variants are shared among diseases can provide novel etiological insight as well as provide evidence of potential shared causal mechanisms between complex traits. In addition, identifying which variants contribute to multiple traits can help decipher which molecular traits (*e.g.*, gene expression) contribute to disease risk [108, 124]; genetic variants that causally alter gene expression, as well as disease risk, can link a particular gene to a given disease.

Genetic overlap has been analyzed both at the genome-wide level and local level, where the latter refers to an analysis done within a given genomic region. *Genetic correlation*, a measure that quantifies the similarity in the genetic effects on pairs of traits, is commonly used for assessing the relationship between two traits and can be applied either genome-wide or to local data [51, 268]. Many of the models for estimating genome-wide genetic correlation assume an *infinitesimal* genetic architecture where all SNPs are assumed to have a very small effect on the trait. In contrast to genetic correlation, *colocalization* methods aim to estimate whether the GWAS association signals for two traits at the same region are due to causal variant(s) shared across the traits or chance[108, 124]. The methods that relax the infinitesimal assumption either assume a single causal variant per region or limit the number of potential causal variants a priori, often due to computational considerations [108, 124] . Although both genetic correlation and colocalization aim to describe the genetic sharing between traits, these methods have been utilized independently of each other.

In this work, we present a unifying statistical model that ties together genetic correlation and colocalization. To accomplish this, we present a fully generative Bayesian statistical model that models the shared as well as distinct genetic variants underlying a pair of traits. The model allows for sparse genetic architectures (where only a small fraction of variants are causally impacting the traits). The model is richly parameterized: allowing us to jointly model global parameters such as the proportion of variants that are causal for both as well for either trait, the trait heritabilities, the correlation of the effect sizes at the causal SNPs, as well as local parameters such as whether the effect of a single SNP on each of the traits.

A challenge of a non-infinitesimal genetic architecture is that it presents a computationally challenging inference problem. Performing inference under this model often involves explicitly enumerating all causal configurations of the SNPs. This exponential search space of $2^{2M}$, where $M$ is the number of SNPs analyzed, proves intractable given the large genetic data sets now available. We propose our method, Unifying Non-Infinitesimal Trait analYsis (UNITY), that relies on Markov Chain Monte Carlo (MCMC) to approximate the posterior probabilities of the model parameters. In this work, we focus on estimating the proportion of shared and trait-specific causal variants since parameters such as heritability and genetic correlation can be estimated using previous methods [51]. Additionally, a key advantage of the method is that it only requires summary level association statistic data, which bypasses many of the privacy concerns associated with individual level data. With the widespread availability of GWAS summary statistics [228], we expect that a method operating only on summary statistics would prove most useful for the research community. Through comprehensive simulations and an analysis of height and BMI, we show that our method can accurately estimate the proportion of shared causal SNPs.

## 2.2 Methods

### 2.2.1 Generative Model

Here we introduce a Bayesian framework for estimating the proportion of shared causal variants between two complex traits. The input of our method is a vector of signed effect sizes for each SNP from each trait. We model the genetic variances from both traits, genetic correlation, non-genetic variance for both traits, and the proportion of causal SNPs. The estimated proportion of causal SNPs shared between the traits is denoted by $p_{11}$, the proportion of causal SNPs specific to trait 1 and trait 2 as $p_{10}$ and $p_{01}$, and the proportion of non-causal SNPs is denoted by $p_{00}$. We denote the heritability as $h_1^2 = \sigma_1^2, h_2^2 = \sigma_2^2$ and environmental noise as $\sigma_{e_1}^2 = \frac{1-h_2^2}{N_1}$, $\sigma_{e_2}^2 = \frac{1-h_1^2}{N_2}$, where $N_1$ and $N_2$ denote the sample sizes for trait 1 and trait 2. Additionally, the number of individuals shared between studies is denoted by $N_s$. Altogether, our model has the following parameters: $\{\sigma_1^2, \sigma_2^2, \rho, \sigma_{e_1}^2, \sigma_{e_2}^2, p_{00}, p_{10}, p_{01}, p_{11}\}$.

We assume that each phenotype, $y_1$ and $y_2$ is a linear function of standardized genotype matrices $X_1$ and $X_2$ with $M$ number of SNPs, SNP effect sizes $\beta_1$ and $\beta_2$, and a noise term denoted by $\epsilon_1, \epsilon_2$, where the noise terms follows a Gaussian distribution:

$$y_{1,i} = \sum_{m=1}^{M} \beta_{1,m} x_{1,m} + \epsilon_1 \qquad y_{2,i} = \sum_{m=1}^{M} \beta_{2,m} x_{2,m} + \epsilon_2$$

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{e_1}^2 & \text{cov}(\epsilon_1, \epsilon_2) \\ \text{cov}(\epsilon_1, \epsilon_2) & \sigma_{e_2}^2 \end{pmatrix}\right)$$

We let the probability of a SNP being causal for every combination of the two traits be $\vec{p} = (p_{00}, p_{10}, p_{01}, p_{11})$. We assume that $\vec{p}$ has a Dirichlet prior, where in practice we set $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.20$:

$$\vec{p} \;\sim\; \text{Dirch}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$$

The effect sizes for each SNP are assumed to be independent, allowing us to model every causal effect size for each trait through a bivariate normal distribution centered at zero with the following covariance matrix:

$$\begin{pmatrix}\gamma_{1,m}\\ \gamma_{2,m}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\frac{\sigma_1^2}{M(p_{11}+p_{10})} & \frac{\sigma_1\sigma_2\rho}{M(p_{11})}\\ \frac{\sigma_1\sigma_2\rho}{M(p_{11})} & \frac{\sigma_2^2}{M(p_{11}+p_{01})}\end{pmatrix}\right)$$

Next, let $C_p$ be a causal indicator vector for trait $p$, where $C_{p,m} = 1$ if SNP $m$ is causal for trait $p$ and 0 otherwise. The true effect sizes for each trait, $\beta_p$, conditioned on a SNP's causal status is the element-wise product of the causal indicator vector and the true causal effect sizes.

$$\begin{pmatrix}\beta_{1,m}\\ \beta_{2,m}\end{pmatrix} \mid \begin{pmatrix}C_{1,m}\\ C_{2,m}\end{pmatrix}, \begin{pmatrix}\gamma_{1,m}\\ \gamma_{2,m}\end{pmatrix} = \begin{pmatrix}\gamma_{1,m} \circ C_{1,m}\\ \gamma_{2,m} \circ C_{2,m}\end{pmatrix}$$

We can model the conditional distribution of the GWAS summary statistics, where $\hat{\beta}_{p,m}$ is the estimated effect size of the $m^{th}$ SNP for a trait [268]:

$$\begin{pmatrix}\hat{\beta}_{1,1:M}\\ \hat{\beta}_{2,1:M}\end{pmatrix} \mid \begin{pmatrix}\beta_{1,1:M}\\ \beta_{2,1:M}\end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix}V\beta_{1,1:M}\\ V\beta_{2,1:M}\end{pmatrix}, \Sigma_e\right)$$

$$\Sigma_e = \begin{pmatrix}\sigma_{e_1}^2 V & \frac{N_s \text{cov}(\epsilon_1,\epsilon_2)}{N_1 N_2} V\\ \frac{N_s \text{cov}(\epsilon_1,\epsilon_2)}{N_1 N_2} V & \sigma_{e_2}^2 N_2 V\end{pmatrix}$$

We denote $V$ as the linkage disequilibrium matrix, which in practice could be estimated from a reference panel. However, when performing inference at the genome-wide level, we can prune the list of SNPs such that they come from independent LD blocks. LD-pruning creates an approximately independent subset of SNPs, which reduces $V$ to the identity matrix.

### 2.2.2 Parameter inference

The true joint posterior distribution is intractable, thus we use Markov chain Monte Carlo to sample from the posterior distribution. We derive a collapsed Gibbs sampling scheme as follows:

$$\begin{aligned}\vec{p}^{(t+1)} &\sim P(\vec{p}\,|\begin{pmatrix}\hat{\beta}_{1,1:M}\\ \hat{\beta}_{2,1:M}\end{pmatrix},\sigma_1^2,\sigma_2^2,\rho,\sigma_{e_1}^2,\sigma_{e_2}^2)\\ &\propto P(\vec{p},\begin{pmatrix}\hat{\beta}_{1,1:M}\\ \hat{\beta}_{2,1:M}\end{pmatrix},\sigma_1^2,\sigma_2^2,\rho,\sigma_{e_1}^2,\sigma_{e_2}^2)\end{aligned}$$

The conditional distribution does not have a closed form. Although it is difficult to sample from, it is simple to compute. To account for this, we derive a Metropolis-Hastings step within our collapsed Gibbs sampler, where the posterior can be written as:

$$\begin{aligned}&P(\begin{pmatrix}\hat{\beta}_{1,1:M}\\ \hat{\beta}_{2,1:M}\end{pmatrix},\sigma_1^2,\sigma_2^2,\rho,\sigma_{e_1}^2,\sigma_{e_2}^2,\vec{p}) \propto \Bigg[\prod_{m=1}^{M}\mathcal{MVN}\left(\begin{pmatrix}\hat{\beta}_{1,m}\\ \hat{\beta}_{2,m}\end{pmatrix};\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma_{e_1}^2 & 0\\ 0 & \sigma_{e_2}^2\end{pmatrix}\right)\cdot(p_{00})\\ &+\mathcal{MVN}\left(\begin{pmatrix}\hat{\beta}_{1,m}\\ \hat{\beta}_{2,m}\end{pmatrix};\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\frac{\sigma_1^2}{M(p_{11}+p_{10})}+\sigma_{e_1}^2 & 0\\ 0 & \sigma_{e_2}^2\end{pmatrix}\right)\cdot(p_{10})\\ &+\mathcal{MVN}\left(\begin{pmatrix}\hat{\beta}_{1,m}\\ \hat{\beta}_{2,m}\end{pmatrix};\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma_{e_1}^2 & 0\\ 0 & \frac{\sigma_2^2}{M(p_{11}+p_{01})}+\sigma_{e_2}^2\end{pmatrix}\right)\cdot(p_{01})\\ &+\mathcal{MVN}\left(\begin{pmatrix}\hat{\beta}_{1,m}\\ \hat{\beta}_{2,m}\end{pmatrix};\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\frac{\sigma_1^2}{M(p_{11}+p_{10})}+\sigma_{e_1}^2 & \frac{\sigma_1\sigma_2}{M(p_{11})}\rho\\ \frac{\sigma_1\sigma_2}{M(p_{11})}\rho & \frac{\sigma_2^2}{M(p_{11}+p_{01})}+\sigma_{e_2}^2\end{pmatrix}\right)\cdot(p_{11})\Bigg]\\ &\times \mathrm{Dirch}(\vec{p};\vec{\lambda})\end{aligned}$$

To sample from $P(\vec{p}|\begin{pmatrix}\hat{\beta}_{1,1:M}\\ \hat{\beta}_{2,1:M}\end{pmatrix},\sigma_1^2,\sigma_2^2,\rho,\sigma_{e_1}^2,\sigma_{e_2}^2)$, we use a random-walk Metropolis-Hastings scheme with the following proposal distribution, where $p^*$ denotes the value from the previous iteration and $B$ is a constant that controls the variance of the proposal distribution. In practice, we found that $B = 10$ yields effective mixing.

$$\vec{p} \sim \text{Dirch}(d_1, d_2, d_3, d_4)$$

$$d_1 = \lambda_1 + (B)(p^*_{00})$$

$$d_2 = \lambda_2 + (B)(p^*_{10})$$

$$d_3 = \lambda_3 + (B)(p^*_{01})$$

$$d_4 = \lambda_2 + (B)(p^*_{11})$$

### 2.2.3 Efficient mixing of MCMC chains

In any practical application of MCMC, the number of iterations, burn-in period, and initialization point are critical to ensuring convergence and accurate estimates. Slow mixing of the MCMC chains can occur if the starting point is at a region of low posterior density. As opposed to selecting a random starting point, we carefully select the initialization of each chain by choosing the set of parameters that yields the highest posterior density. We use the Limited-memory BFGS algorithm to determine the maximum a posteriori estimates for $p_{00}, p_{10}, p_{01}, p_{11}$. We repeat this 10 times, initializing the optimization algorithm with random starting values drawn from the prior. We compute the posterior density of all 10 candidate starting values and select the set that yields the highest density. This set of parameters is then used as the starting point for our MCMC chain. In addition, to diagnose convergence, we use 100 Markov chains all initialized using the scheme described above. Our final estimate is the mean of all samples drawn from the 100 chains.

### 2.2.4 Note on runtime

We assessed the performance based on the number of seconds per iteration of the MCMC sampler. The main computation is calculating the likelihood at each iteration, which is directly dependent on the number of SNPs per trait. The complexity of the algorithm is $\mathcal{O}(m)$, where $m$ is the number of SNPs. We empirically demonstrate that our method is linear in the number of SNPs through simulation (Figure 2.1). In addition, the runtime is invariably connected to the number of iterations required for the MCMC to converge. We find that using the maximum a posteriori probability (MAP) estimate as an initialization value leads to fast convergence, requiring only 500 iterations in practice.

## 2.3 Results

### 2.3.1 UNITY generalizes colocalization and genetic correlation

UNITY provides a novel generalized framework to jointly model GWAS summary statistics data of two complex traits, incorporating fundamental genetic parameters, such as heritability and genetic correlation, and makes minimal assumptions in inference procedures. Since UNITY assumes a non-infinitesimal model, it allows for very sparse genetic architectures, i.e. by setting $p_{00} \approx 1$. However, this non-infinitesimal model can also be generalized to the infinitesimal model by setting $p_{00} \approx 0, p_{10} \approx 0, p_{01} \approx 0, p_{11} \approx 1$.

We discuss a comparison of the parameters of UNITY with those obtained by other methods that perform cross-trait analysis and the underlying assumptions of each method. We first analyze the cross-trait LD score regression model [51], which estimates genome-wide genetic correlation based on the random-effect model, making the implicit assumption that every SNP has a non-zero effect. In contrast to cross-trait LD score regression, UNITY assumes a generalized non-infinitesimal model, explicitly modeling a sparse genetic architecture. We also compare UNITY with methods that do not make the infinitesimal model assumption. While models such as PleioPred explicitly model the proportion of trait-specific and shared causal variants $p_{00}, p_{10}, p_{10}, p_{11}$, the main goal of this method is to perform genetic risk prediction [128] rather than estimating these proportions.

We compare UNITY with COLOC [108] and eCAVIAR [124], Bayesian methods to assess the evidence of colocalization, i.e. whether GWAS signals of two traits driven the same underlying causal variants. Both methods explicitly model $\vec{p} = (p_{00}, p_{10}, p_{10}, p_{11})$ [108, 124]. However, COLOC makes the simplifying assumption that there is at most one causal variant at a region [108], allowing it to not explicitly model LD. And although eCAVIAR allows for multiple causal variants and explicitly models LD, it caps the maximum number of causal variants at 6 per region for computational efficiency [124]. In comparison with these methods, UNITY allows for any number of causal variants while making the assumption that there is no LD between the SNPs. We outline a summary of the relationship between UNITY and all methods described in Table 2.1.

To empirically demonstrate the benefit of the relaxed assumptions of UNITY as compared to current methods, we conduct a modest comparison against COLOC [108]. We simulated 100 regions of 500 SNPs with multiple causal variants. We perform colocalization analysis over all of the regions using COLOC. When there are causal variants independently associated with each trait and shared variants, COLOC estimates that the association within the region is driven only by two independent variants, where one is specific to trait 1 and the other is specific to trait 2. Because COLOC assumes at most one causal variant per region, the method is unable to distinguish between a variant that independently drives only one trait versus a variant that is colocalized when both cases are present. For completeness, we also included a simulation that follows COLOC's assumption of the one-causal setting. The full table listing these results in outlined in Table 2.2. However, we are unable to directly compare estimates with COLOC because there is not a clear mapping between the estimates of COLOC and the estimated parameters of UNITY, thus any direct comparison would be unfair comparison due to the mismatch in the models.

### 2.3.2 Simulations

We generated summary statistics for 500 SNPs from two synthetic GWAS. First, we simulated standardized genotype matrices for two traits. The two corresponding phenotypes were simulated under a linear model such that for each phenotype $p$, the $i^{th}$ individual's phenotype follows $y_{p,i} = \sum_{m=1}^{M} \beta_{p,m} x_m + \epsilon_p$. The causal effect sizes for each SNP, $\gamma_{p,m}$, were drawn jointly from a multivariate normal distribution where $h_1^2, h_2^2, \rho$ denote the heritability of each trait and the genetic correlation. We denote the number of SNPs as $M$ and the proportion of causal variants for each trait as $p_{10}, p_{01}$ and the proportion of shared casuals as $p_{11}$:

$$\begin{pmatrix}\gamma_{1,m} \\ \gamma_{2,m}\end{pmatrix} \sim \begin{pmatrix}\frac{\sigma_1^2}{M(p_{11}+p_{10})} & \frac{\sigma_1^2\sigma_2^2\rho}{M(p_{11})} \\ \frac{\sigma_1^2\sigma_2^2\rho}{M(p_{11})} & \frac{\sigma_2^2}{M(p_{11}+p_{01})}\end{pmatrix}$$

To simulate causal SNPs, we drew an $M \times 4$ matrix from a multinomial distribution parameterized by $\vec{p}$ where the $m^{th}$ row of values denotes whether a SNP is causal for neither trait, only trait 1, only trait 2, or neither trait. Using this, we constructed two $M \times 1$ causal indicator vectors $C_1, C_2$, where $C_{1,m}, C_{2,m} = 1$ if the $m^{th}$ SNP was causal for both traits, $C_{1,m} = 1, C_{2,m} = 0$ if the SNP was only causal for trait 1, $C_{1,m} = 0, C_{2,m} = 1$ if it was only causal for trait 2, and $C_{1,m}, C_{2,m} = 0$ if the SNP was non-causal. To get the true effect sizes, we multiplied element-wise $\beta_1 = C_1 \circ \gamma_1$ and $\beta_2 = C_2 \circ \gamma_2$ where we are essentially zeroing out any entry from the causal effect vector where a SNP is non-causal.

To compute the estimated GWAS effect sizes, $\hat{\beta}_p$, we assumed $\text{cov}(\epsilon_1, \epsilon_2) = 0$, so random noise terms $\epsilon_1, \epsilon_2$ were drawn from two normal distributions $\mathcal{N}(0, \frac{1-h_1^2}{N_1})$ and $\mathcal{N}(0, \frac{1-h_2^2}{N_2})$ respectively. We assume that the SNPs being used at the genome-wide level will be LD-pruned such that there is very little or no correlation structure. Thus, we set the LD matrix $V = I_M$, where $I_M$ is an $M \times M$ identify matrix and draw the estimated effect sizes from a conditional distribution of the GWAS summary statistics, as described in Methods.

First, we confirm that our method accurately predicts the proportion of causal variants under varying sample sizes and heritability estimates. We tested a variety of simulation frameworks where we fixed the genetic correlation and heritabilities of the two traits and ran each simulation for 500 iterations and used the first quarter of the iterations as burn-in. We vary the proportion of causal variants contributing to only trait 1 ($p_{10}$), the proportion of causal variants for only trait 2 ($p_{01}$), and the proportion of casuals contributing to both traits ($p_{11}$). As shown in Figure 2.2, we can see that UNITY performs robustly across each scenario.

Next, to assess how UNITY performs with varying levels of heritability, we continued to fix $\rho = 0$, but varied the values of the heritability. Note that we used low heritability values due to the low number of simulated SNPs (M=500). From Figure 2.3, we can see that the estimates reflect the prior distribution of $(p_{00}, p_{10}, p_{01}, p_{11})$ when the heritability is very low. We also show in Figure 2.4 that our estimates are invariant to the correlation between phenotypes.

To assess the role of sample size in our inference, we performed simulations where we varied the number of individuals from 1,000 to 250,000. We find that the recommended sample size should be at least greater than 50,000 individuals to yield precise results (Figure 2.5). Additionally, to further assess the performance of the method, we also performed simulations where $h_1^2 \neq h_2^2$ and when $p_{10} \neq p_{01}$. Through simulation, we demonstrate that our method is robust to these scenarios, with detailed results provided in Figure 2.6 and Figure 2.7.

Finally, through simulations, we empirically demonstrate that our method is well calibrated under the null hypothesis, defined as (1) $p_{10} = 0$, (2) $p_{01} = 0$, and (3) $p_{11} = 0$. To demonstrate this, we simulated 100,000 SNPs with 100,000 individuals where $h_1^2 = 0.25, h_2^2 = 0.25, \rho = 0$. For each hypothesis, we set the parameter of interest exactly to 0 and then simulated 2% causal variants between the remaining parameters. For example, for the null hypothesis (1), the corresponding set of parameters would be $p_{10} = 0, p_{01} = 0.01, p_{11} = 0.01$. Using UNITY, we estimated the null parameter and report the posterior mean and standard deviation below. Note that UNITY estimates the null parameter very close to zero, but not exactly zero. This is because there is a nonzero prior on the set of parameters, making it not possible to be exactly zero, but can instead be asymptotically close (Figure 2.3).

### 2.3.3 LD pruning to identify approximately independent SNPs

To rigorously assess the role of LD in our model, we demonstrate a sufficient LD-pruning scheme through simulations. To model a realistic LD structure, we used SNPs from 1000 Genomes [73] to compute the LD for each of the approximately independent LD blocks identified in Berisa et al [37]. We filtered rare SNPs with MAF $\leq 0.05$ and used 1 million SNPs sampled across the LD blocks. We chose only a subset of 1 million SNPs because this closely reflects the number of SNPs genotyped on SNP arrays. We simulated the GWAS effect sizes as outlined in Section 3.1, where the heritabilities for each trait were set to $h_1^2 = 0.50$ and $h_2^2 = 0.50$ (which is similar to the estimated SNP heritability for height), genetic correlation $\rho = 0$.

To assess the role of LD-pruning, we divided the genome into K kilobase non-overlapping windows and selected a SNP from each window. We varied K to assess the minimal window size necessary to create a subset of approximately independent SNPs. In addition, We used cross-trait LD Score regression to quantify the heritabilities for both traits and the genetic correlation after pruning, which were subsequently used in the inference. Through simulations, we determined that a 5KB window provides precise estimates (Table 2.4).

### 2.3.4 Empirical analysis of BMI and Height

We downloaded GWAS summary data for both Height and BMI from the GIANT consortium [277, 14] where each study has $> 170,000$ individuals. First, we overlapped each GWAS by rsid to get SNPs present in both studies. Then for each trait, we filtered out SNPs with a minor allele frequency $\leq 0.05$. Additionally, we LD pruned by taking a SNP from every 5KB window.

We used cross-trait LD Score to estimate the heritability and genetic correlation parameters: $h_H^2 = 0.2390, h_B^2 = 0.1566, \rho = -0.0845$. Denoting Height as the first trait and BMI as the second, we estimated the proportion of causal variants for each trait as, $p_{00} = 0.9519, p_{10} = 0.0062, p_{01} = 0.01579, p_{11} = 0.0262$. We summarize the distribution of estimated causal SNPs in Figure 2.8.

Our results are consistent with the calculations of BMI and Height. Since BMI is a function of an individual's height and weight, we expect all of the contributing variants for Height to also contribute to BMI. UNITY predicts more BMI-only specific variants than Height-only variants. We hypothesize that the BMI-specific variants are those that contribute to weight, whereas the variants that contribute to height in the BMI data set were already captured in the $p_{11}$ estimate. In principle, we'd expect $p_{10}$ to be zero since SNPs contributing to height also contribute to BMI. We expect this could be due to the nonzero prior on $p_{10}$.

Because of this, the estimate can never truly be zero but can be asymptotically close.

## 2.4 Discussion

In this work, we introduce a statistical framework for quantifying the relationship between two complex traits. The key advantage of our method is that it makes very few assumptions about the data and few restrictions during inference. Rather than relying on assumptions about a trait's genetic architecture, we let the data describe the underlying genetics. By using a Metropolis-Hastings sampling framework, we can calculate a variety of likelihoods without breaking any conjugate prior pairings. For example, although we choose to model the causal effect sizes through a multivariate normal, one could choose another distribution, and the sampling procedure would still hold even if the new distribution did not have a conjugate prior. Additionally, we hypothesize that the collapsed Gibbs sampling would yield faster convergence than a traditional Gibbs sampling since many of the parameters are highly correlated. Finally, by operating exclusively on GWAS summary statistic data, we aim to encourage future large-scale meta-analyses, since obtaining individual level is not always readily available.

We conclude with several limitations and potential future directions of our framework. First, as the size of genetic datasets grows, sub-sampling methods such as MCMC may prove computationally intractable. Alternatives include using adaptive MCMC to accelerate mixing and convergence or variational methods that do not require sub-sampling. Additionally, we have yet to rigorously quantify the effects of LD in our model in practice for local inference. Although decorrelating the SNPs by multiplying by the inverse of the LD matrix, we note that many times the LD matrix is not invertible and thus we will have to assess how approximations to the inverse LD matrix perform in our model. We leave rigorous comparison between UNITY and other relevant methods and applying UNITY on a large number of traits to find new biological insights, for future work.

Additionally, recent integrative methods have shown that the incorporation of a variants functional genomic context can improve both power and accuracy in identifying potential causal variants [230, 145, 170, 129]. Large-scale initiatives such as the ENCODE [75] and ROADMAP [251] projects have provided comprehensive databases of tissue-specific functional genomic annotations. Combining this rich atlas of functional data and the genetic information from GWAS will likely uncover novel insights into disease biology. We leave the incorporation of functional elements as a potential direction for future work.

## 2.5 Tables

| Method | $h^2$ | $\rho$ | $\vec{p}$ | misc. |
|---|---|---|---|---|
| UNITY | * | * | * | |
| Cross-trait LD Score regression [51] | * | * | $p_{11} \approx 1$ | |
| PleioPred [128] | * | * | * | infers $\vec{p}$ to estimate effect sizes |
| COLOC [108] | – | – | * | max 1 causal |
| eCAVIAR [124] | – | – | * | max 6 causals |

Table 2.1: Displayed is a summary of current methods that perform joint trait analysis and the relationship to the parameters in UNITY. Boxes with an (*) denote the values that a method models. Note that this summary is not exhaustive

| | Parameters | H0 | H1 | H2 | H3 | H4 |
|---|---|---|---|---|---|---|
| one causal | $p_{10} = p_{01} = 0, p_{11} = \frac{1}{M}$ | 14.29% | 17.84% | 16.55% | 0.13% | 51.19% |
| multiple causals | $p_{10} = p_{01} = p_{11} = 0.01$ | 4.76% | 13.71% | 9.10% | 63.27% | 9.17% |

Table 2.2: To empirically demonstrate the benefit of the relaxed assumptions of UNITY as compared to current methods, we conduct a modest comparison against COLOC. We simulated 100 regions of M=500 SNPs under two simulation frameworks with the proportion parameters outlined in the second column and $h_1^2 = 0.00125, h_2^2 = 0.00125, \rho = 0, N_1 = 100,000, N_2 = 100,000$. COLOC calculates the posterior probability of a region corresponding to one of the 5 hypothesis - H0: no associated with either trait, H1: association with only trait 1, H2: association with only trait 2, H3: association with both traits driven by two independent SNPs, and H4: association with both trait 1 and trait 2 driven by one shared SNP (i.e. colocalized). We report the average posterior probability calculated over the 100 regions for each of the hypotheses.

| Hypothesis | Null parameter | Mean | SD |
|---|---|---|---|
| 1 | $p_{10}$ | 0.0006 | 0.0023 |
| 2 | $p_{01}$ | 0.0004 | 0.0005 |
| 3 | $p_{11}$ | 0.0002 | 0.0003 |

Table 2.3: We present the posterior means and standard deviations estimated when the proportion of causal variants is set exactly to zero for trait 1 and trait 2, and when the shared proportion is exactly zero.

| Pruning window (K) | | $p_{00}(0.99)$ | $p_{10}(0.0025)$ | $p_{01}(0.0025)$ | $p_{11}(0.0050)$ |
|---|---|---|---|---|---|
| no pruning | Mean | 0.986472 | 2.591e-06 | 0.01352 | 3.165e-06 |
| | SD | 0.09199 | 7.685e-06 | 0.09199 | 1.28e-05 |
| 1KB | Mean | 0.976628 | 2.515e-06 | 0.02337 | 3.168e-06 |
| | SD | 0.1369 | 7.53e-06 | 0.1369 | 1.28e-05 |
| 5KB | Mean | 0.999993 | 2.432e-06 | 2.582e-06 | 2.43e-06 |
| | SD | 9.391e-06 | 3.854e-06 | 4.091e-06 | 3.898e-06 |
| 10KB | Mean | 0.999992 | 2.56e-06 | 2.631e-06 | 2.763e-06 |
| | SD | 9.811e-06 | 3.915e-06 | 3.978e-06 | 4.043e-06 |
| 20KB | Mean | 0.999991 | 2.985e-06 | 3.195e-06 | 3.259e-06 |
| | SD | 1.088e-05 | 4.091e-06 | 4.6e-06 | 4.249e-06 |
| 30KB | Mean | 0.999985 | 5.208e-06 | 5.18e-06 | 5.152e-06 |
| | SD | 1.307e-05 | 5.205e-06 | 5.689e-06 | 5.061e-06 |
| 40KB | Mean | 0.999982 | 6.177e-06 | 6.16e-06 | 6.282e-06 |
| | SD | 1.335e-05 | 5.474e-06 | 5.923e-06 | 5.644e-06 |
| 50KB | Mean | 0.999979 | 6.908e-06 | 6.961e-06 | 6.933e-06 |
| | SD | 1.327e-05 | 5.287e-06 | 6.013e-06 | 6.157e-06 |

Table 2.4: To model a realistic LD structure, we used SNPs from 1000 Genomes to compute the LD for approximately 2,000 independent LD blocks. We simulated GWAS effect sizes as outlined in section 3.1 where the heritabilities for each trait was set to $h_1^2 = 0.50$ and $h_2^2 = 0.50$, genetic correlation $\rho = 0$. We varied the non-overlapping window length, K, to assess the minimal window size necessary to create a subset of approximately independent SNPs. Our results demonstrate that using a 5KB window gives more precise estimates while retaining the highest number of SNPs.

## 2.6 Figures

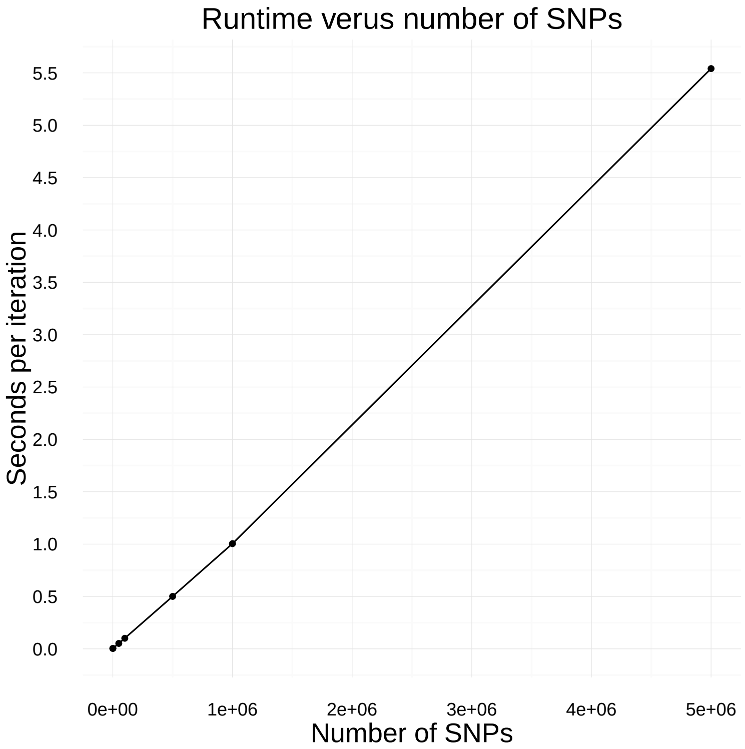

Figure 2.1: The complexity of our algorithm is $\mathcal{O}(M)$, where $M$ is the number of SNPs for each trait. We varied the total number of SNPs from 100 to 5,000,000 and then performed MCMC for 100 iterations and recorded the total amount of time necessary for sampling. This total time divided by the number of iterations is reported on the y-axis.

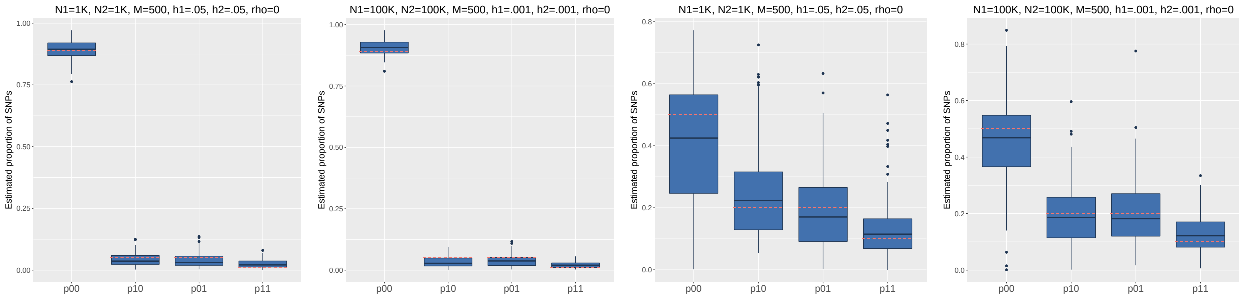

Figure 2.2: We estimate the proportion of causal variants under four simulation frameworks where we vary the sample size [N], heritability [$h_1 = h_2$], and proportion of causal variants. First, we first simulated values where the total proportion of causal variants is low: $p_{00} = 0.89$, $p_{10} = 0.05, p_{01} = 0.05, p_{11} = 0.01$ with low sample size and high heritability: $h_1^2 = .05, h_2^2 = .05, \rho = 0, N_1 = 1000, N_2 = 1000$ , as shown in (a). Second, we tested the model with the same proportion of causal variants but given a large sample size and smaller heritability: $h_1^2 = .001, h_2^2 = .001, \rho = 0, N_1 = 100,000, N_2 = 100,000$ , shown in (b). Third, we simulated data with a higher proportion of causal variants, $p_{00} = 0.50$, $p_{10} = 0.20$, $p_{01} = 0.20$, $p_{11} = 0.10$. Using the same sets of heritabilities and sample sizes from the first two simulations, we tested the prediction accuracy of our model. Box (c) denotes the simulation with low sample size and high heritability, and box (d) denotes the simulation with high sample size and low heritability. The dotted red lines denote the true proportion of causal SNPs in each simulation.

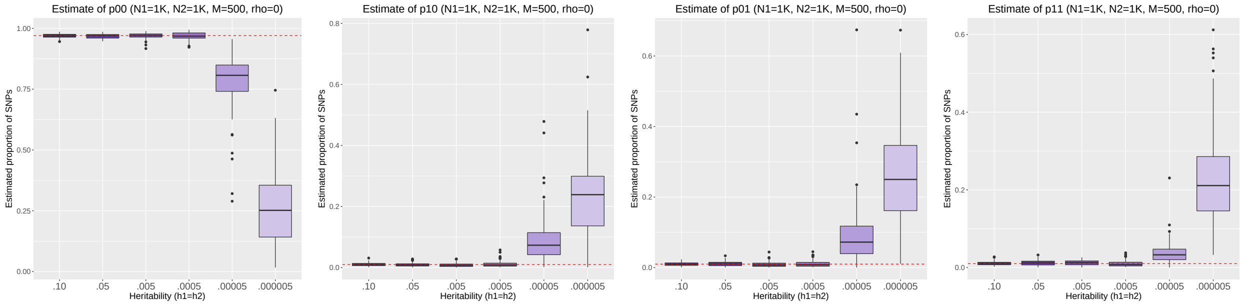

Figure 2.3: We simulate the following proportion of causal variants $p_{00} = 0.97$, $p_{10} = 0.01$, $p_{01} = 0.01$, $p_{11} = 0.01$ and vary the heritability $[h_1 = h_2]$ while fixing $\rho, N, M$. We vary the heritability from .01 to $5e-7$ and plot the estimated proportion of non-causal variants, the proportion of causal variants for trait 1, the proportion of causal variants for trait 2, and the proportion of shared causal variants (d). We note that as the heritability goes down, the data becomes less informative and the estimates reflect the prior.

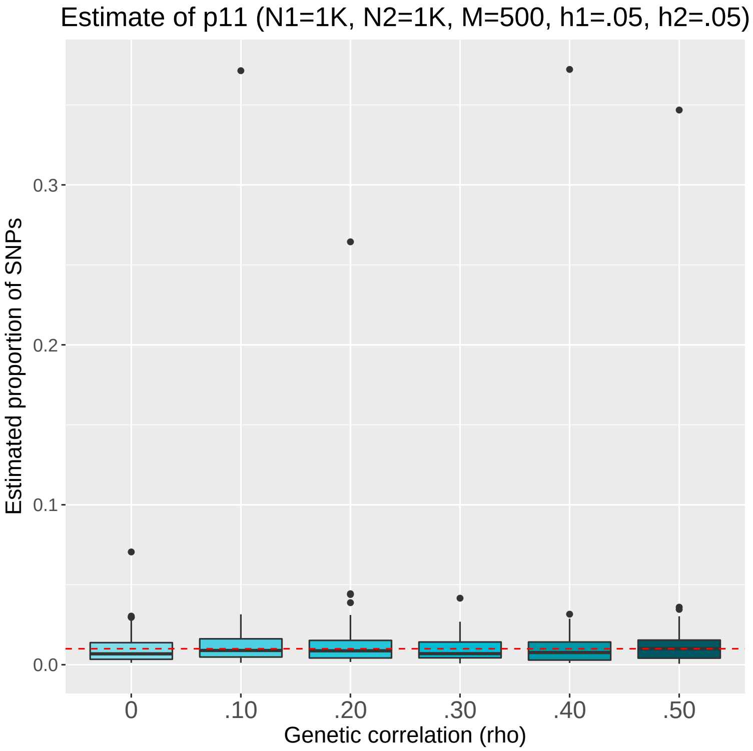

Figure 2.4: We simulate the following proportion of causal variants $p_{00} = 0.97$, $p_{10} = 0.01$, $p_{01} = 0.01$, $p_{11} = 0.01$ and vary the genetic correlation from 0 to 0.50 while fixing $\rho, N, M$,$[h_1 = h_2]$. We only show the estimate of $p_{11}$ since this would be the only estimate directly affected by the presence of genetic correlation.

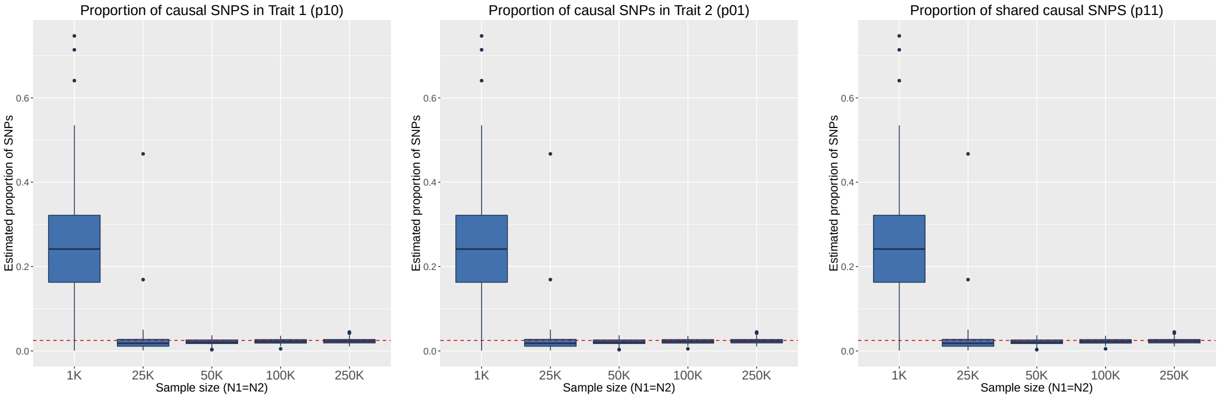

Figure 2.5: To assess the role of sample size in our inference, we performed simulations where we varied the number of individuals from 1,000 to 250,000. We simulated 100,000 SNPs where $h_1^2 = 0.25, h_2^2 = 0.25, \rho = 0.25, p_{10}, p_{01}, p_{11} = 0.01$. This was repeated for 100 independent simulations, and we report the posterior means for each simulation in the plots above. Note that the variance of our estimates increases when the sample size is under 25,000 individuals. We recommend users have at least 50,000 individuals for each trait to yield robust estimates.

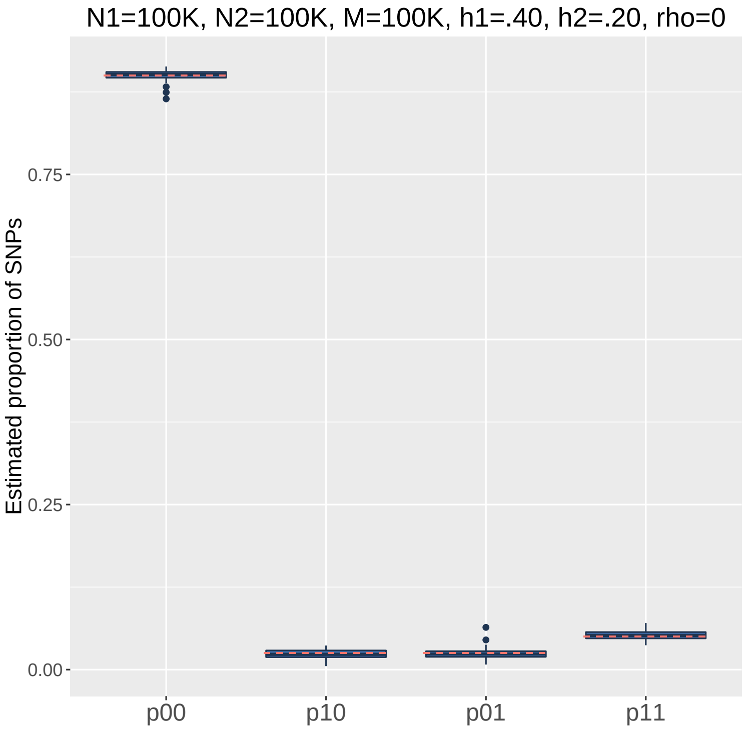

Figure 2.6: To assess whether our estimates are invariant to differing levels of heritability between traits, we performed simulations where $h_1^2 \neq h_2^2$. This was repeated for 100 independent simulations, and we report the posterior means for each simulation in the plots above.

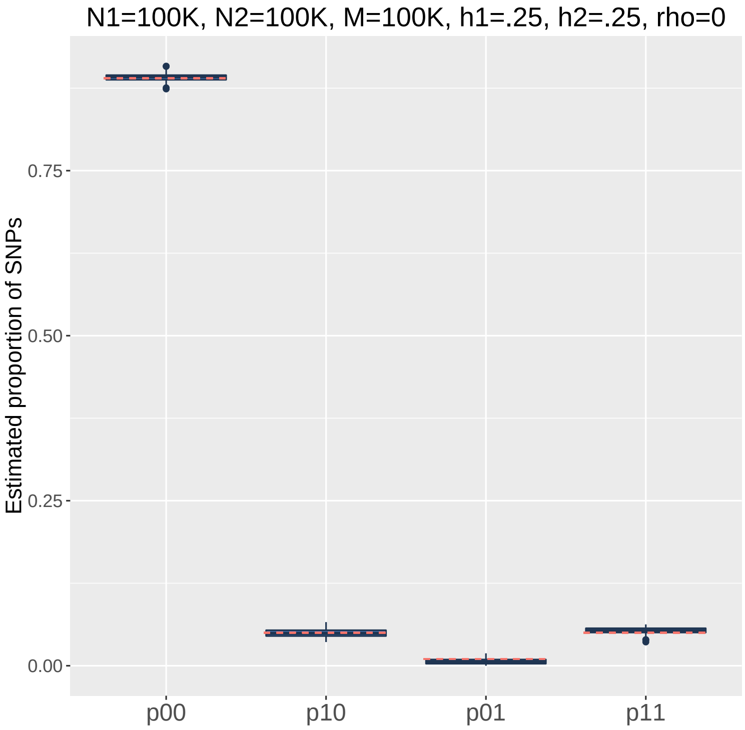

Figure 2.7: To assess whether our estimates are invariant to an unequal trait-specific proportion of causal SNPs, we performed simulations where $p_{10} \neq p_{01}$. This was repeated for 100 independent simulations, and we report the posterior means for each simulation in the plots above.

.

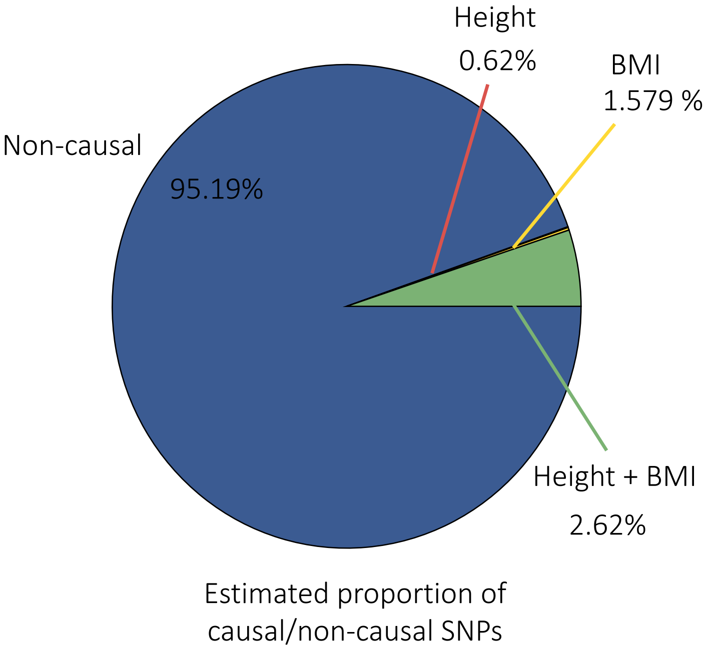

Figure 2.8: We show the distribution of estimated non-causal and causal SNPs from the Height and BMI analysis.

# CHAPTER 3

# Estimation of regional polygenicity from GWAS provides insights into the genetic architecture of complex traits

## 3.1 Introduction

*Polygenicity, i.e.*, the proportion of SNPs with nonzero effects on a trait, is a key quantity in efforts to understand the genetic architecture of complex traits. Accurate estimates of genome-wide polygenicity can be used to improve the prediction accuracy of polygenic risk scores[65, 323], quantify the strength of selection acting on a trait[322, 222], or better understand the biological complexity of the pathways driving disease risk[48, 178]. A major challenge in estimating polygenicity from genome-wide association study (GWAS) data arises due to the correlations between nearby SNPs, *i.e.* linkage disequilibrium (LD). In the presence of LD, methods for estimating polygenicity need to search over all possible causal status configurations at each SNP which, in turn, leads to an intractable computation for regions that harbor even a modest number of SNPs. Several methods implicitly model polygenicity in the context of phenotype prediction[211, 179, 328, 129, 128] whereas other methods explicitly aim to estimate polygenicity [322, 323], with recent methods overcoming the computational bottleneck by making simplifying model assumptions about the relationship between LD and polygenicity [323]. While all previous studies have focused on genome-wide polygenicity and its variation across traits [323], identification of genomic regions that are important for trait variation requires an understanding of how the number of causal SNPs varies across the genome (*regional polygenicity*).

In this work, we propose a statistical framework, Bayesian Estimation of Variants in a Region (BEAVR), to estimate regional polygenicity for a complex trait. Our approach estimates the proportion of causal variants in a given region ($p_r$) using marginal effect sizes from GWAS and in-sample LD information. In this work, we define 'causal variants' as a set of variants measured in a given GWAS study that have either a nonzero effect on the trait or tag unmeasured variants through LD that also have a nonzero effect. This particular definition does not imply a causal biological relationship nor formal causation as defined in causal inference. Thus, the estimates of polygenicity are defined with respect to the set of variants in the analyzed GWAS. This is similar to the definition of SNP-heritability estimates which are also specific to each set of variants and cannot be extrapolated to other sets of SNPs [267, 126, 52, 319].

The Bayesian model in BEAVR imposes a prior on the true SNP effect sizes where the probability of a non-zero true effect size at each SNP in the region is given by $p_r$[115, 206]. The observed GWAS effect sizes are obtained as a noisy combination of the unobserved true SNP effect sizes [123, 330]. We use Markov chain Monte Carlo (MCMC)[49] to approximate the posterior probability of the regional polygenicity parameter. This inference problem is computationally challenging as it requires disentangling correlations between SNPs due to LD. Leveraging the insight that the genetic architectures of most traits are likely to be sparse (so that most SNPs are not causal), we obtain a substantially more efficient MCMC algorithm that allows us to infer regional polygenicity across a large number of SNPs.

We validate our approach using extensive simulations and find that our method accurately estimates polygenicity in realistic settings; BEAVR estimates yield a relative bias $< 2\%$ across all simulations whereas existing methods obtain biased estimates, particularly in simulations with high degrees of polygenicity (*i.e.* $p_r > 5\%$). Next, we estimate regional polygenicity across 6-Mb regions for five quantitative anthropometric and blood pressure traits in the UK Biobank ($N = 290,641$ unrelated British individuals) restricting to genotyped SNPs with MAF $> 1\%$. Consistent with previous works [323], we find that all analyzed traits are highly polygenic at the genome-wide scale: over one-third of regions harbor at least one causal SNP across all traits. The proportion of regions containing at least one causal SNP (typically defined as regions with significant heritability) has been used as a proxy for polygenicity in earlier studies [180, 267]; we find that the proportion of regions containing at least one causal SNP is much higher than the estimated polygenicity. For example, while 79.6% of regions contain at least one causal SNP for height, the genome-wide polygenicity is estimated to be 3.07%. Additionally, we observe wide variation in regional polygenicity: on average across all analyzed traits, 48.9% of regions contain at least 5 causal SNPs while 5.44% of regions contain at least 50 causal SNPs, demonstrating the additional information provided from estimates of regional polygenicity. Finally, we find that within traits, regional SNP-heritability is proportional to regional polygenicity, suggesting that variation in heritability across the genome is largely driven by variation in the number of causal SNPs.

## 3.2 Methods

### 3.2.1 Generative model

We assume that the trait measured in individual $i$, $y_i$, is a linear function of standardized genotypes $\boldsymbol{x_i} = (x_{i,1}, \cdots, x_{i,M})$ measured at $M$ SNPs with true SNP effect sizes $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)$ and an independent additive noise term $\epsilon_i$.

$$y_i = \sum_{m=1}^{M} \beta_m x_{i,m} + \epsilon_i, i \in \{1, \cdots N\}$$

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$$

We model a non-infinitesimal trait architecture in which a subset of the $M$ SNPs are causal by imposing a spike-and-slab prior on the causal effect sizes $\boldsymbol{\beta}$ [211, 323, 322]. We represent the causal statuses across the SNPs as $\boldsymbol{c} = (c_1, \cdots, c_M)$. Here, $c_m = 1$ if SNP $m$ is a causal SNP with probability $p$ and 0 otherwise. Thus, $p$ denotes the proportion of causal SNPs or the *polygenicity*.

The Gaussian slab is parametrized with mean 0 and variance $\frac{h^2_{GW}}{Mp}$ where $h^2_{GW}$ is the genome-wide heritability. We draw independent Gaussian random variables for each of the $M$ SNPs: $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_M)$. The effect size $\beta_m$ is $\gamma_m$ if SNP $m$ is causal and 0 otherwise.

$$\gamma_m \sim \mathcal{N}(0, \frac{h^2_{GW}}{Mp}) \tag{3.1}$$

$$\beta_m \mid c_m, \gamma_m = \begin{cases} \gamma_m & \text{if } c_m = 1 \\ 0 & \text{if } c_m = 0 \end{cases} \tag{3.2}$$

We model the conditional distribution of the GWAS effect sizes given the true effect sizes where $\hat{\beta}_m$ is the estimated marginal effect size of SNP $m$ for the trait.

$$\hat{\boldsymbol{\beta}}|\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{V}\boldsymbol{\beta}, \boldsymbol{V}\sigma^2_e) \tag{3.3}$$

Here the covariance matrix is parametrized by the environmental noise $\sigma^2_e$ and the correlations among SNPs, *i.e.* the linkage disequilibrium (LD) matrix $\boldsymbol{V}$. The variance of the environmental noise term is parameterized by $\sigma^2_e = \frac{1-h^2_{GW}}{N}$, where $N$ is the number of individuals in the study.

We impose a symmetric Beta prior on the polygenicity parameter, $p$.

$$p \sim Beta(\alpha, \alpha) \tag{3.4}$$

In practice, we use $\alpha = 0.2$ to put a higher weight on the tails of the Beta distribution.

In this work, we focus on accurately estimating the proportion of causal variants in a given region $r$ (regional polygenicity, $p_r$). We assume that the above proposed genome-wide generative model holds when applied only within a specific region of the genome. This includes modeling the heritability only within that region ($h^2_r$) instead of the genome-wide heritability ($h^2_{GW}$). Modeling each region separately also assumes that there are no correlations between regions, such as correlations due to long-range LD. This assumption is reasonable when regions are chosen to correspond to LD blocks or when regions are sufficiently large such that correlations with adjacent regions may be ignored. Therefore, the LD matrix used in the regional model would only be the LD computed from SNPs within that particular region ($\boldsymbol{V}_r$). Additionally, although our framework naturally estimates the SNP effect sizes and posterior inclusion probabilities (*i.e.*, the probability that a given SNP is causal), we focus in this work on the posterior probability of $p_r$.

The posterior probability of the model parameters of interest $(p_r, \boldsymbol{\gamma}_r, \boldsymbol{c}_r)$ for a given region $r$ is given by:

$$P(p_r, \boldsymbol{\gamma}_r, \boldsymbol{c}_r \mid \hat{\boldsymbol{\beta}}_r, \alpha, h_r^2) \propto P(p_r \mid \alpha) P(\boldsymbol{c}_r \mid p_r) P(\boldsymbol{\gamma}_r \mid h_r^2, p_r) P(\hat{\boldsymbol{\beta}}_r \mid \boldsymbol{\gamma}_r, \boldsymbol{c}_r, h_r^2) \tag{3.5}$$

### 3.2.2 Inference

We use Markov Chain Monte Carlo (MCMC) to approximate the posterior probability as defined in Eq 3.5. Specifically, we derive a Gibbs sampler [106] to sample from the posterior distribution of the regional polygenicity $p_r$ and latent variables $(\boldsymbol{c}_r, \boldsymbol{\gamma}_r)$. The method takes as input the marginal effect sizes from GWAS for a single trait in a region $r$: $(\hat{\boldsymbol{\beta}}_r)$, the matrix of SNP correlations or LD per region $(\boldsymbol{V}_r)$, an estimate of the SNP heritability in that region $({h_r}^2)$, and the sample size of the GWAS $(N)$. As output, we estimate the posterior probability of the regional polygenicity for region $r$ $(p_r)$.

#### 3.2.2.1 Transforming GWAS effect sizes

To facilitate efficient inference, we transform the marginal effects from GWAS: $\tilde{\boldsymbol{\beta}}_r \equiv \boldsymbol{V}_r^{-\frac{1}{2}} \boldsymbol{\beta}_r$. The conditional probability of these transformed effects is given by:

$$\tilde{\boldsymbol{\beta}}_r \mid \boldsymbol{\beta}_r \sim \mathcal{N}(\boldsymbol{V}_r^{\frac{1}{2}} \boldsymbol{\beta}_r, \boldsymbol{I}_{M_r} \sigma_e^2)$$

Here, $\boldsymbol{I}_{M_r}$ is the identity matrix of size $M_r \times M_r$ where $M_r$ is the number of SNPs in region $r$. We note that this is a one-time transformation that is performed before running the sampler.

These transformed effects can be efficiently computed and stored for each genomic region.

#### 3.2.2.2 Sampling $\boldsymbol{\gamma}_r$ and $\boldsymbol{c}_r$

We recall that the true effect size at SNP $m$ in region $r$ ($\beta_{r,m}$) is given by a spike-and-slab prior parametrized by the causal effect size ($\gamma_{r,m}$) and the causal status at that SNP ($c_{r,m}$) (see Eq 3.2). We choose to sample $\gamma_{r,m}$ and $c_{r,m}$ together in a block step in the Gibbs sampler update.

Let $\boldsymbol{\theta}_r = \{(\boldsymbol{\gamma}_{\neg r,m}, \boldsymbol{c}_{\neg r,m}), h_r^2, p_r, \alpha\}$, where $\boldsymbol{\gamma}_{\neg r,m}$ denotes all effect sizes except for the effect of the $m^{th}$ SNP; this similarly follows for $\boldsymbol{c}_{\neg r,m}$.

$$P(\gamma_{r,m}, c_{r,m} \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) = P(\gamma_{r,m} \mid c_{r,m}, \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r)P(c_{r,m} \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r)$$

We are interested in the posterior probability of the causal effect size $\gamma_{r,m}$ when $c_{r,m} = 1$ since $P(\gamma_{r,m} \mid c_{r,m} = 0) = 0$ due to the spike-and-slab prior. This can be expressed as:

$$P(\gamma_{r,m} \mid c_{r,m} = 1, \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) \propto P(\tilde{\boldsymbol{\beta}}_r \mid \gamma_{r,m}, c_{r,m} = 1, \boldsymbol{\theta}_r)P(\gamma_{r,m} \mid c_{r,m} = 1, \boldsymbol{\theta}_r)$$

Working with the transformed GWAS effect sizes, the posterior distribution of $\gamma_{r,m}$ becomes univariate Gaussian with the following mean and variance. Here we denote $\boldsymbol{r}_{r,m} = \tilde{\boldsymbol{\beta}}_r - \boldsymbol{V}_r^{\frac{1}{2}}\boldsymbol{\gamma}_r \circ \boldsymbol{c}_r + \boldsymbol{V}_{r,m}^{\frac{1}{2}}\gamma_{r,m}c_{r,m}$, which is the residual from subtracting the effects of all SNPs except for SNP $m$ (here $\boldsymbol{V}_{r,m}^{\frac{1}{2}}$ denotes column $m$ of the matrix $\boldsymbol{V}_r^{\frac{1}{2}}$). We define $\sigma_{r,g}^2 = \frac{h_r^2}{M_r p_r}$ and $\sigma_e^2 = \frac{1-h_r^2}{N}$ for the region-specific model. See the appendix for full derivation details.

$$P(\gamma_{r,m} \mid c_{r,m} = 1, \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) = \mathcal{N}(\gamma_{r,m}; \mu_{r,m}, \sigma_{r,m}^2) \tag{3.6}$$
$$\frac{1}{\sigma_{r,m}^2} = \frac{1}{\sigma_{r,g}^2} + \frac{1}{\sigma_e^2}\boldsymbol{V}_{r,m}^{\frac{1}{2}\top}\boldsymbol{V}_{r,m}^{\frac{1}{2}}$$
$$\mu_{r,m} = \sigma_{r,m}^2\frac{1}{\sigma_e^2}\boldsymbol{r}_{r,m}^\top\boldsymbol{V}_{r,m}^{\frac{1}{2}}$$

We sample $c_{r,m} \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r$ from a Bernoulli distribution with parameter $P(c_{r,m} = 1 \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r)$:

$$
\begin{aligned}
P(c_{r,m} = 1 \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) &= \int P(c_{r,m} = 1, \gamma_{r,m} \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) d\gamma_{r,m} \\
&= \int \frac{P(\tilde{\boldsymbol{\beta}}_r \mid \gamma_{r,m}, c_{r,m} = 1, \boldsymbol{\theta}_r) P(\gamma_{r,m}, c_{r,m} = 1 \mid \boldsymbol{\theta}_r)}{P(\tilde{\boldsymbol{\beta}}_r \mid \boldsymbol{\theta}_r)} d\gamma_{r,m} \\
&= \frac{(p_r)\sqrt{\frac{\sigma^2_{r,m}}{\sigma^2_{r,g}}} \exp\left\{\frac{1}{2\sigma^2_{r,m}}\mu^2_{r,m}\right\}}{(p_r)\sqrt{\frac{\sigma^2_{r,m}}{\sigma^2_{r,g}}} \exp\left\{\frac{1}{2\sigma^2_{r,m}}\mu^2_{r,m}\right\} + (1 - p_r)} \\
&= d_{r,m}
\end{aligned}
$$

#### 3.2.2.3 Sampling $p_r$

The complete conditional posterior distribution of $p_r$ depends not only on the causal status of each SNP ($c_{r,m}$), but also on the latent variable ($\gamma_{r,m}$) since $p_r$ parameterizes the variance term of $\gamma_{r,m}$. We sample from this distribution using a random-walk Metropolis-Hastings step[49]. We use a Beta distribution as a proposal distribution:

$$
\begin{aligned}
p^*_r &\sim Q(p^*_r \mid p_r) \\
&= Beta\left(\alpha + Cp_r, \alpha + C(1 - p_r)\right)
\end{aligned}
$$

Here, $C$ is a constant that controls the variance of the proposal distribution. In practice, we found that $C = 10$ yields effective mixing.

### 3.2.3 Leveraging sparsity of the genetic architecture to improve the computational efficiency

The key computational bottleneck in the Gibbs sampling scheme involves computing the mean of the posterior distribution of the causal effect size at SNP $m$ ($\mu_{r,m}$ in Eq 3.6). Specifically, the matrix computations associated with the residual term, $\boldsymbol{r}_{r,m} = \tilde{\boldsymbol{\beta}}_r - \boldsymbol{V}_r^{\frac{1}{2}} \boldsymbol{\gamma}_r \circ \boldsymbol{c}_r + \boldsymbol{V}^{\frac{1}{2}}_{r,m} \gamma_{r,m} c_{r,m}$ , naively scales as $\mathcal{O}(M_r^2)$ due to the middle term, which is a matrix of size $M_r \times M_r$ multiplied by a vector of size $M \times 1$. Because this computation must be performed for every SNP, the overall complexity of the sampler is $\mathcal{O}(M_r^3)$ if implemented in this straightforward fashion. Below, we will break down the posterior mean term such that the complexity of computing $\boldsymbol{r}_{r,m}$ will only be $\mathcal{O}(K_r)$, where $K_r$ is the number of causal SNPs in the region, and the complexity of the sampler will be $\mathcal{O}(K_r M_r)$. This is accomplished by two steps: i) breaking the equation into constant terms that do not need to be updated at every iteration of the sampler, ii) leveraging the expected sparsity of the true causal vector and only performing computations over the causal SNPs.

Writing out the posterior mean term and expanding, we have:

$$
\begin{aligned}
\mu_{r,m} &= \frac{\sigma^2_{r,m}}{\sigma^2_e} \boldsymbol{r}^\top_{r,m} \boldsymbol{V}^{\frac{1}{2}}_{r,m} \\
&= \frac{\sigma^2_{r,m}}{\sigma^2_e} \left[ \tilde{\boldsymbol{\beta}}_r - \mathbf{V}^{\frac{1}{2}}_r \gamma_r \circ \mathbf{c}_r + \mathbf{V}^{\frac{1}{2}}_{r,m} \gamma_{r,m} c_{r,m} \right]^\top \boldsymbol{V}^{\frac{1}{2}}_{r,m} \\
&= \frac{\sigma^2_{r,m}}{\sigma^2_e} \left[ \tilde{\boldsymbol{\beta}}_r - \sum_{m \neq l}^{M_r} \boldsymbol{V}^{\frac{1}{2}}_{r,l} \gamma_{r,l} c_{r,l} \right]^\top \boldsymbol{V}^{\frac{1}{2}}_{r,m} \\
&= \frac{\sigma^2_{r,m}}{\sigma^2_e} \left[ \tilde{\boldsymbol{\beta}}^\top_r \boldsymbol{V}^{\frac{1}{2}}_{r,m} - \sum_{l \neq m, c_{r,l}=1}^{M_r} \boldsymbol{V}^{\frac{1}{2}\,\top}_{r,l} \boldsymbol{V}^{\frac{1}{2}}_{r,m} \gamma_{r,l} c_{r,l} \right]
\end{aligned}
$$

The first term, $\tilde{\boldsymbol{\beta}}^\top_r \boldsymbol{V}^{\frac{1}{2}}_{r,m}$, is composed of the vector of GWAS effect sizes and a vector of the LD matrix corresponding to the $m^{th}$ SNP, neither of which are updated within the sampler. Second, the term $\boldsymbol{V}^{\frac{1}{2}}_{r,l} \boldsymbol{V}^{\frac{1}{2}}_{r,m}$ can also be pre-computed since it is only the product of two columns within the LD matrix. Aside from the variance terms at the beginning of the equation, which are only scalars, the only term that varies at each iteration of the sampler is $\gamma_{r,l} c_{r,l}$ since both the effect size and causal status need to be re-sampled at each iteration. Since this term is wrapped in a summation over $M_r$ SNPs, the complexity of computing $\mu_{r,m}$ is currently $\mathcal{O}(M_r)$. However, even with this simplification, the overall complexity of the sampler is $\mathcal{O}(M_r^2)$ since this mean term must be computed at every SNP at every iteration.

To further simplify the computation, we can leverage the observation that most complex traits contain only a small proportion of causal SNPs ($K_r$) in each region. As the sampler converges to the stationary distribution, we would expect the causal status vector ($\mathbf{c}_r$) to be sparse, where $K_r << M_r$. When this occurs, the summation term will only include a few non-zero terms. By only subtracting the non-zero terms, this term is simply reduced to the number of causal variants and the complexity becomes $\mathcal{O}(K_r)$. Even though this computation must be done at each SNP, the overall complexity of the sampler is only $\mathcal{O}(K_r M_r)$ which is tractable under the assumption of $K_r << M_r$.

### 3.2.4 Simulation analysis

#### 3.2.4.1 Simulations for marginal effects using LD information

Using pre-computed LD information, we generated marginal effect sizes for a given region from synthetic GWAS that reflect a variety of genetic architectures. We denote the number of SNPs in a region as $M_r$ and the regional polygenicity as $p_r$. We denote the causal indicator status of each SNP in each region as $c_{r,m} \in \{0, 1\}$, where $c_{r,m} = 1$ if the $m^{th}$ SNP is causal and 0 otherwise for $m = 1, \cdots, M_r$ and regions $r = 1, \cdots, R$.

The causal status of a SNP is generated from:

$$c_{r,m} \sim Ber(p_r)$$

If $c_{r,m} = 1$, the effect size of SNP $m$ within the $r^{th}$ region is drawn from a univariate Gaussian distribution with mean 0 and variance equal to the regional heritability ($h^2_r$) divided by the number of casual SNPs:

$$\beta_{r,m} \sim \begin{cases} 0, \quad c_{r,m} = 0, \\ \mathcal{N}(0, \frac{h^2_r}{M_r p_r}), \quad c_{r,m} = 1 \end{cases}$$

Marginal association statistics for the region are then generated from the following model:

$$\hat{\boldsymbol{\beta}}_r \mid \boldsymbol{\beta}_r \sim \mathcal{N}\left(\boldsymbol{V}_r \boldsymbol{\beta}_r, \boldsymbol{V}_r \sigma^2_e\right)$$

Here, the environmental noise is a function of the sample size and heritability of the trait, $\sigma^2_e = \frac{1-h^2_r}{N}$. We use regional LD computed with genotypes from $337,205$ unrelated (less related than third-degree relatives), white, British individuals ($M_r = 1,000$ array SNPs) from the UK Biobank[55]. The LD matrix for a region is computed as $\boldsymbol{V}_r = \frac{\boldsymbol{X}_r^\top \boldsymbol{X}_r}{N}$, where $\boldsymbol{X}_r$ is the genotype matrix using only SNPs within region $r$.

Using the framework above, we generated marginal effect sizes where we varied the regional polygenicity from $p_r = 0.005, 0.01, 0.05,$ and $0.10$, genome-wide heritability from $h^2_{GW} = 0.10, 0.25,$ and $0.50$, and the sample size from $N = 50\text{K}, 500\text{K}, 1\text{M}$ individuals, which is comparable to the sample sizes of many current GWAS studies [163, 175]. For each simulated region, we set the number of SNPs per region to $1,000$. For the regional heritability parameter, we used the simulated genome-wide heritability scaled by the number of SNPs in the region, $M_r$, and the number of SNPs on the array, $M$: $h^2_r = \frac{h^2_{GW} M_r}{M}$.

To estimate the regional polygenicity, we ran BEAVR for 1,000 iterations with a burn-in of 250 iterations. We used the same LD information that was used for simulation (*i.e.* ”in-sample” LD). We also computed regional polygenicity using GENESIS[323]. We ran GENESIS using the default parameter settings and LD information from 1000 Genomes [74]. We used both the 2-component and 3-component settings when running GENESIS. We note that the implementation of GENESIS uses the 1000 Genomes LD matrix as a default and there is no option to specify an alternative LD matrix. We averaged the performance of each method across 100 replicates.

#### 3.2.4.2 Simulations for marginal effects computed from individual genotype and phenotype data

Using SNP data ($M = 9,564$ array SNPs from chromosome 22, $N = 337\text{K}$ individuals) from a group of unrelated, self-identified British, white ancestry individuals from the UK Biobank[55], we simulated marginal effects by generating phenotypes from real genotype array data. For this analysis, the set of unrelated individuals is defined as pairs of individuals with kinship coefficient $< \frac{1}{2}^{(9/2)}$ (greater than third-degree relatives) [55]. Then we performed ordinary least squares to estimate the marginal effect size of each SNP. Given the standardized genotype matrix $\boldsymbol{X}$ and the genome-wide SNP heritability $h^2_{GW}$, phenotypes are generated as follows.

We set the genome-wide proportion of causal variants to be $p = 0.01$. We denote the causal indicator status of each SNP as $c_m \in \{0, 1\}$, where $c_m = 1$ if the $m^{th}$ SNP is causal and 0 otherwise for $m = 1, \cdots, M$. Standardized effects and phenotypes are generated from the following model. Note that $\sigma_m^2 = 0$ if $c_m = 0$.

$$\sigma_m^2 = c_m \frac{h_{GW}^2}{Mp}$$

$$(\beta_1, \cdots, \beta_M)^\top \sim \mathcal{N}\left(0, \text{diag}(\sigma_1^2, \cdots, \sigma_M^2)\right)$$

$$(y_1, \cdots, y_N)^\top \mid \boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, (1 - h_{GW}^2)\boldsymbol{I}_N\right)$$

Finally, given the phenotypes for all individuals, $\boldsymbol{y} = (y_1, \cdots, y_N)^\top$ and genotypes $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_N^\top)^\top$, we compute marginal association statistics through the OLS estimator, $\hat{\boldsymbol{\beta}} = \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{y}$.

We generated 100 sets of marginal effect sizes where we fixed $p = 0.01$ and $h_{GW}^2 = 0.50$. We then estimated the regional polygenicity within each 6-Mb window for chromosome 22 (M=9,564 array SNPs) using BEAVR. This windowing formed 6 consecutive regions. We used HESS (Heritability Estimator from Summary Statistics)[267], a method for estimating regional heritability at a single region from GWAS summary statistics, to estimate the regional heritability which is then used as input for BEAVR. HESS is run with all default parameters and the same LD matrices used in the simulation framework (*i.e.* in-sample LD). We finally ran BEAVR for 1,000 iterations with a burn-in of 250 iterations and using the same LD information that was used for simulation.

### 3.2.5 Analysis of UK Biobank phenotypes

We estimated the partitioned polygenicity for five complex traits in the UK Biobank[55] across 6-Mb windows. We limited our analyses to unrelated individuals with self-identified British, white ancestry. Here, the set of unrelated individuals is defined as pairs of individuals with kinship coefficient $< \frac{1}{2}^{(9/2)}$ (greater than third-degree relatives)[55]. We additionally excluded individuals with putative sex chromosome aneuploidy. All genotypes were standardized, where for each SNP $m$ and individual $n$, we computed $x_{nm} = (g_{nm} - 2f_m)/\sqrt{2f_m(1 - f_m)}$, where $g_{nm} \in \{0, 1, 2\}$ is the number of minor alleles and $f_m$ is the in-sample minor allele frequency (MAF). We then used PLINK[63] (https://www.cog-genomics.org/plink2) to exclude SNPs with MAF $< 0.01$, genotype missingness $> 0.01$, and SNPs that fail the Hardy-Weinberg test at significance threshold $10^{-7}$. We obtained a final set of $N = 290,641$ individuals for our analyses.

Marginal association statistics were computed through OLS using PLINK. Age, sex, and the top 20 genetic PCs were used as covariates in the regression, where these top 20 PCs were pre-computed by the UK Biobank from a superset of $488,295$ individuals. Additional covariates were used for waist-to-hip ratio (adjusted for body mass index (BMI)) and diastolic/systolic blood pressure (adjusted for cholesterol-lowering medication, blood pressure medication, insulin, hormone replacement therapy, and oral contraceptives).

The genome is then divided into 6-Mb windows. Using HESS[267], we estimated the regional heritability within each window for each trait. HESS is run with all default parameters specified and in-sample LD. Using BEAVR and the computed regional heritability estimates, we estimated the regional polygenicity in each 6-Mb window. To initialize the MCMC sampler, we must set initial values for the vector of causal statuses, causal effect sizes, and regional polygenicity $(\boldsymbol{c}_r, \boldsymbol{\gamma}_r, p_r)$. For each SNP $m$, if the z-score estimated from GWAS is $\geq 3.5$, then $c_{r,m}$ is initialized to 1 and 0 otherwise. Each causal effect size is drawn from the prior distribution (see Eq 3.1). The initial value of $p_r$ is set to the proportion of 1's in the initialized causal status vector. We ran the Gibbs sampler for $1,000$ iterations and the first 250 samples were discarded as burn-in. For each region, we computed the posterior mean and posterior standard deviation for $p_r$ from the MCMC samples.

### 3.2.6 Annotations in regression analysis

We performed a multivariate regression of the heritability on the estimated number of SNPs from BEAVR, the number of causal SNPs, and genomic annotations within a region. The genomic annotations include the number of genes, median $B$ value (a measure of background selection), and functional annotations[96]. We computed the number of protein-coding genes within a region using the protein-coding gene sets that have been defined in previous work[102]. If a gene body overlapped two regions, we included the presence of the gene in both regions. Using previously computed $B$ values[203], we computed the median $B$ value of all the SNPs in a region. This quantity was used as the annotation value for that particular region. We additionally included a combination of binary and continuous functional annotations[96]. For each region, we computed the median annotation value for continuous annotations and the proportion of variants with a binary annotation.

## 3.3 Results

### 3.3.1 Simulations

We compare BEAVR to GENESIS[323], an approach that employs a spike-and-slab mixture model to capture both large and small effect sizes at causal SNPs in order to estimate polygenicity at a genome-wide scale (see Methods). To be applicable in genome-wide settings, GENESIS assumes that LD patterns are independent of the probability of a SNP belonging to different mixture components which, in turn, leads to a scalable algorithm. As shown in Figure 3.1, BEAVR obtains approximately unbiased estimates of polygenicity across each scenario (relative bias $< 2\%$ across the simulations). Both the two and three mixture component models from GENESIS obtain relatively unbiased estimates when the true polygenicity is low but demonstrate a severe downward bias in the high polygenicity setting (relative bias $> 70\%$ when $p_r = 0.10$). This observation is consistent with our hypothesis that not fully modeling LD limits the ability of GENESIS to accurately estimate parameters, consistent with previously reported downward bias when GENESIS was run with external LD information [323].

Next, we assessed the robustness of our approach to sample size and heritability. We vary the genome-wide heritability to be 0.10 and 0.25 and the sample size to be 50K and 1 million individuals (Figure 3.2) to fully explore the limitations of our method. We note that when the regional polygenicity $p_r$ is high, BEAVR demonstrates a downward bias either when sample sizes are relatively small ($N = 50$K individuals) (relative bias 56% and 80% for $p_r = 0.05$ and $p_r = 0.10$ when $h^2_{GW} = 0.50$ ) or when the heritability is low ($h^2_{GW} = 0.10$) (relative bias 54% and 73% for $p_r = 0.05$ and $p_r = 0.10$ for $N = 500$K). These biases likely arise due to the causal effect sizes being similar in magnitude to the environmental noise, making it difficult to correctly identify the causal status of a SNP. Thus, we recommend applying BEAVR to heritable traits measured in large sample sizes.

Next, we performed simulations where GWAS marginal effects are computed from phenotypes simulated from individual genotypes and the regional heritability is estimated directly from the data. Specifically, we simulate phenotypes using individual genotypes for $N = 337$K individuals from the UK Biobank. Each phenotype is simulated to have $h^2_{GW} = 0.50$ and polygenicity $p = 0.01$. We limit our simulations to SNPs from chromosome 22 ($M = 9,564$ SNPs) as each chromosome would be analyzed separately in real data analyses. We then estimate the marginal effect sizes. We divide the simulated data into consecutive regions of 6-Mb for a total of 6 regions, where each region contains 1,000 SNPs on average. We use estimates of regional heritability from GWAS marginal effects (using HESS [267]; see Methods) as input to BEAVR. We find that BEAVR obtains relatively unbiased estimates of polygenicity across all regions (Figure 3.3A); relative bias $< 2\%$ across simulations). These simulations indicate that the polygenicity estimates obtained by BEAVR are robust to heritability estimates that are used as input as well as when LD spans regions. The LD does not significantly affect the estimates likely because the correlation due to LD tends to diminish with genomic distance.

We also explored the robustness of BEAVR to the number of SNPs in the region. Using a simulated GWAS with genome-wide heritability $h^2_{GW} = 0.50$, sample size $N = 500$K, and polygenicity $p_r = 0.01$, we vary the size of the region from $M_r = 500, 1\text{K}, 5\text{K}$ SNPs. From Figure 3.3B, we can see that the estimates of $p_r$ tend to be unbiased across regions of various sizes although the standard errors tend to increase in smaller regions (relative bias $13\%, 1\%,$ and $1.2\%$ for $M_r = 500, 100, 5000$ SNPs). This trend occurs because larger regions have a higher number of SNPs to inform the posterior distribution, meaning that there will be higher certainty in the posterior estimates. Additionally, if a region is small, there is a larger impact on the estimated polygenicity when misidentifying causal SNPs due to the small denominator of SNPs in the region. For example, misidentifying a single causal SNP from a set of 10 SNPs will have a greater impact on the bias of polygenicity estimates compared to a set of 1,000 SNPs. These results suggest that BEAVR could potentially be applied to regions of varying length and be used to estimate regional polygenicity around genes or within larger LD blocks.

We next assess the sensitivity of our results when using different hyper-parameters for our prior on the polygenicity parameter $p_r$. Using a simulated GWAS with genome-wide heritability $h^2_{GW} = 0.50$, sample size $N = 500\text{K}$, and polygenicity $p_r = 0.01$, we vary our choice of hyper-parameter for the prior on $p_r$: $\alpha = 0.2, 1, 2$. We find that the accuracy of our results is relatively robust to the choice of prior (Figure 3.3C); we use $\alpha = 0.2$ for all subsequent analyses.

Although we assume that causal effect size distributions follow a Gaussian distribution, there are likely traits that do not follow this assumption. We next evaluate the performance of BEAVR when the true causal effect sizes deviate from these the assumption of normality. We next estimate regional polygenicity across 3 causal effect size distributions defined by a mixture of Gaussian distributions with the following set of variance components and mean 0: $[1 \times 10^{-3}, 1 \times 10^{-4}]$; $[1 \times 10^{-4}, 1 \times 10^{-5}]$; $[1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}]$. We assume a sample size of $N = 500\text{K}$, total regional polygenicity $p_r = 0.01$, and assume the true heritability of the region is known. The number of causal SNPs is spread equally amongst all the mixture components. We see that for causal effect sizes drawn from the distribution with larger variances components (*e.g.* $1 \times 10^{-3}$), our estimates are relatively unbiased.

However, for distributions with smaller variance components (*e.g.* $1 \times 10^{-5}$ ), we start to see a downward bias proportional to the fraction of SNPs drawn from the distribution with the smaller variance component(s). Thus, it is not necessarily the exact shape of the distribution of effect sizes, but the magnitude of the causal effect sizes, which affects the accuracy of the estimates.

### 3.3.2 Effect of out-of-sample LD

Although we recommend using in-sample LD when computing estimates of regional polygenicity, we also investigate the scenario where only LD derived from a reference panel is available. We simulate two scenarios: i) reference panel LD is computed from genotypes from individuals of a similar continental population as the target GWAS population but from a separate study (*e.g.* European ancestry individuals from the 1000 Genomes Project); ii) reference panel LD is computed with genotypes from a specific cohort/study and the target GWAS is also conducted with a subset of data from the same the cohort/study or a different version of the study (e.g. 'White, British' individuals from the UK Biobank). This second scenario closely reflects situations where many groups separately apply for freezes of data from the same study yet share GWAS summary statistics across applications.

We simulate the first scenario by simulating 1,000 GWAS regions with $M = 1,000$ SNPs, regional polygenicity $p_r = 0.01$, regional heritability, $h_r^2 = 0.0001$ (corresponding to a genome-wide heritability of 0.50), a sample size of $N = 500$K, and use a LD matrix computed from $337,205$ genotypes from unrelated individuals within the 'White, British' population from the UK Biobank[55]. However, inference is then performed using a reference panel derived LD matrix computed from 503 European ancestry individuals from the 1000 Genomes Project[74]. We find that when using LD from these separate studies, BEAVR fails to accurately estimate the regional polygenicity. Although the reference panel is constructed using individuals of European ancestry, these individuals were sampled from multiple subcontinental ancestries in Europe (*e.g.* Italy, Spain, Finland). In comparison, the target GWAS population from the UK Biobank is ancestrally homogeneous since it is limited to 'White, British' individuals within the UK.

The second scenario uses the same simulation parameters as above, except the GWAS effect sizes are computed using a LD matrix derived from 168,602 individuals from the unrelated, 'White, British' population within the UK Biobank. Inference is then performed using LD estimated from a separate, non-overlapping set of 168,602 individuals also from the unrelated, 'White, British' population within the UK Biobank. When using LD computed from a separate set of individuals from the same study, we find that our estimates are approximately unbiased. Our findings show that one can perform inference using a reference panel constructed from a separate set of individuals than used in the GWAS when both sets of individuals are from the same study (*e.g.* UK Biobank). These findings suggest that LD reference panels cannot solely be matched based on the continental ancestry level but need to be matched on a much finer scale. Additionally, differences in study designs between the genotypes used for the LD reference panel and the genotypes used when performing the GWAS may also contribute to discrepancies between the estimated LD structure.

### 3.3.3 Computational efficiency

BEAVR uses Gibbs sampling [49] to estimate the posterior probability of the regional polygenicity parameter. A naive implementation of the Gibbs sampler has a per-iteration computational complexity of $\mathcal{O}(M_r^2)$, where $M_r$ is the number of SNPs in the region. By leveraging the expected sparsity of the causal status at each SNP, we can improve the run-time of the algorithm to $\mathcal{O}(M_rK_r)$, where $K_r$ is the number of causal SNPs in the region. Figure 3.4A shows that this improvement leads to a 12-fold improvement in run-time for a region with $5,000$ SNPs. To assess how the number of causal SNPs affects the efficiency of our algorithm, we generated simulated GWAS data for 1,000 SNPs and varied the regional polygenicity from $p_r = 0.005, 0.01, 0.02, 0.03, 0.04, 0.05$ and observe efficiency gains across the range of parameters (Figure 3.4B). The optimization of our method makes it possible to efficiently analyze regions of various sizes as well as densely imputed regions with thousands of variants.

### 3.3.4 Contrasting genome-wide and regional polygenicity across complex traits

We applied BEAVR to estimate regional polygenicity from marginal effect size estimates for five anthropometric and blood pressure traits from the UK Biobank (see Table 3.1). Marginal association statistics were computed for each of these traits from a subset of unrelated individuals identified as White British (see Methods). We applied BEAVR by dividing the genome into a total of 470 6-Mb regions where each region has on average $1,000$ SNPs. Since BEAVR requires an estimate of LD between the SNPs, we used in-sample LD, *i.e.*, LD computed on the White British subset of the UK Biobank. We additionally used HESS [267] to estimate regional heritability. Since BEAVR produces a posterior distribution of the regional polygenicity, we report a region to have nonzero polygenicity if the posterior mean - ($2\times$ posterior standard deviations) does not overlap 0. Furthermore, we estimate the genome-wide polygenicity for a trait as the sum of the posterior means of regional polygenicity across all regions.

Consistent with previous estimates of genome-wide polygenicity [323], we observe that all the analyzed traits are highly polygenic. Across the traits, we observe that over one-third of the regions in the genome contain at least one causal SNP, and overall each of the traits is estimated to harbor at least $1,000$ causal SNPs (Table 3.1). We also observe variation across traits: for height, nearly 80% of the regions contain at least one causal SNP and the total number of causal SNPs could be as high as $15,000$ while blood pressure traits are estimated to harbor about $2,500 - 3,000$ causal SNPs. Our estimates for the proportion of causal SNPs for height is significantly higher than previously reported[323] (Table 3.1): the 95% credible interval estimated by BEAVR is $(3.0\%, 3.2\%)$ while the estimates from prior work [323] are 0.9% with standard error 0.1%. We hypothesize that this difference is due, in part, to our method capturing smaller effect sizes by fully modeling LD, which is consistent with our simulations, but could also arise from the differences in SNP sets and GWAS summary statistics analyzed.

Previous studies have used the proportion of genomic regions with nonzero heritability as a proxy for polygenicity since nonzero heritability requires at least one causal SNP in the region[180, 267]. However, the distribution of regional heritability does not fully reflect the distribution of regional polygenicity (Figure 3.5). Across the traits, the proportion of regions containing at least one causal SNP is substantially higher than the estimated proportion of causal SNPs across the genome (Table 3.1): while $\approx 80\%$ of regions contain at least one causal SNP for height, we estimate that $\approx 3\%$ of the SNPs are casual. Further, we observe wide variation in regional polygenicity where, across the analyzed traits, nearly 50% of regions contain at least 5 causal SNPs and about 5% of regions contain at least 50 causal SNPs (Figure 3.5). These results demonstrate the additional information that can be obtained from estimates of regional polygenicity.

### 3.3.5 Heritability is proportional to the number of causal SNPs

Previous studies have documented a linear relationship between chromosome length and the per-chromosome heritability for multiple traits suggesting that the architecture of these traits is highly polygenic[320, 267]. We replicate this relationship between the number of SNPs and the heritability in a genomic region for each trait ($p$-value $= 1.22 \times 10^{-13}$; $R^2 = 0.162$ averaged across traits; Table 3.3). In addition, we observe that a linear regression of heritability on the number of causal SNPs in the region is significant ($p$-value $= 2.60 \times 10^{-21}$; $R^2 = 0.278$ averaged across traits) (Table 3.3). We also observe that the number of causal SNPs in a region better explains regional heritability than the number of overall SNPs in the region. This ranges from approximately the same $R^2$ in systolic blood pressure to nearly three times in WHR (Table 3.3). The slope of the regression of regional heritability on the number of causal SNPs averaged across traits is $1.63 \times 10^{-5}$, which can be interpreted as the heritability per additional causal SNP (Table 3.4). Performing multiple regression, we find that both the number of SNPs and the number of causal SNPs have a significant relationship to the heritability in a region (average $p$-value $= 3.60 \times 10^{-39}$; $R^2 = 0.374$). We hypothesize that the number of SNPs and causal SNPs together explains more of the variation in heritability than the number of causal SNPs alone due, in part, to inaccurate estimates of the number of causal SNPs and regional heritability as well as possible misspecifications in the model assumed by BEAVR.

We further investigate the relationship between genomic annotations and heritability as well as the number of causal SNPs in a region. Including the number of genes, median $B$ value, and functional annotations [96] as covariates in the regression (see Methods), only the number of causal SNPs remains significant (average $p$-value $= 6.37 \times 10^{-11}$, $p$-value $<$ 0.05/(number of annotations)) while the total number of SNPs in the region remains significant for 3 out of 5 traits (Table 3.5). None of the other genomic annotations are significant after the multiple testing correction.

While the expected regional heritability can be partly explained by the number of causal SNPs, we also observe regions that have disproportionately high heritability given the number of estimated causal SNPs (Figure 3.6). These outlier regions (defined as regions with an absolute studentized residual larger than 3) are likely to harbor SNPs with larger effect sizes compared to other regions. Consistent with this hypothesis, 24 out of 27 outlier regions contain at least one genome-wide significant SNP for at least one trait. This proportion is significantly higher than a randomly chosen set of 27 regions ($p$-value $< \frac{1}{1{,}000}$). Taken together, our analyses indicate that the heritability of a trait is composed of a mixture of small-effect SNPs as well as some SNPs with relatively larger effects.

Finally, we also investigate whether the gene density in a region plays a role in the observed regional polygenicity estimates. We perform a likelihood ratio test between the following two models to assess the effect of gene density on the number of causal SNPs ($M_{C_r}$) after adjusting for both regional heritability and the number of SNPs: $H_0 : M_{C_r} \sim h_r^2 + M_r; H_1 : M_{C_r} \sim h_r^2 + M_r + \#\text{genes}$. As shown in Table 3.6, we find that only the likelihood ratio test for height is significant after adjusting for the number of tested traits ($p$-value $< \frac{0.05}{5}$). This observation could be due to the fact that we included all protein-coding genes in the analysis regardless of the specific biological mechanism of each gene. For example, when analyzing BMI, one would expect regions with genes related to lipids or metabolism to harbor more causal variants than genes related to seemingly unrelated biological mechanisms. This observed effect of gene density on the polygenicity of height is consistent with the hypothesis that genes throughout the genome, regardless of the specific biological function, contribute to the variance of height. Previous work has shown that height is one of the most polygenic traits with numerous causal variants spread throughout the genome[319]. This idea that numerous genes, regardless of functional mechanism, have an effect on a trait is related to the recently proposed 'omnigenic' model[48].

## 3.4 Discussion

In this work, we propose BEAVR, a novel, scalable method to estimate regional polygenicity from GWAS effect size estimates in a Bayesian framework. We employ a fast inference algorithm that enables efficient inference while fully accounting for LD. Applying BEAVR to anthropometric and blood pressure traits in the UK Biobank, we observe that all of the analyzed traits are highly polygenic. At least a third of 6-Mb regions harbor at least one causal variant with this fraction rising as high as 80% for height. We find that the proportion of regions containing at least one causal SNP, which is often used as a proxy for polygenicity in previous studies, is much higher than our estimates of the proportion of causal SNPs. Additionally, we observe wide variation in regional polygenicity with an average of 48.9% of regions across the analyzed traits containing at least 5 causal SNPs and 5.44% of regions containing at least 50 causal SNPs. Finally, we find that the number of causal SNPs better explains variation in SNP heritability across regions compared to the total number of SNPs.

The observed polygenic architecture of complex traits supports the hypothesis that the majority of trait variation is modulated by variants distributed across the genome. Trait heritability is largely driven by the number of causal variants and most of these variants are spread uniformly across the genome. This finding suggests that a large proportion of genes have at least some, although limited, effect on a trait. These findings are consistent with the recently proposed omnigenic model which suggests that disease risk is driven by a combination of a small number of primary 'core' genes and numerous 'peripheral' genes which are connected to core genes via highly interconnected gene networks [48].

We conclude by discussing limitations of our study and directions for future work. First, our model assumes that the causal effects are drawn from a single Gaussian distribution. This assumption can be relaxed and other distributions (such as mixtures of Gaussians) can be used instead. Second, our estimates of genome-wide polygenicity assume that the LD matrix is block structured which allows us to estimate genome-wide polygenicity by applying our method to regions corresponding to LD blocks. Finally, our analyses in the UK Biobank were limited to array data and thus the set of SNPs used in our analyses are missing true causal SNPs that were not typed. We leave a more thorough investigation of this scenario and analyses on imputed data as future work.

## 3.5 Appendix

### 3.5.1 Additional derivations

### 3.5.2 Sampling $\gamma_{r,m}, c_{r,m}$

We derive a Gibbs sampler to sample from the posterior distribution of each parameter $\{c_{r,m}, \gamma_{r,m}, p_r\}$. Because the causal status and effect size of a SNP are highly correlated, we sample $(\gamma_{r,m}, c_{r,m})$ together in a block.

Let $\boldsymbol{\theta}_r = \{(\boldsymbol{\gamma}_{\neg r,m}, \boldsymbol{c}_{\neg r,m}), h_r^2, p_r, \alpha\}$, where $\boldsymbol{\gamma}_{\neg r,m}$ denotes all effect sizes except for the effect of the $m^{th}$ SNP; this similarly follows for $\boldsymbol{c}_{\neg r,m}$. We denote $\sigma_{r,g}^2 = \frac{h_r^2}{M_r p_r}$ and $\sigma_e^2 = \frac{1-h_r^2}{N}$. The derivation for each marginal posterior distribution, $P(\gamma_{r,m} \mid \cdot)$ and $P(c_{r,m} \mid \cdot)$ is given separately below. By the chain rule note that:

$$P(\gamma_{r,m}, c_{r,m} \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) = P(\gamma_{r,m} \mid c_{r,m}, \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) P(c_{r,m} \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r)$$

Additionally, for convenience we denote $\boldsymbol{r}_{r,m} = \tilde{\boldsymbol{\beta}}_r - \boldsymbol{V}_r^{\frac{1}{2}} \boldsymbol{\gamma}_r \circ \boldsymbol{c}_r + \boldsymbol{V}_{r,m}^{\frac{1}{2}} \gamma_{r,m} c_{r,m}$, which is the residual from subtracting the effects of all SNPs except for SNP $m$.

Deriving the first term, $P(\gamma_{r,m} \mid c_{r,m}, \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r)$, we break up the expression into the cases when $c_{r,m} = 1$ and $c_{r,m} = 0$:

$$
\begin{aligned}
P(\gamma_{r,m} \mid c_{r,m} = 1, \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) \propto\ & P(\tilde{\boldsymbol{\beta}}_r \mid \gamma_{r,m}, c_{r,m} = 1, \boldsymbol{\theta}_r) P(\gamma_{r,m} \mid \boldsymbol{\theta}_r, c_{r,m} = 1) \\
&= \exp\Big\{ -\frac{1}{2\sigma_e^2} (\boldsymbol{r}_{r,m} - \boldsymbol{V}_{r,m}^{\frac{1}{2}} \gamma_{r,m})^\top (\boldsymbol{r}_{r,m} - \boldsymbol{V}_{r,m}^{\frac{1}{2}} \gamma_{r,m}) \Big\} \exp\Big\{ -\frac{1}{2\sigma_{r,g}^2} \gamma_{r,m}^2 \Big\} \\
&= \exp\Big\{ -\frac{1}{2\sigma_e^2} (\boldsymbol{r}_{r,m}^\top \boldsymbol{r}_{r,m} - 2\boldsymbol{r}_{r,m}^T \boldsymbol{V}_{r,m}^{\frac{1}{2}} \gamma_{r,m} + \boldsymbol{V}_{r,m}^{\frac{1}{2}^\top} \boldsymbol{V}_{r,m}^{\frac{1}{2}} \gamma_{r,m}^2) - \frac{1}{2\sigma_{r,g}^2} \gamma_{r,m}^2 \Big\} \\
& [\text{drop constants that don't depend on } \gamma_{r,m}] \\
&= \exp\Big\{ -\frac{1}{2\sigma_e^2} (-2\boldsymbol{r}_{r,m}^\top \boldsymbol{V}_{r,m}^{\frac{1}{2}} \gamma_{r,m} + \boldsymbol{V}_{r,m}^{\frac{1}{2}^\top} \boldsymbol{V}_{r,m}^{\frac{1}{2}} \gamma_{r,m}^2) - \frac{1}{2\sigma_{r,g}^2} \gamma_{r,m}^2 \Big\} \\
& [\text{common denominators}] \\
&= \exp\Big\{ -\frac{1}{2\sigma_e^2 2\sigma_{r,g}^2} (-2\boldsymbol{r}_{r,m}^\top \boldsymbol{V_m}^{\frac{1}{2}} 2\sigma_{r,g}^2 \gamma_{r,m} + \boldsymbol{V}_{r,m}^{\frac{1}{2}^\top} \boldsymbol{V}_{r,m}^{\frac{1}{2}} 2\sigma_{r,g}^2 \gamma_{r,m}^2 + 2\sigma_e^2 \gamma_{r,m}^2) \Big\} \\
&= \exp\Big\{ -\frac{\gamma_{r,m}^2}{2} \left( \frac{1}{\sigma_{r,g}^2} + \frac{1}{\sigma_e^2} \boldsymbol{V}_{r,m}^{\frac{1}{2}^\top} \boldsymbol{V}_{r,m}^{\frac{1}{2}} \right) + \gamma_{r,m} \left( \frac{1}{\sigma_e^2} \boldsymbol{r}_{r,m}^\top \boldsymbol{V}_{r,m}^{\frac{1}{2}} \right) \Big\} \\
&= \exp\Big\{ -\frac{\gamma_{r,m}^2}{2} (a) + \gamma_{r,m} (b) \Big\} \\
& \Big( a = -\frac{1}{2\sigma_{r,m}^2}, b = \frac{\mu_{r,m}}{\sigma_{r,m}^2} \Big) \\
&= \mathcal{N}(\mu_{r,m}, \sigma_{r,m}^2) \\
& \frac{1}{\sigma_{r,m}^2} = \frac{1}{\sigma_{r,g}^2} + \frac{1}{\sigma_e^2} \boldsymbol{V}_{r,m}^{\frac{1}{2}^\top} \boldsymbol{V}_{r,m}^{\frac{1}{2}} \\
& \mu_{r,m} = \sigma_{r,m}^2 \frac{1}{\sigma_e^2} \boldsymbol{r}_{r,m}^T \boldsymbol{V}_{r,m}^{\frac{1}{2}}
\end{aligned}
$$

$$P(\gamma_{r,m} \mid c_{r,m} = 0, \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) = \delta_0(\gamma_{r,m})$$

The bottom line follows because the effect size of a non-causal SNP is 0 ($c_{r,m} = 0$).

Deriving the second term, $P(c_{r,m} \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r)$:

$$
\begin{aligned}
P(c_{r,m} = 1 \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) &= \int P(c_{r,m} = 1, \gamma_{r,m} \mid \boldsymbol{\theta}_r, \tilde{\boldsymbol{\beta}}_r) d\gamma_{r,m} \\
&= \int \frac{P(\tilde{\boldsymbol{\beta}}_r \mid \gamma_{r,m}, c_{r,m} = 1, \boldsymbol{\theta}_r) P(\gamma_{r,m}, c_{r,m} = 1 \mid \boldsymbol{\theta}_r)}{P(\tilde{\boldsymbol{\beta}}_r \mid \boldsymbol{\theta}_r)} d\gamma_{r,m} \\
&= \int \frac{P(\tilde{\boldsymbol{\beta}}_r \mid \gamma_{r,m}, c_{r,m} = 1, \boldsymbol{\theta}_r) P(\gamma_{r,m} \mid c_{r,m} = 1\boldsymbol{\theta}_r) P(c_{r,m} = 1 \mid \boldsymbol{\theta}_r)}{P(\tilde{\boldsymbol{\beta}}_r \mid \boldsymbol{\theta}_r)} d\gamma_{r,m} \\
&= \frac{P(c_{r,m} = 1 \mid \boldsymbol{\theta}_r)}{P(\tilde{\boldsymbol{\beta}}_r \mid \boldsymbol{\theta}_r)} \int P(\tilde{\boldsymbol{\beta}}_r \mid \gamma_{r,m}, c_{r,m} = 1, \boldsymbol{\theta}_r) P(\gamma_{r,m} \mid c_{r,m} = 1, \boldsymbol{\theta}_r) d\gamma_{r,m} \\
&\text{[denominator does not depend on } c_{r,m}\text{]} \\
&= P(c_{r,m} = 1 \mid \boldsymbol{\theta}_r) \int \Big[ \frac{1}{\sqrt{2\pi\sigma^2_{r,m}}} \exp\Big\{ -\frac{1}{2\sigma^2_{r,m}} (\gamma_{r,m} - \mu_{r,m})^2 \Big\} \Big] d\gamma_{r,m} \\
&\quad \times \sqrt{2\pi\sigma^2_{r,m}} \exp\Big\{ -\frac{1}{2\sigma_e^2} \boldsymbol{r}^\top_{r,m} \boldsymbol{r}_{r,m} + \frac{1}{2\sigma^2_{r,m}} \mu^2_{r,m} \Big\} \\
&= P(c_{r,m} \mid \boldsymbol{\theta}_r) \frac{\sqrt{2\pi\sigma^2_{r,m}}}{\sqrt{2\pi\sigma_e^2}\sqrt{2\pi\sigma^2_{r,g}}} \exp\Big\{ -\frac{1}{2} \frac{\boldsymbol{r}^\top_{r,m} \boldsymbol{r}_{r,m}}{\sigma_e^2} + \frac{1}{2\sigma^2_{r,m}} \mu^2_{r,m} \Big\}
\end{aligned}
$$

To sample $c_{r,m}$, we draw $c_{r,m} \sim \text{Bern}(d_{r,m})$, where $d_{r,m}$ is defined as follows:

$$d_{r,m} = \frac{\frac{\sqrt{2\pi\sigma^2_{r,m}}}{\sqrt{2\pi\sigma^2_e}\sqrt{2\pi\sigma^2_{r,g}}} P(c_{r,m}=1) \exp\Big\{ -\frac{1}{2}\frac{\boldsymbol{r}^\top_{r,m}\boldsymbol{r}_{r,m}}{\sigma^2_e} + \frac{1}{2\sigma^2_{r,m}}\mu^2_{r,m} \Big\}}{\frac{\sqrt{2\pi\sigma^2_{r,m}}}{\sqrt{2\pi\sigma^2_e}\sqrt{2\pi\sigma^2_{r,g}}} P(c_{r,m}=1 \mid \boldsymbol{\theta}_r) \exp\Big\{ -\frac{1}{2}\frac{\boldsymbol{r}^\top_{r,m}\boldsymbol{r}_{r,m}}{\sigma^2_e} + \frac{1}{2\sigma^2_{r,m}}\mu^2_{r,m} \Big\} + \frac{1}{\sqrt{2\pi\sigma^2_e}} P(c_{r,m}=0) \exp\Big\{ -\frac{1}{2\sigma^2_e}\boldsymbol{r}^\top_{r,m}\boldsymbol{r}_{r,m}}$$

$$= \frac{\frac{\sqrt{2\pi\sigma^2_{r,m}}}{\sqrt{2\pi\sigma^2_e}\sqrt{2\pi\sigma^2_{r,g}}} (p_r) \exp\Big\{ -\frac{1}{2}\frac{\boldsymbol{r}^\top_{r,m}\boldsymbol{r}_{r,m}}{\sigma^2_e} + \frac{1}{2\sigma^2_{r,m}}\mu^2_{r,m} \Big\}}{\frac{\sqrt{2\pi\sigma^2_{r,m}}}{\sqrt{2\pi\sigma^2_e}\sqrt{2\pi\sigma^2_{r,g}}} (p_r) \exp\Big\{ -\frac{1}{2}\frac{\boldsymbol{r}^\top_{r,m}\boldsymbol{r}_{r,m}}{\sigma^2_e} + \frac{1}{2\sigma^2_{r,m}}\mu^2_{r,m} \Big\} + \frac{1}{\sqrt{2\pi\sigma^2_e}} (1-p_r) \exp\Big\{ -\frac{1}{2\sigma^2_e}\boldsymbol{r}^\top_{r,m}\boldsymbol{r}_{r,m} \Big\}}$$

[break up terms over exp]

$$= \frac{\frac{\sqrt{2\pi\sigma^2_{r,m}}}{\sqrt{2\pi\sigma^2_e}\sqrt{2\pi\sigma^2_{r,g}}} (p_r) \exp\Big\{ -\frac{1}{2}\frac{\boldsymbol{r}^\top_{r,m}\boldsymbol{r}_{r,m}}{\sigma^2_e} \Big\} \exp\{\frac{1}{2\sigma^2_{r,m}}\mu^2_{r,m} \Big\}}{\frac{\sqrt{2\pi\sigma^2_{r,m}}}{\sqrt{2\pi\sigma^2_e}\sqrt{2\pi\sigma^2_{r,g}}} (p_r) \exp\Big\{ -\frac{1}{2}\frac{\boldsymbol{r}^\top_{r,m}\boldsymbol{r}_{r,m}}{\sigma^2_e} \Big\} \exp\{\frac{1}{2\sigma^2_{r,m}}\mu^2_{r,m} \Big\} + \frac{1}{\sqrt{2\pi\sigma^2_e}} (1-p_r) \exp\Big\{ -\frac{1}{2\sigma^2_e}\boldsymbol{r}^\top_{r,m}\boldsymbol{r}_{r,m} \Big\}}$$

[common exp terms and constants from top/bottom]

$$= \frac{(p_r)\sqrt{\frac{\sigma^2_{r,m}}{\sigma^2_{r,g}}} \exp\Big\{\frac{1}{2\sigma^2_{r,m}}\mu^2_{r,m}\Big\}}{(p_r)\sqrt{\frac{\sigma^2_{r,m}}{\sigma^2_{r,g}}} \exp\Big\{\frac{1}{2\sigma^2_{r,m}}\mu^2_{r,m}\Big\} + (1-p_r)}$$

In summary, to jointly sample $(\gamma_{r,m}, c_{r,m})$, one first samples $c_{r,m}$. Then depending on if $c_{r,m} = 1$ we sample $\gamma_{r,m}$, and if $c_{r,m} = 0$ we set $\gamma_m$ to a point mass at 0:

$$c_{r,m} \sim \text{Bern}(d_{r,m})$$

$$\gamma_{r,m} \sim \begin{cases} \mathcal{N}(\mu_{r,m}, \sigma^2_{r,m}) & \text{if } c_{r,m} = 1 \\ 0 & \text{if } c_{r,m} = 0 \end{cases}$$

## 3.6 Tables

Table 3.1: **Genome-wide estimates of polygenicity and total SNP heritability.**

| **Trait** | % | $\mathbf{h^2_{GW}}$ | **p** | $\mathbf{M_c}$ |
|---|---|---|---|---|
| *BMI* | 66.2 | 0.30 (0.004) | (0.017, 0.018) | $(7.67 \times 10^3, 8.41 \times 10^3)$ |
| *Height* | 79.6 | 0.64 (0.004) | (0.030, 0.032) | $(1.37 \times 10^4, 1.45 \times 10^4)$ |
| *Waist-hip ratio* | 40.6 | 0.18 (0.004) | (0.007, 0.008) | $(3.12 \times 10^3, 3.57 \times 10^3)$ |
| *Diastolic blood pressure* | 35.5 | 0.16 (0.004) | (0.006, 0.007) | $(2.54 \times 10^3, 3.01 \times 10^3)$ |
| *Systolic blood pressure* | 34.9 | 0.17 (0.004) | (0.006, 0.007) | $(2.58 \times 10^3, 3.01 \times 10^3)$ |

Table 3.2: We report the percentage of 6-Mb regions containing at least one causal SNP under the column '%'. Genome-wide estimates of polygenicity and heritability were computed by aggregating estimates across all regions. The standard error is reported for genome-wide heritability estimates. Here, $p$ denotes the proportion of causal SNPs and $M_c$ denotes the total number of causal SNPs (we report the 95% posterior credible interval for each of these parameters).

Table 3.3: **Linear relationship between heritability, number of SNPs, number of causal SNPs, and genomic annotations.**

| **Trait** | $\mathbf{R^2(h_r^2 \sim M_r)}$ | $\mathbf{R^2(h_r^2 \sim M_{Cr})}$ | $\mathbf{R^2(h_r^2 \sim M_r + M_{Cr})}$ | $\mathbf{R^2(h_r^2 \sim}$ **all-annotations**$)$ |
|---|---|---|---|---|
| *BMI* | $0.182\ (3.67 \times 10^{-22})$ | $0.226\ (7.64 \times 10^{-28})$ | $0.347\ (6.38 \times 10^{-44})$ | $0.532\ (1.20 \times 10^{-27})$ |
| *Height* | $0.172\ (5.34 \times 10^{-21})$ | $0.447\ (2.96 \times 10^{-62})$ | $0.501\ (3.30 \times 10^{-71})$ | $0.647\ (3.23 \times 10^{-47})$ |
| *Waist-hip ratio* | $0.105\ (6.13 \times 10^{-13})$ | $0.295\ (2.05 \times 10^{-37})$ | $0.352\ (1.08 \times 10^{-44})$ | $0.540\ (8.23 \times 10^{-29})$ |
| *Diastolic blood pressure* | $0.183\ (2.60 \times 10^{-22})$ | $0.254\ (1.08 \times 10^{-31})$ | $0.359\ (7.33 \times 10^{-46})$ | $0.530\ (2.09 \times 10^{-27})$ |
| *Systolic blood pressure* | $0.168\ (2.09 \times 10^{-20})$ | $0.169\ (1.30 \times 10^{-20})$ | $0.311\ (1.77 \times 10^{-38})$ | $0.532\ (1.11 \times 10^{-27})$ |

In the first column, we model the linear relationship between the heritability of a trait and the number of SNPs across all regions of the genome. We report the coefficient of determination ($R^2$). The relationship is significant for all traits ($p$-values are reported in parentheses). We observe a similar trend relating the heritability and number of causal SNPs in a region. We perform a multivariate regression to assess the relationship between the heritability and both the number of SNPs and the number of causal SNPs in a region. Finally, in the last column, we perform a multivariate regression of heritability on the number of SNPs, number of causal SNPs, number of genes, median $B$ value, and functional annotations[96]. Significant annotations are listed in Table 3.5.

| Trait | OLS slope (CI) | Number outlier regions |
|---|---|---|
| BMI | $1.078 \times 10^{-5}$ $(1.1 \times 10^{-6}$ ) | 4 |
| Height | $2.874 \times 10^{-5}$ $(1.48 \times 10^{-6})$ | 9 |
| Waist-hip ratio | $1.781 \times 10^{-5}$ $(1.27 \times 10^{-6})$ | 4 |
| Diastolic blood pressure | $1.337 \times 10^{-5}$ $(1.06 \times 10^{-6})$ | 5 |
| Systolic blood pressure | $1.078 \times 10^{-5}$ $(1.06 \times 10^{-6})$ | 5 |

Table 3.4: Linear relationship between the number of causal SNPs and heritability. We model the linear relationship between the number of causal SNPs for a trait and the heritability across all regions of the genome. We report the slope of the regression and the standard error. The slope can be interpreted as the expected per-SNP heritability contribution per causal SNP. The last column reports the number of 'outlier' regions for each trait, defined as a region with an absolute studentized residual greater than 3.

| Trait | Annotation | p-value |
|---|---|---|
| BMI | $M_{C_r}$ | 1.11E-13 |
|  | $M_r$ | 3.33E-04 |
| Height | $M_{C_r}$ | 1.79E-28 |
|  | $M_r$ | 4.91E-04 |
| Waist-hip ratio | $M_{C_r}$ | 3.73E-21 |
| Diastolic blood pressure | $M_{C_r}$ | 6.79E-19 |
|  | $M_r$ | 3.38E-04 |
| Systolic blood pressure | $M_{C_r}$ | 3.18E-10 |

Table 3.5: Covariates that are associated with regional heritability $\mathbf{h_r^2}$. We perform a multivariate regression of heritability on the number of SNPs, number of causal SNPs, number of genes, median $B$-statistic, and non-cell-type-specific annotations. Only the number of causal SNPs ($M_{C_r}$) remains significant for all traits after the multiple testing correction (average $p$-value $=6.37 \times 10^{-11}$), and the number of SNPs ($M_r$) remains significant for 3 out of 5 traits after the multiple testing correction.

| Trait | p-value |
|---|---|
| Systolic blood pressure | 0.028 |
| Height | $3.93 \times 10^{-7}$ |
| Waist-hip ratio | 0.070 |
| BMI | 0.073 |
| Diastolic blood pressure | 0.096 |

Table 3.6: Likelihood ratio test assessing the role of gene density in regional polygenicity estimates. We perform a likelihood ratio test between the following two models to assess the effect of gene density on the number of causal SNPs ($M_{C_r}$) after adjusting for both regional heritability and the number of SNPs ($H_0 : M_{C_r} \sim h_r^2 + M_r; H_1 : M_{C_r} \sim h_r^2 + M_r + \#\text{genes}$).

## 3.7 Figures

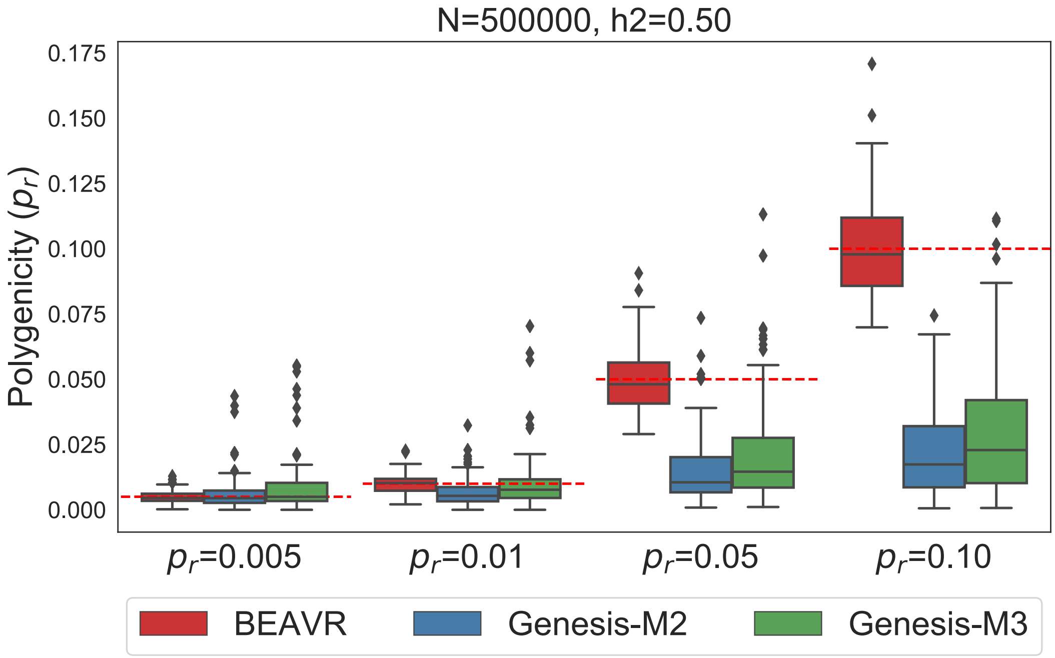

Figure 3.1: **BEAVR is relatively unbiased in simulated data.** We ran 100 replicates ($M = 1,000$ SNPs, $N = 500$K individuals) where the genome-wide heritability was set to $h^2_{GW} = 0.5$ and the true polygenicity of the region was $p_r = 0.005, 0.01, 0.05, 0.10$. We compared BEAVR to GENESIS-M2 and GENESIS-M3 which employs a spike-and-slab model with either 2 or 3 components (point-mass and either 1 or 2 slabs). All methods are unbiased when the polygenicity is low ($p_r = 0.005, 0.01$). However, when polygenicity is higher ($p_r = 0.05, 0.10$), both GENESIS-M2 and GENESIS-M3 are severely downward biased whereas BEAVR provides unbiased estimates across all settings. Dashed red lines denote true regional polygenicity values in each setting.

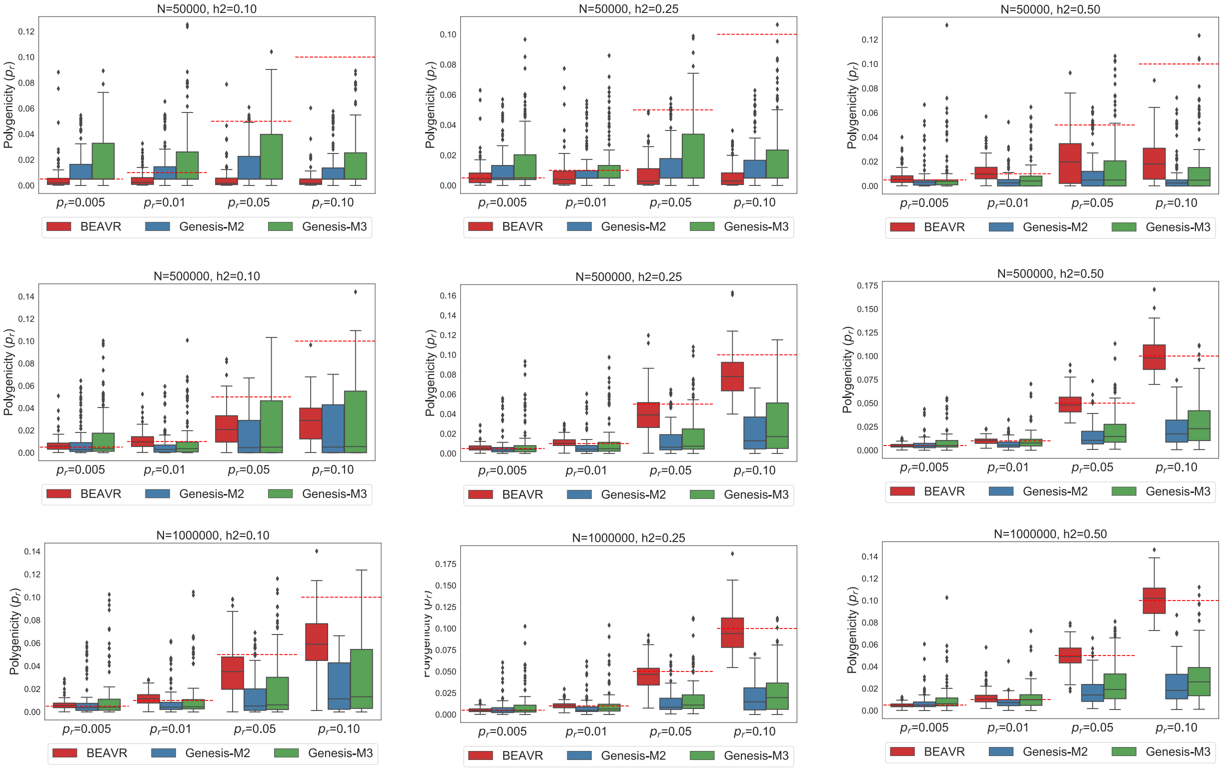

Figure 3.2: **BEAVR is relatively unbiased across various genetic architectures.** We ran 100 replicates where we vary the genome-wide heritability to be $h^2_{GW} = 0.10, 0.25, 0.5$, the polygenicity of the region to be $p_r = 0.005, 0.01, 0.05, 0.10$, and the sample size $N = 50\text{K}, 500\text{K}, 1$ million individuals. We compared BEAVR to GENESIS-M2 (2-component) and GENESIS-M3 (3-component). The x-axis denotes the simulated values for the regional polygenicity and the y-axis denotes the estimated values across 100 replicates. Dashed red lines denote the true regional polygenicity value in each setting.

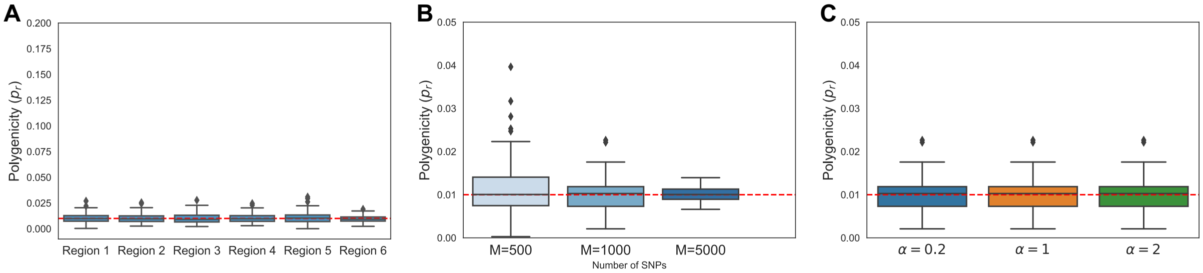

Figure 3.3: **BEAVR is robust in realistic settings.** **(A)** Using SNP data from chromosome 22 ($M = 9,564$ array SNPs, $N = 337$K individuals), we simulated 100 replicates where the genome-wide heritability was $h^2_{GW} = 0.50$ and $p = 0.01$. We divided the data into 6-Mb consecutive regions for a total of 6 regions and estimated the regional heritability using external software (HESS[267]). Using BEAVR and the estimated regional heritability, we estimated the regional polygenicity to be unbiased across all regions. **(B)** We ran 100 replicates where the genome-wide heritability is fixed $h^2_{GW} = 0.50$, polygenicity $p_r = 0.01$, sample size $N = 500$K, and then varied the number of SNPs in the region from $M = 500, 1\text{K}, 5\text{K}$ SNPs. We used BEAVR to estimate the polygenicity in each region and found our results to be unbiased across all regions. **(C)** We set the genome-wide heritability to $h^2_{GW} = 0.50$, regional polygenicity $p_r = 0.01$, and sample size $N = 500$K. We find that the accuracy of our results is invariant to our choice of prior hyper-parameter ($\alpha$).

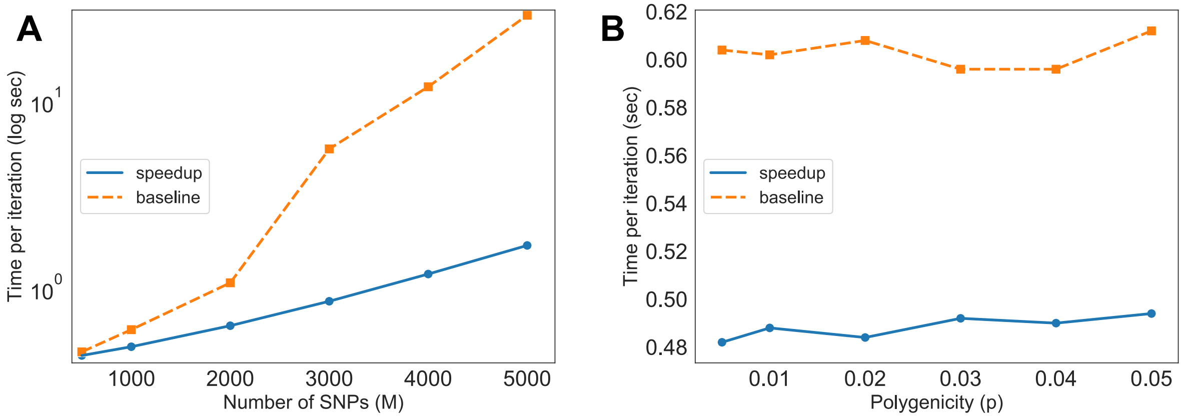

Figure 3.4: **BEAVR is computationally efficient. (A)** We show the run-time in terms of seconds per iteration of the Gibbs sampler (log-scale). We compare the version of BEAVR with the algorithmic speedup outlined in Methods ('speedup') versus the straightforward implementation ('baseline'). We vary the number of SNPs in the region while fixing the polygenicity of each region to $p_r = 0.01$. **(B)** We show the runtime of the sampler when the number of SNPs in the region is fixed to $M = 1,000$ and we vary the polygenicity.

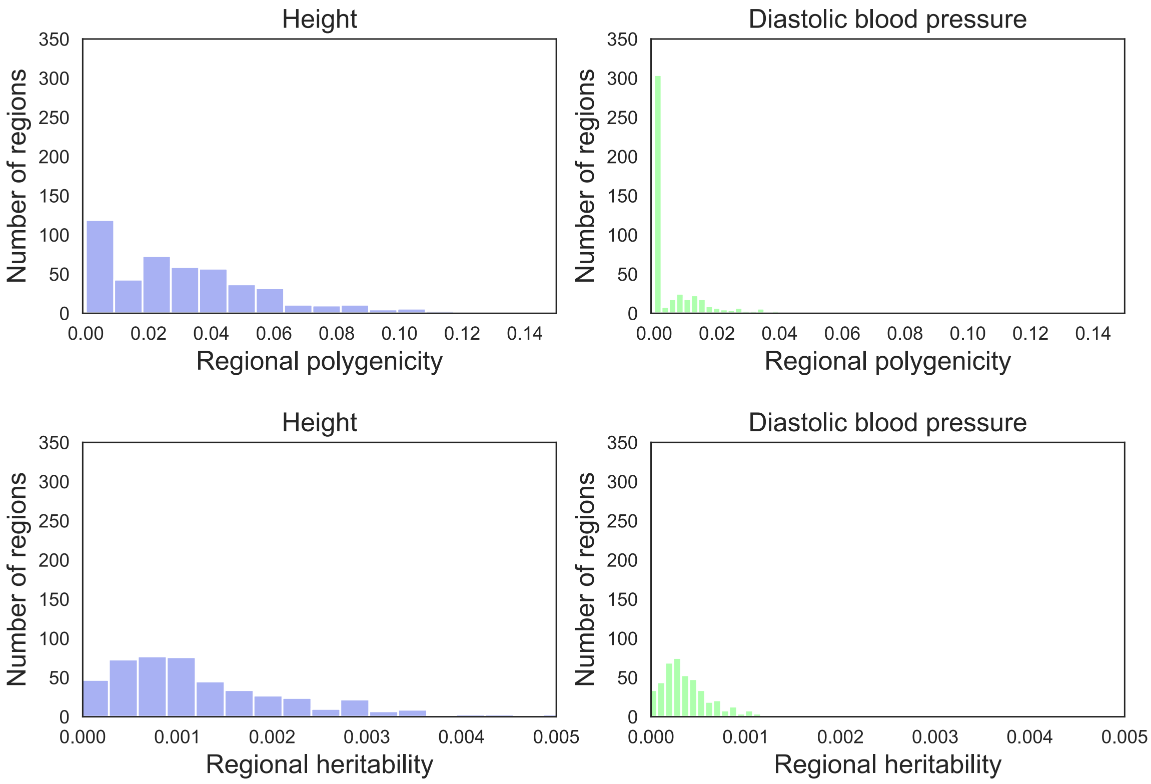

Figure 3.5: **Distribution of regional polygenicity and heritability.** We divide the genome into 6-Mb regions and report the posterior mean of the regional polygenicity for each region across height and diastolic blood pressure. Using external software[267], we report the distribution of regional heritability for each trait.

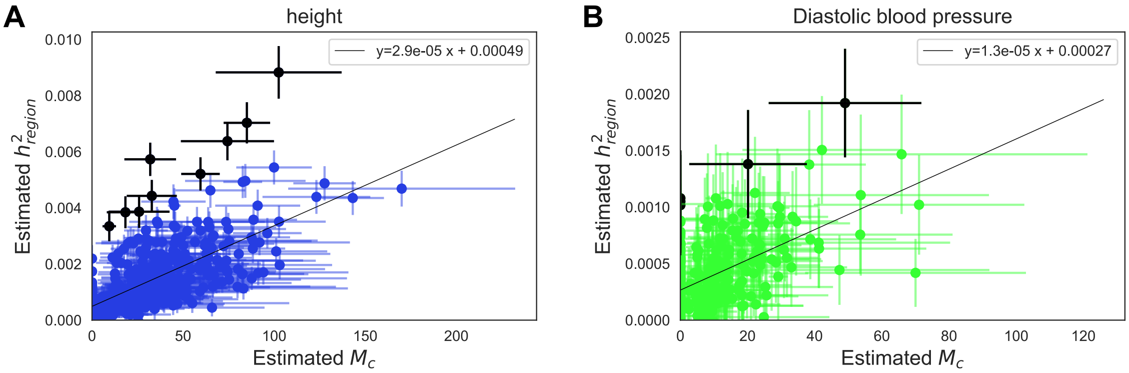

Figure 3.6: **Heritability is proportional to the number of causal SNPs in a region.** We show the relationship between the number of causal SNPs and heritability for each region across height and diastolic blood pressure. We fit a linear regression for each trait and report the slope of the regression, which can be interpreted as the increase of heritability per additional causal SNP. Horizontal error bars represent two posterior standard deviations around our estimates for the number of causal SNPs. Vertical error bars represent twice the standard error around the estimates of regional heritability. Dots in black denote outlier regions which have an absolute studentized residual larger than 3.

# CHAPTER 4

# The UCLA ATLAS Community Health Initiative: promoting precision health research in a diverse biobank

## 4.1 Introduction

The UCLA ATLAS Community Health Initiative (ATLAS), named for its location “At LA”, aims to recruit 150,000 participants from across the UCLA Health System, with the goal of creating California’s largest genomic resource for translational and precision medicine research. Each biosample is linked with the patient’s electronic health record (EHR) from UCLA Health via the UCLA Data Discovery Repository (DDR), a database containing a de-identified version of electronic health records. Participants are recruited from 18 UCLA Health medical centers, laboratories, and clinics located throughout the greater Los Angeles area. Participants watch a short video outlining the goals of the initiative and document their choice of whether they wish to consent to participation[213, 154]. Biological samples are collected during routine clinical lab work performed at any UCLA Health laboratory and then genotyped using a customized Illumina Global Screening Array (GSA)[2] (see Methods).

Both biological samples and EHR information are de-identified to protect patient privacy. As of September 2021, the initiative has enrolled 90,400 participants through the consent process and successfully genotyped 39,300 samples. Comprehensive details on the biobanking and consenting processes are described in prior work[154, 213]. In this work, we describe quality control pipelines for genotype curation and phenotype extraction from the medical records for the purpose of large-scale genotype and phenotype scans. To establish the genotyping QC pipelines, we present the first freeze of the data containing genotypes and phenotypes collected and processed up to September 2020, resulting in a total of N=27,987 samples.

## 4.2 Results

The UCLA Health System includes 2 hospitals and a total of 210 primary and specialty outpatient locations located primarily in the greater Los Angeles area. In total, the UCLA Health System serves approximately 5% of Los Angeles County population. An electronic form of health records was implemented throughout the UCLA Health System in 2013, where a variety of clinical information is recorded, such as laboratory tests, medications and prescriptions, diagnoses, and hospital admissions. A version of this information has been de-identified and approved for research purposes. The de-identification process removes some clinical data including names, family relationships, geographic information, exact dates, as well as exact ages for those at the extremes of age (>90 years old).

The average age of participants, defined as a participant's age recorded in the EHR as of September 2021, is 55.6 (SD: 17.2) years with an average medical record length of 11.6 (SD: 8.5) years. We use phecodes, a coding system that maps diagnosis codes (i.e. ICD-9 and ICD-10 codes) to more clinically meaningful phenotypes[85], to construct phenotypes from the EHR. The median number of unique phecodes per participant is 68 whereas the mean is 85.2 (SD: 65.0). This skewed mean is consistent with the presence of individuals with many more healthcare interactions than the average person in the general population, a pattern that has been well described in the literature[216].

Participants' self-identified race and ethnicity (SIRE) information are also recorded within the DDR where participants select a single option for their race and a single separate option for their ethnicity from multiple-choice lists. The majority of patients in ATLAS self-identify as White race (61.4%) and Non-Hispanic/Latino ethnicity (75.4%), although a substantial proportion of individuals report being of an Asian race (9.67%) or of Hispanic/Latino, Spanish, or Mexican ethnicity (14.1%). A full list of the provided race/ethnicity fields within the DDR and a summary of the ATLAS demographic information can be found in Table 4.1.

We regret that the term 'White/Caucasian' is a preset multiple-choice option under the race field within the medical records. The scientific and medical communities have since denounced this specific terminology due to its erroneous origins and historically racist implications[97, 238, 236] but it is still built into the language of many documents and surveys, such as those within electronic health record systems. In presenting our analyses, we omit the inclusion of the term 'Caucasian' when describing race and list the specific 'White/Caucasian' field only as 'White'. Furthermore, we strongly discourage the connection of the term 'Caucasian' with the discussion of race, a social construct separate from biology, and emphasize that the term does not have any biological implications.

### 4.2.1 Genotype generation and quality control

The ATLAS initiative continuously recruits new participants and batches of genotype samples are being processed on a rolling basis in monthly installments of approximately 1,000 samples per batch. Genotyping was performed at the UCLA Neuroscience Genomics Core using a custom genotyping array constructed from the Global Screening Array with the multi-disease drop-in panel[2] under the GRCh37 assembly. An additional set of "Pathogenic" and "Likely Pathogenic" variants selected from ClinVar[159] were additionally added to the chip design. The first freeze of genotype data presented in this work combines samples from 15 separate batches yielding a total of 697,023 SNPs and 27,987 individuals. Principal component analysis (PCA)[136] was used to visualize the variation across batches and did not show any evidence of batch effects.

We next describe the quality control pipeline used to filter out low-quality SNPs and samples while also considering the diverse ancestral backgrounds represented in ATLAS. In this work, we aim to focus on describing only the common genetic variation and leave a further in-depth analysis of rare variation in ATLAS to future work as sample sizes continue to grow. First, we excluded poor-quality SNPs with $> 5\%$ missingness as well as monomorphic SNPs and strand ambiguous SNPs, defined as those with A/T or C/G alleles. Samples with $> 5\%$ missingness were also removed. We estimated kinship coefficients using KING 2.2.2[190] and found 38 duplicate samples, 357 parent-offspring, 128 first-degree, and 166 second-degree relatives. This level of relatedness is not surprising since members of a family tend to attend the same health center. For the sets of duplicate samples, we removed the sample with the higher missing rate. A summary of the quality control pipeline and the number of filtered SNPs and individuals is outlined in Figure 4.1. Following sample- and variant-level quality control, M=673,130 genotyped SNPs remained across N=27,946 individuals (N=27,291 unrelated individuals).

After genotyping QC, we inferred biological sex using the '–sex-check' function with default thresholds implemented in PLINK 1.912 which estimates the X chromosome homozygosity or F statistic (Female: F $>$ 0.20, Male: F $>$ 0.80). We find that 45.5% of genotypes yielded a male call and 53.9% a female call while 0.6% of samples were estimated to be unknown (Table 4.1). For the group of individuals with unknown inferred sex, the mean F statistic was 0.27 (SD: 0.10). The sex of these individuals likely could not be inferred because the F statistics were slightly over the threshold. Next, using self-identified information from the EHR, we find that 45.1% of individuals self-identify as male and 54.9% self-identify as female (Table 1). Within the EHR, this specific field is labeled as 'Sex' and has a list of pre-determined multiple-choice fields where participants select one of the following options: 'Male', 'Female', 'Other', 'Unknown', '*Unspecified', 'X'. The mean F statistics for individuals who self-identified as male and female were 0.96 (SD: 0.06) and 0.06 (SD: 0.09) respectively. There were not any individuals in the current data who self-identified as one of the other listed options. We also observe that 0.04% of individuals who were inferred to be biologically male do not self-identify as male as reported from the EHR. This comparison is a common heuristic used to determine sample mismatch. However, this small deviation does not appear to reflect a systematic sample mismatch and instead could describe transgender and gender-nonconforming13 individuals. We retain these samples with appropriate documentation and encourage researchers utilizing the ATLAS data to perform further sex-based filtering based on their specific analysis criteria.

The final step of genotyping QC involves genotype imputation to the TOPMedFreeze5 reference panel, a multi-ancestry dataset assembled from over 50,000 ancestrally diverse genome[285], using the Michigan Imputation Server15. Overall, approximately 300 million SNPs and indels were used as the backbone for genotype imputation. The imputation process yielded a total of 230 million imputed SNPs from the ATLAS data. We found that SNPs with a lower minor allele frequency (MAF) tended to have lower imputation quality ($r^2$) scores. This demonstrates that rare SNPs were more difficult to accurately impute within ATLAS (Figure 4.2A) which is consistent with prior findings[325, 199, 234]. Due to this observation, SNPs with imputation r2 $<$ 0.90 or MAF $<$ 1% were pruned from the data, leaving a total of 7.9 million well-imputed SNPs across 27,946 individuals for follow-up analyses (Figure 4.1).

When performing genome-wide association studies, we stratified individuals by genetic ancestry groups and then performed an additional level of QC separately within each ancestry group. We limited analyses to the subset of 27,291 unrelated individuals ($>$ 2nd degree) and performed ancestry inference (see 'Genetic ancestry inference'), where each individual was assigned to one continental genetic ancestry cluster: European (N=18,023), African (N=1,340), Admixed American (N=4,930), East Asian (N=2,495), and South Asian ancestry (N=402). At this time, we omitted GWAS analyses within the South Asian ancestry group due to the limited sample size. Individuals who could not be clustered into a specific genetic ancestry group (N=756) were also omitted from GWAS analyses. Within each ancestry group, samples identified as heterozygosity outliers (+/- 3 SDs from the mean) were removed and SNPs that failed the Hardy-Weinberg equilibrium test (p-value ¡1 $\times$ $10^{-12}$) were also removed. Finally, we limited analyses to only SNPs with MAF ¿ 1% within each ancestry group, yielding a total of N=17,874 individuals and M=6.9 million SNPs within the European ancestry group, N=1,337 individuals and M=6.6 million SNPs within the African group, N=4,776 and M=7.2 million SNPs within the Admixed American group, and N=2,459 individuals and M=5.4 million SNPs within the East Asian group.

### 4.2.2 Genetic ancestry inference

The ATLAS data presents a unique resource to study genomic medicine across an ancestrally diverse set of individuals within a single medical system. Genetic ancestry information is necessary for numerous types of genetic and epidemiological studies, such as genome-wide association studies and polygenic risk score estimation. The EHR contains self-identified demographic information such as race and ethnicity, but these concepts are distinct from genetic ancestry, which describes the biological history of one's genome with little to no relation to cultural aspects of identity [309, 47]. Although self-identified race/ethnicity and genetic ancestry are correlated [249, 291], populations constructed from these two concepts are not analogous and capture distinct information. A thorough discussion of the role of ancestry within the ATLAS data can be found in ref [135].

Instead, we use PCA to identify population structure in ATLAS solely from genetic information as means to correct for genetic stratification in large-scale genotype/phenotype association studies. PCA produces a visual summary of the observed genetic variation which can then be used to describe population structure across the samples. First, we merged genotypes from ATLAS with individuals from the 1000 Genomes Project reference panel[7], which contains genotypes of individuals sampled from various populations: European, African, Admixed American, East Asian, and South Asian. We limited analyses to individuals in ATLAS unrelated to the 2nd degree (N=27,291) and performed PCA analyses on the ATLAS genotyped data (M=673,130) merged with individuals from the 1000 Genomes dataset. The top 10 PCs were then computed using the FlashPCA 2.0 software[8].

After projecting the PCs into two-dimensional space, we use the samples from 1000 Genomes to define clusters of individuals corresponding to each continental ancestry group. The first two PCs capture the variation between European, African, and East Asian ancestries. PCs 2 and 3 can approximately delineate individuals with Admixed American ancestry whereas PCs 4 and 5 can cluster individuals with South Asian ancestry (Figure 4.2B). Cluster thresholds were visually determined by comparing the overlap of the 1000 Genomes reference panel samples to ATLAS samples in PC space. Individuals who fell into multiple ancestry groups or could not be classified into any of the defined ancestry groups were labeled as 'Admixed or other ancestry'.

We find that 64.5% (N=18,023) of individuals are inferred to be of European ancestry, 4.8% (N=1,340) of African ancestry, 17.8% (N=4,930) of Admixed American ancestry, 8.9% (N=2,495) of East Asian ancestry, 1.5% (N=402) South Asian ancestry, and 2.7% (N=756) were characterized as 'Admixed or other ancestry' (Table 4.1). As expected, the inferred ancestry clusters were largely concordant with the self-identified race and ethnicity information provided in the EHR: 90.5% of individuals within the European ancestry group self-identified as White/Caucasian, 92.1% of the African ancestry group self-identified as Black or African American, 90.4% of the East Asian ancestry group self-identified as an Asian race, and 77.6% of the Admixed American ancestry group self-identified as either Hispanic or Latino, Puerto Rican, Mexican, or Cuban ethnicity. We observe that most individuals who self-identified to be of African American race tended to fall along the cline between the African and European ancestry clusters, demonstrating that genetic ancestry, in particular for admixed populations, often lies on a continuum rather than within discrete categorizations. These analyses demonstrate how the pairing between self-identified information and inferred genetic ancestry is not one-to-one, further emphasizing the important distinction between these two concepts. A full analysis on exploring the ancestral diversity within ATLAS is described in ref [135].

### 4.2.3 EHR-based phenotyping through the phecode system

In this work, we utilized phenotypes derived from the EHR in the form of phecodes, a mapping of ICD codes to a collapsed set of more clinically descriptive groupings (Denny et al. 2010). Phecodes allow for systematic phenotyping across a large number of individuals for numerous clinical phenotypes and provide a level of consistency when collaborating across multiple institutions. Additionally, the phecode mapping provides a list of control exclusion phecodes which typically excludes similar or related phecodes to the case phecode. Using both ICD-9 and ICD-10 codes, we constructed 1,866 phecodes using a previously defined ICD-phecode-mapping (Phecode Map 1.2) [83] resulting in a binary phenotype where a patient is a case if the specific phecode occurs at least once within their medical record. Controls are defined as individuals without the occurrence of the case phecode. An additional stricter definition of controls also restricts individuals with the occurrence of any phecode from the case phecode's control exclusion list.

Out of all individuals in ATLAS (N=27,946), over 99% of individuals have at least one phecode and 30.8% have over 100 distinct phecodes. The distribution of phecodes varies across different demographic groups in ATLAS (Figure 4.3). Older patients tended to have more phecodes; individuals under the age of 18 had an average of 57.38 (SD: 49.80) unique phecodes and individuals over the age of 64 had an average of 109.98 (SD: 70.34) unique phecodes. We limited subsequent genetic analyses to phecodes with $>$ 100 cases in ATLAS, resulting in a total of 1,330 phecodes.

To further demonstrate the potential of the EHR-derived phecodes connected with genetic data, we focus on a set of 7 traits to illustrate downstream genetic analyses: asthma, chronic obstructive pulmonary disease (COPD), gout, heart failure (HF), idiopathic pulmonary fibrosis (IPF), cerebral artery occlusion with cerebral infarction (stroke), and venous thromboembolism (VTE). As shown in Figure 4.3, the prevalence of certain phecodes varies across sex, age, and genetic ancestry. For example, gout is observed at a much higher frequency in males compared to females (76.4% cases) and tends to be diagnosed in individuals over the age of 64 (59.8% cases). We also observe a high proportion of cases of heart failure within the African ancestry group (freq(all-ATLAS)=0.044, freq(AFR-ATLAS)=0.079; p-value=$2.4 \times 10^{-6}$) and cases of gout within the East Asian ancestry group (freq(all-ATLAS)=0.048, freq(EAS-ATLAS)=0.066; p-value=$8.0 \times 10^{-4}$) compared to the prevalence across all ATLAS individuals.

### 4.2.4 Genome-wide association studies across 7 traits and 4 ancestry groups

As an example of the utility of ancestrally diverse genetic data linked with EHR-based phenotypes (phecodes), we perform GWAS for 7 well-studied traits within each of the 4 continental ancestry groups in ATLAS. The traits represent a wide variety of diseases: asthma, chronic obstructive pulmonary disease (COPD), gout, heart failure (HF), idiopathic pulmonary fibrosis (IPF), cerebral artery occlusion with cerebral infarction (stroke), and venous thromboembolism (VTE) (see 'EHR-based phenotyping through the phecode system').

We performed association testing using SAIGE [327], a generalized mixed-model approach that accounts for unbalanced case-control ratios as well as infers and accounts for sample relatedness. Given that many disease phenotypes suffer from case-control imbalance, such as gout (N-case=810, N-control=15,831) and idiopathic pulmonary fibrosis (N-case=700, N-control: 15,941) within the European ancestry group in ATLAS, SAIGE is an advantageous inference method for association testing in ATLAS. In this work, we computed association statistics across 7.3 million SNPs and 27,190 unrelated individuals for 7 traits. Association studies for all 7 traits were performed separately within each of the 4 continental ancestry groups using SAIGE (Table 4.1) for a total of 28 analyses. Self-identified sex (as reported in the EHR) and current age (as of September 2021) were used as covariates, as well as age*age and age$*$sex interaction terms. We additionally used the first 10 principal components that were re-computed only on individuals from each ancestry group. Overall, GWAS associations are well-calibrated and do not exhibit strong evidence of test statistic inflation (average across all 28 analyses $\lambda_{GC} = 0.98$, SD($\lambda_G C$)=0.01) (Figure 4.2C). We found 26 genome-wide significant SNPs (p-value $< 5 \times 10^{-8}$) within the European ancestry group (gout, heart failure, venous thromboembolism), 1 within the African ancestry group (asthma), and 8 within the Admixed American ancestry group (gout, stroke), for a total of 35 significant SNPs (Figure 4.4A).

We next compared the associated regions identified in ATLAS to those reported in previous studies, specifically those listed in the GWAS Catalog [53] and the meta-analyses performed through the Global Biobank Meta-analysis Initiative (GBMI)[326]. To avoid biasing our results, we used the GBMI summary statistics that were computed across all other contributing biobanks but omitted ATLAS data from the meta-analysis computation. To construct regions comparable across all of the studies for a given trait, we performed the following procedure. First, we aggregated all SNPs that reached genome-wide significance in at least one of the datasets (i.e. ATLAS, GBMI summary statistics, GWAS Catalog). We then performed a greedy approach by selecting the most significant SNP and created a 1Mb window (500Kb on each side) around this top SNP. All other genome-wide significant SNPs within this window were removed from the list and this procedure was performed until all significant SNPs are accounted for within a region. We defined an individual GWAS for a trait as having a significantly associated region if at least one genome-wide significant SNP fell into one of the constructed regions. Using this process, we found a total of 10 significantly associated regions in ATLAS across the 28 GWAS analyses. Out of these 10 regions, 7 were also reported both in the GWAS Catalog as well as in the GBMI meta-analysis (Figure 4.4B). Finally, when comparing the separate analyses for the 7 traits across the 4 genetic ancestry groups in ATLAS, we did not find any significant associations (SNPs or regions) occurring in multiple populations. This observation could be due to the current limited sample sizes or potentially different genetic architectures across ancestries.

In addition, we identified an association for gout on chromosome 1 exclusively within the AMR group. This association has not been previously identified in any previous gout association study. We replicated this association within the AMR group in a subsequent version of the ATLAS data with an increased sample size (N=40K individuals). A PheWAS within ATLAS at this SNP reveals associations with the“Gout” and “Gout and other crystal arthropathies” phenotypes exclusively within the AMR population as well. This provides evidence of potential differences in genetic architecture between populations for gout risk.

To further assess the congruence of genetic effects estimated in ATLAS to those from more mature EHR-linked biobanks with larger sample sizes, we compared GWAS effect sizes for the 7 traits between ATLAS and BioVU [252] within the European ancestry group. Considering nominally significant SNPs associated with each trait with p-value ¡ $1 \times 10^{-6}$ in either study, we find a strong, significant positive correlation (Pearson correlation = 0.92, p-value$< 2.2 \times 10^{-16}$) between effect sizes in BioVU and ATLAS (Figure 4.4C). Although association statistics for the BioVU study were computed using PLINK 2.0 [64] and association statistics for ATLAS were computed using SAIGE, it is encouraging that we observe a positive correlation despite the differences in association testing methods. As shown in Figure 4.4C, we see that the effects in ATLAS are slightly depressed towards the null, though this may reflect smaller sample sizes in ATLAS compared to BioVU.

### 4.2.5 Phenome-wide association studies

EHR-linked biobanks also offer the opportunity to contextualize putative associations within the clinical phenome through phenome-wide association studies (PheWAS) [85] as well as provide a valuable step for validating phenotype quality control. ATLAS has an extensive and diverse set of clinical phenotypes from non-ascertained cohorts which is critical for performing unbiased association tests. We limited our analyses to phecodes with >100 cases within ATLAS, resulting in a total of 1,330 phecodes describing the clinical phenome at UCLA. To demonstrate the utility of this diverse set of clinical phenotypes, we performed a PheWAS at rs6025, a missense variant within the F5 gene identified from the ATLAS GWAS of venous thromboembolism in the European ancestry group and also documented in many previous studies [119, 276, 174]. We performed an association between rs6025 and 1,330 phecodes and found phenotypic associations with 'iatrogenic pulmonary embolism and infarction' and 'other venous embolism and thrombosis'. Because embolisms generally form in the veins of the leg and then travel to the lungs where they can potentially cause infarctions, these associated phenotypes are consistent with the current understanding of the pathophysiology of venous thromboembolism and pulmonary embolisms [263]. This demonstrates that despite modest sample sizes across many of the phenotypes, we can recapitulate findings consistent with expected disease biology, making PheWAS a valuable tool in investigating the shared genetic architecture across clinical traits. We provide a web browser containing the PheWAS associations from ATLAS as a resource to the public (https://atlas-phewas.mednet.ucla.edu/).

### 4.2.6 Biobank contributions

The ancestral diversity represented in ATLAS plays a key role in the expansion of the catalog of genetic variation used in precision medicine efforts such as polygenic risk scores. Despite its nascency, ATLAS has already contributed to many multi-ancestry disease mapping initiatives, such as the Global Biobank Meta-analysis Initiative (GBMI) [326] and COVID-19 Host Genetics Initiative [78]. Although ATLAS constitutes approximately 1% of the total sample size for the GBMI meta-analysis (N=27,946 samples out of approximately 2.6 million total GBMI samples), we observe a large contribution of samples from diverse ancestral populations within ATLAS to GBMI. For example, ATLAS contributes larger proportions of the African (range of proportions across 7 traits: 3% - 14%) and Admixed American ancestry (22% - 32%) samples when compared to the total sample size in GBMI (Table 4.2). Additionally, for several phenotypes, ATLAS represents even larger proportions of the total African (AFR) and Admixed American (AMR) ancestry-specific case numbers (e.g., idiopathic pulmonary fibrosis, gout, and heart failure in both AFR and AMR). In addition to GBMI, ATLAS accounted for 73.4% of the Admixed American samples utilized in the primary analysis from the COVID-19 Host Genetics Initiative. This enrichment of AFR and AMR samples from ATLAS can facilitate meta-analytic disease mapping in these historically underrepresented populations and expand the genetic understanding of diverse ancestries.

In the future, we aim to perform phenotyping composed of EHR elements in addition to diagnosis codes, such as laboratory values, medications, and clinical notes. We also plan to incorporate additional types of genomic information such as exome sequencing and methylation data. As sample sizes continue to grow, the ATLAS Community Health Initiative will enable rigorous genetic and epidemiological studies, with the specific aim to accelerate genomic medicine in diverse populations.

## 4.3 Discussion

The ATLAS biobank provides a valuable resource for the biomedical community with numerous future opportunities. In the future, we aim to perform phenotyping composed of EHR elements in addition to diagnosis codes, such as laboratory values, medications, and clinical notes. We also plan to incorporate additional types of genomic information such as exome sequencing and methylation data. Furthermore, although this analysis focused on describing only common variants, we plan to investigate the rare variants in ATLAS as sample sizes continue to grow. We hope that the inclusion of rare variants in both genome- and phenome-wide association studies can increase our power to detect novel associations as well as explore more ancestry-specific effects. We hope to also leverage the typed ClinVar variants to examine the role of genetic ancestry in pathogenic and likely pathogenic variants. Additionally, we plan to create a catalog of polygenic risk score (PRS) weights for EHR-derived phenotypes across each genetic ancestry group, creating one of the largest and most ancestrally diverse PRS resources.

## 4.4 Tables

|  |  | ATLAS | Asthma | COPD | Gout | HF | IPF | Stroke | VTE |
|---|---|---|---|---|---|---|---|---|---|
| Sample size |  | 27,946 | 4702 | 2927 | 1342 | 2212 | 1139 | 1402 | 2543 |
| Age (years) |  | 55.6 | 55.8 | 67.1 | 66.3 | 66.3 | 65.2 | 66.5 | 60.6 |
|  |  | (17.2) | (17.5) | (14.1) | (13.8) | (15.8) | (13.6) | (15.0) | (16.3) |
| Self-reported sex | Male | 45.1% | 37.8% | 52.3% | 77.2% | 59.4% | 46.3% | 52.1% | 53.3% |
|  | Female | 54.9% | 62.2% | 47.7% | 22.8% | 40.6% | 53.7% | 48.0% | 46.7% |
|  | ‘Other’, ‘Unknown’, *Unspecified, ‘X’ | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| Inferred biological sex | Male | 45.5% | 37.7% | 51.7% | 76.2% | 58.4% | 46.0% | 51.8% | 52.2% |
|  | Female | 53.9% | 60.7% | 46.5% | 22.1% | 39.7% | 51.6% | 46.3% | 45.3% |
|  | Unknown | 0.6% | 0.5% | 0.3% | 0.5% | 0.5% | 0.6% | 0.9% | 0.6% |
| Self-reported race | White | 61.4% | 64.6% | 64.2% | 55.5% | 59.0% | 61.6% | 59.9% | 61.5% |
|  | Black, African American | 4.8% | 6.3% | 6.5% | 8.5% | 8.3% | 7.0% | 7.2% | 8.0% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Asian, Asian Indian, Chinese, Filipino, Indonesian, Japanese, Korean, Pakistani, Thai, Pakistani, Taiwanese, Vietnamese, Asian-Other | 9.7% | 7.7% | 7.7% | 11.8% | 7.6% | 8.6% | 8.2% | 6.5% |
| | American Indian, Alaska Native | 0.3% | 0.4% | 0.3% | 0.3% | 0.3% | 0.6% | 0.6% | 0.3% |
| | Native Hawaiian, Guamian or Chamorro, Samoan, Other Pacific Islander | 0.3% | 0.5% | 0.2% | 0.5% | 0.5% | 0.6% | 0.4% | 0.4% |
| | Other race | 12.9% | 10.9% | 9.8% | 10.5% | 13.8% | 12.7% | 12.2% | 14.6% |
| | Unknown, Declined to Specify | 10.6% | 2.3% | 1.1% | 1.4% | 0.8% | 0.9% | 1.6% | 0.9% |
| Self-reported ethnicity | Non-Hispanic/Latino | 75.4% | 76.7% | 77.7% | 76.7% | 72.6% | 75.5% | 73.9% | 72.3% |
| | Hispanic/Latino, Cuban, Hispanic/Spanish origin, Mexican, Mexican American, Chicano/a, Puerto Rican | 14.1% | 13.7% | 10.7% | 10.3% | 17.1% | 15.6% | 14.7% | 18.9% |
| | Unknown, Declined to Specify | 10.5% | 2.3% | 1.2% | 1.5% | 0.6% | 0.8% | 1.6% | 1.0% |
| Inferred genetic ancestry | European continental ancestry | 64.5% | 63.4% | 67.4% | 59.6% | 57.7% | 61.0% | 60.4% | 57.7% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | African continental ancestry | 4.8% | 6.5% | 6.7% | 8.5% | 8.1% | 7.1% | 7.4% | 8.2% |
| | Admixed American continental ancestry | 17.8% | 17.5% | 13.9% | 14.1% | 20.6% | 19.0% | 19.3% | 23.0% |
| | East Asian continental ancestry | 8.9% | 7.1% | 7.7% | 12.4% | 7.4% | 8.1% | 7.8% | 6.1% |
| | South Asian continental ancestry | 1.5% | 1.5% | 0.8% | 1.1% | 1.4% | 1.6% | 1.1% | 1.0% |
| | Admixed or other ancestry | 2.7% | 4.1% | 3.6% | 4.5% | 4.9% | 3.2% | 4.1% | 4.0% |
| Medical record length (years) | | 11.6 (8.5) | 13.12(8.5) | 13.4 (8.4) | 14.34 (8.3) | 13.2 (8.4) | 12.6 (8.1) | 13.3 (8.7) | 12.6 (8.3) |
| Number of unique ICD codes | | 86.7 (66.4) | 114.78 (76.5) | 149.4 (83.2) | 139.7 (84.1) | 164.7 (85.4) | 158.5 (81.2) | 148.1 (86.4) | 154.2 (87.4) |
| Number of phenotypes (phecodes) | Mean (standard deviation) | 85.2 (65.0) | 114.75 (76.5) | 149.44 (83.2) | 139.7 (84.1) | 164.67(85.4) | 158.5 (81.2) | 148.1 (86.4) | 154.2 (87.4) |
| | Median | 68 | 97 | 138 | 123 | 157 | 151 | 138 | 141 |

Table 4.1: Summary of UCLA ATLAS demographics. We provide summary statistics describing the UCLA ATLAS population computed from data available in the electronic health records and genotype data. Results are computed over all N = 27,946 individuals from ATLAS as well as separately within each trait.

| Trait | Abbrev. | Ancestry | UCLA-case | GBMI-case | Ratio |
|---|---|---|---|---|---|
| Asthma | Asthma | EUR | 3051 | 101311 | 1.04 |
|  |  | AFR | 289 | 5051 | 1.97 |
|  |  | AMR | 760 | 4069 | 6.42 |
|  |  | EAS | 308 | 18549 | 0.57 |
| Chronic obstructive pulmonary disease | COPD | EUR | 2005 | 51644 | 1.14 |
|  |  | AFR | 187 | 1978 | 2.77 |
|  |  | AMR | 384 | 1503 | 7.49 |
|  |  | EAS | 208 | 19044 | 0.32 |
| Gout | Gout | EUR | 810 | 20702 | 1.16 |
|  |  | AFR | 105 | 1312 | 2.38 |
|  |  | AMR | 179 | 557 | 9.55 |
|  |  | EAS | 155 | 10425 | 0.44 |
| Heart failure | HF | EUR | 1301 | 28795 | 1.51 |
|  |  | AFR | 174 | 1367 | 4.26 |
|  |  | AMR | 423 | 1170 | 12.11 |
|  |  | EAS | 144 | 12665 | 0.38 |
| Idiopathic pulmonary fibrosis | IPF | EUR | 700 | 5229 | 1 |
|  |  | AFR | 76 | 169 | 3.37 |
|  |  | AMR | 204 | 319 | 4.79 |
|  |  | EAS | 89 | 1210 | 0.55 |
| Cerebral artery occlusion with cerebral infarction | Stroke | EUR | 855 | 15842 | 2.48 |
|  |  | AFR | 100 | 1161 | 3.96 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  |  | AMR | 248 | 903 | 12.64 |
|  |  | EAS | 105 | 23345 | 0.21 |
| Venous thromboembolism | VTE | EUR | 1503 | 15970 | 1.11 |
|  |  | AFR | 195 | 1466 | 1.57 |
|  |  | AMR | 543 | 1037 | 6.18 |
|  |  | EAS | 132 | 193 | 8.07 |

Table 4.2: UCLA ATLAS contributes a substantial proportion of non-European ancestry samples to global meta-analyses. We show the case sample sizes across 7 traits for ATLAS and across the entire GBMI study, stratified by genetic ancestry. The last column reports the ratio of the proportion of ancestry-specific samples in ATLAS compared with the proportion of total samples from the GBMI meta-analyses.

## 4.5 Figures

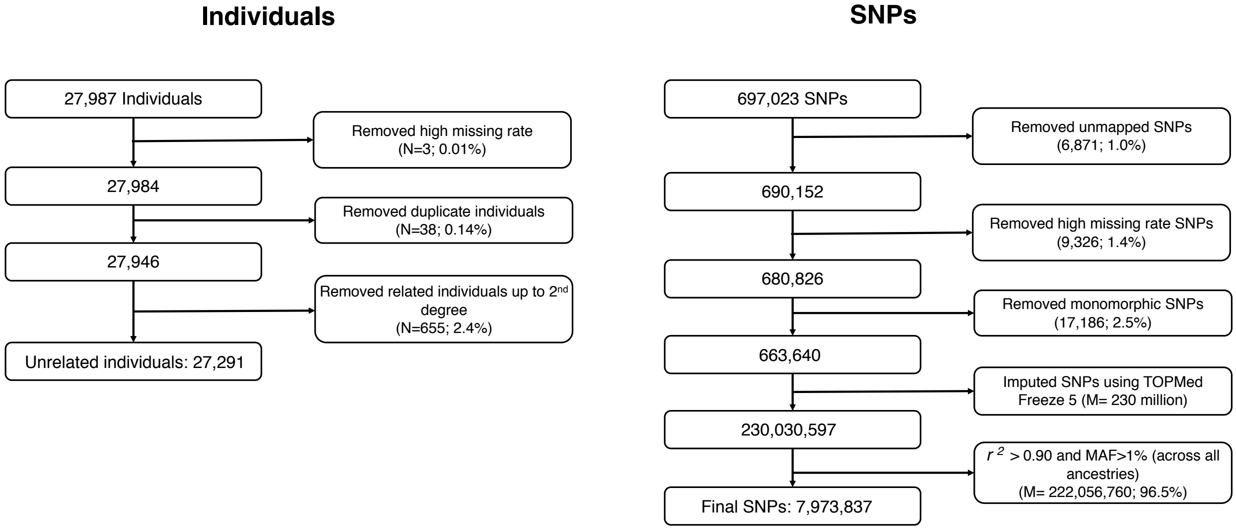
Individuals
27,987 Individuals
Removed high missing rate (N=3; 0.01%)
27,984
Removed duplicate individuals (N=38; 0.14%)
27,946
Removed related individuals up to 2nd degree (N=655; 2.4%)
Unrelated individuals: 27,291
SNPs
697,023 SNPs
Removed unmapped SNPs (6,871; 1.0%)
690,152
Removed high missing rate SNPs (9,326; 1.4%)
680,826
Removed monomorphic SNPs (17,186; 2.5%)
663,640
Imputed SNPs using TOPMed Freeze 5 (M= 230 million)
230,030,597
r 2 > 0.90 and MAF>1% (across all ancestries) (M= 222,056,760; 96.5%)
Final SNPs: 7,973,837

Figure 4.1: Summary of genotype quality control pipeline. We outline the quality control pipeline for the genotype samples and list the number of excluded samples (left) and SNPs (right) at each step.

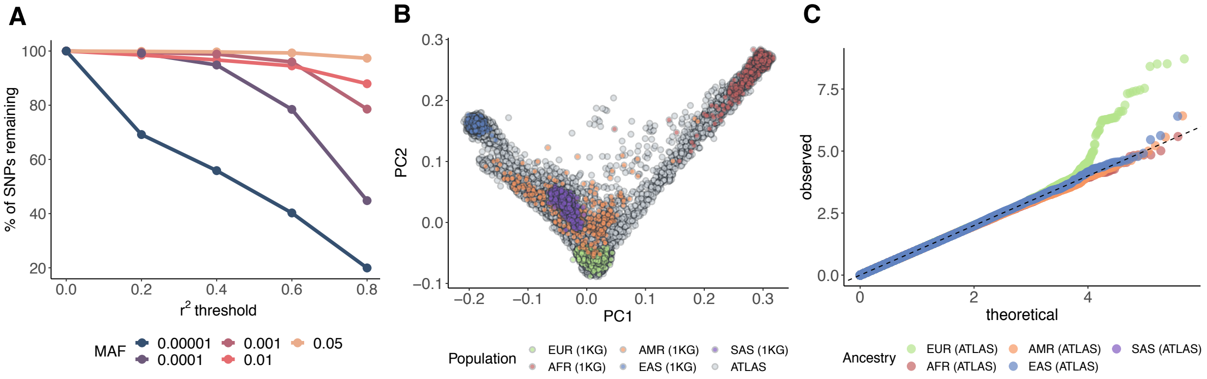

Figure 4.2: Genotyped and imputed data from ATLAS are of high quality. In A) we show the 230 million imputed SNPs stratified by minor allele frequency. SNPs are binned by the estimated imputation $r^2$ scores, and then we report the percentage of remaining SNPs after applying the $r^2$ threshold. B) shows the projected genetic PCs 1 and 2 of unrelated individuals in ATLAS (N=27,291) in gray. Samples from 1000 Genomes are shaded by continental genetic ancestry: European (EUR), African (AFR), Admixed American (AMR), East Asian (EAS), and South Asian (SAS). In C) we show the QQ-plots from the GWAS of gout across the African, Admixed American, East Asian, and European continental ancestry groups within ATLAS.

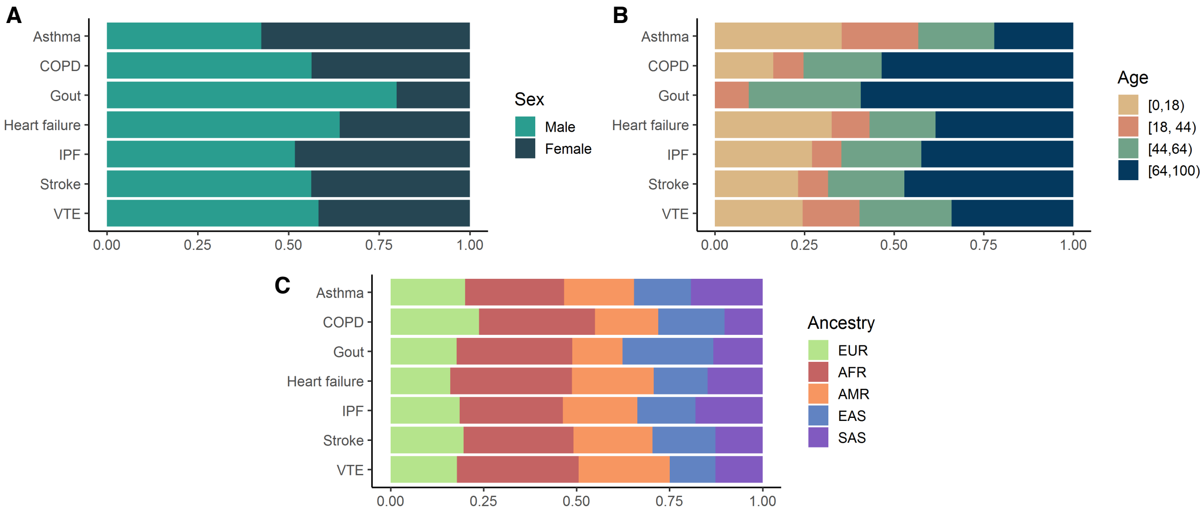

Figure 4.3: Distribution of phenotypes across different demographic groups in ATLAS. We show the distribution of 7 traits across A) sex, B) age groups and C) inferred genetic ancestry. Sex information is derived from the EHR.

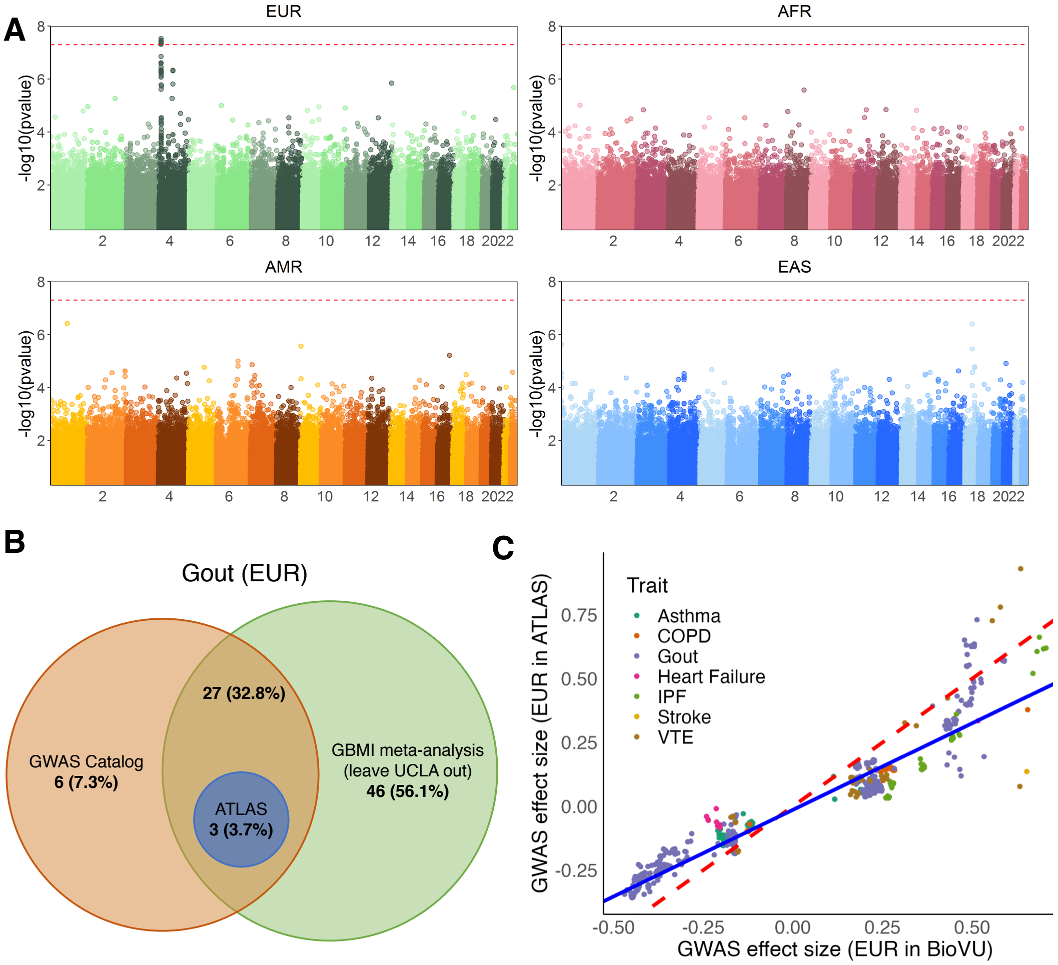

Figure 4.4: Genome-wide association studies across 7 traits and 4 continental ancestry groups recapitulate known associations. In A) we provide Manhattan plots from the GWAS of gout across the European (EUR), African (AFR), Admixed American (AMR), and East Asian (EAS) continental ancestry groups in ATLAS. The red dotted line denotes genome-wide significance (p-value $< 5 \times 10^{-8}$). In B) we show the overlap of genome-wide significant regions for gout computed from ATLAS within the European ancestry group, previous associations listed in the GWAS Catalog, and associations identified in the GBMI meta-analysis. C) shows a scatterplot of GWAS effect sizes of SNPs associated with each trait in either ATLAS or BioVU at p-value $< 1 \times 10^{-6}$. Points are colored by trait. The red line shows the 45-degree line through the origin, and the blue line shows the estimated trend for these points (Pearson correlation=0.92).

# CHAPTER 5

# Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative

## 5.1 Background

Linking electronic health records (EHRs) to patient genomic data within biobanks in a de-identified fashion has the potential to significantly advance genomic discoveries and precision medicine efforts (e.g., population screening, identifying drug targets) [168, 209, 31, 9]. However, the underrepresentation of minoritized populations in biomedical research [283, 212, 239, 274, 116, 245, 98] raises concerns that advancements in precision medicine may widen disparities in access to high-quality health care [10, 195, 260]. For example, European-ancestry individuals constitute approximately 16% of the global population, yet account for almost 80% of all genome-wide association study (GWAS) participants [195]. As a direct result of this imbalance, existing methods to predict disease risk from genetics (e.g., polygenic risk scores) are vastly inaccurate in individuals of non-European ancestry[195, 80] thus forming a barrier to advancing genomic medicine to benefit patients of all ancestries The University of California, Los Angeles (UCLA) Health medical system is located in Los Angeles, one of the most ethnically diverse cities in the world. There is no ethnic majority: 48.5% of Los Angeles residents self-identify as Hispanic or Latino, 11.6% as Asian, and 8.9% as Black or African American; additionally, 37% of Los Angeles residents are neither U.S. nationals, nor U.S. citizens at birth [298]. Therefore, the UCLA Health patient population and the availability of digital health data captured in EHRs from a single medical system presents a unique opportunity to increase the inclusion of underrepresented minorities in biomedical research. In this study, we investigate the role of genetic ancestry in a disease context within the UCLA ATLAS Community Health Initiative (or ATLAS for brevity), a biobank embedded within the UCLA Health medical system composed of de-identified, EHR-linked genomic data from a diverse patient population [134, 154]. The current initiative aims to collect genomic data from over 150,000 individuals; currently, this consists of N=36,736 individuals genotyped at M=667,191 SNPs genome-wide using the Illumina global screening array (GSA) [2] and then imputed to ¿8 million SNPs using a multi-ancestry imputation panel [285]. A detailed description describing the recruitment, consent process, sample collection, and genotype and phenotype quality control are discussed in prior works [134, 155, 213].

The EHR contains a de-identified extract of medical records (billing codes, laboratory values, etc.) as well as demographic information such as self-identified race and ethnicity information. It is important to note that self-identified race and ethnicity (SIRE) represent social constructs that capture shared values, cultural norms, and behaviors of subgroups [46] are distinct concepts from genetic ancestry which refer to the ancestral history of one's genome. This difference is even more relevant for individuals self-describing as multi-racial (and/or admixed) where genetic ancestry bears little correlation to SIRE [306, 47]. Understanding the interplay of genetic factors (such as genetic ancestry) with social determinants of health (as inferred from self-reports) is still mired in the confounding overlaps between race, socioeconomic status, and disease, but serves as a critical step in mapping and predicting disease risk across individuals of all ancestries.

In this work, we leverage the unique genomic diversity of our single-center cohort to explore the interconnected effects of self-identified race/ethnicity and genetic ancestry on clinical phenotypes. We cluster individuals by genetically inferred ancestry (GIA) within the EHR-linked biobank, systematically construct phenotypes from EHR, and compute disease associations using multi-ancestry pipelines for both genome-wide and phenome-wide association studies (PheWAS). We find that genetically-derived and self-identified information yield distinct subpopulations, emphasizing the distinction between GIA and SIRE. We leverage genetic and self-identified data to find extensive variation of subcontinental ancestry within ATLAS across European American (EA), East Asian American (EAA), Hispanic Latino American (HL), and African American (AA) GIA groups. For example, we find clusters of individuals with recent inferred ancestry from Filipino, Chinese, Japanese, and Korean ancestries among the EAA cluster. Such subcontinental clusters also stratify individuals according to disease groups thus emphasizing their utility in biomedical research. We perform both ancestry-specific GWAS and meta-analyses across GIA groups and recapitulate known genomic risk regions. We perform PheWAS on significant regions and describe genetic associations for liver-related phenotypes in multiple ancestry groups as well as associations with neurological and neoplastic phenotypes that are associated exclusively in the HL GIA group. These results underscore how the utility of large-scale genetic analyses and deep phenotyping in diverse populations can make substantial medical contributions to population health.

## 5.2 Methods

### 5.2.1 Study population

The UCLA Health System includes two hospitals (520 and 281 inpatient beds) and 210 primary and specialty outpatient locations predominantly located in Los Angeles County. The UCLA Data Discovery Repository (DDR) contains de-identified patient EHRs that have been collected since March 2nd, 2013, under the auspices of the UCLA Health Office of Health Informatics Analytics and the UCLA Institute of Precision Health. Currently, the DDR contains longitudinal records for more than 1.5 million patients (inpatient and outpatient), including basic patient information (height, weight, gender), diagnosis codes, laboratory tests, medications, prescriptions, hospital admissions, and procedures. The UCLA ATLAS Community Health Initiative includes the EHR-linked biobank within the UCLA Health System. Currently, there are more than 37,000 genotyped participants with their de-identified EHR linked through the DDR. participation is voluntary and privacy is protected by de-identifying the samples. Additional information regarding recruitment, consent, sample processing, and quality control pipelines can be found in previous work [134, 213]. Patient Recruitment and Sample Collection for Precision Health Activities at UCLA is an approved study by the UCLA Institutional Review Board IRB#17-001013 (UCLA IRB).

#### 5.2.1.1 Self-identified demographic information

Self-identified demographic information is collected as a part of clinical care which is then translated to the EHR. Participants self-identify race and ethnicity via two distinct drop-down fields where there are pre-determined multiple-choice fields for race and ethnicity. At this time, only one selection from each category can be chosen as a patient's primary race and ethnicity [3]. We group together race/ethnicity pairings to form 'self-identified race/ethnicity' (SIRE) groupings. Patients also report their 'Preferred Language' from pre-determined multiple-choice fields within the EHR. See the section "Notes on terminology and naming conventions" for a more detailed discussion of terminology used for SIREs.

#### 5.2.1.2 Notes on terminology and naming conventions

In this section, we explicitly discuss the origin of the terminology and naming conventions used throughout this manuscript with respect to genetic ancestry, race, and ethnicity. We refer to Peterson et al. [229] for a more comprehensive description of terms for GWAS in ancestrally diverse populations.

The term 'genetic ancestry' refers to the characterization of the population(s) from which an individual is descended and describes the genetic relationship implied by shared, large segments of genomic DNA between an individual and these ancestors [196]. Throughout this work, we reserve this term to describe individuals with information about the origin of their recent biological ancestors. For instance, we treat populations represented in genetic reference panels (e.g. 1000 Genomes Project [7]) as instances of genetic ancestry since the information describing the origin of the recent biological ancestors represented in the samples is known.

Much of this work involves inferring the genetic ancestry information for a set of individuals. We introduce the term ‘genetically inferred ancestry (GIA)’ to describe the genetic characterization of individuals within a group who likely share recent biological ancestors as inferred by a method of choice. We emphasize that GIA differs from genetic ancestry in that GIA depends on the inference method (e.g. clustering) and choice of reference data (e.g. 1000 Genomes).

The terms “Native American genetic ancestry” and “Native American GIA” refer to ancestry and/or recent biological ancestors from individuals originating from indigenous groups originally from North America, Central American, and South America. The term “Native American race” refers to the definition used by the US Census, “ a person having origins in any of the original peoples of North and South America (including Central America) and who maintains tribal affiliation or community attachment” [301]. We recognize that individuals in this group may prefer other terms such as “American Indians”. To be clear, the identification of subjects as Native American GIA is not meant to imply a tribal status.

In the context of this work, the term “African genetic ancestry” describes individuals whose recent biological ancestors originated from the continent of Africa. “African American (AA) GIA” refers to an admixed group of individuals within the United States who have recent biological ancestors inferred to be of African ancestry and thus have partial or total ancestry originating from Africa. The term “Admixed American ancestry” refers to those with recent biological ancestors from European, African, and Native American ancestries that achieved admixture in North America, Central America, and South America. Thus, Admixed American ancestry contains global proportions of all three ancestry groups. “Hispanic Latino American (HL) GIA” refers to the group of admixed individuals within the United States whose recent biological ancestors were inferred to be individuals of Admixed American ancestry. “European ancestry” refers to individuals with recent biological ancestors with origins in continental Europe. “European American (EA) GIA” refers to individuals within the United States with recent biological ancestors inferred to be of European ancestry. Thus, partial or total ancestry originating from Europe. “East Asian ancestry” and “South Asian ancestry” refers to individuals with recent biological ancestors from East Asia and South Asia respectively. “East Asian American (EAA) GIA” and “South Asian American (SAA) GIA” refers to individuals within the United States with recent biological ancestors inferred to be of East Asian ancestry or South Asian ancestry.

We disapprove that the term “White/Caucasian” is a preset multiple-choice option under the race field within the medical records and renounce its usage due to its erroneous origins and historically racist implications. We strongly discourage the connection of the term ‘Caucasian’ with the discussion of race, a social construct separate from biology, and emphasize that the term does not have biological implications[273]. For subsequent analyses presented in this work, we use “White” to refer to the “White/Caucasian” category. Although this terminology is still built into the language of many documents and surveys, such as EHRs, we make this change to avoid perpetuating its usage within the field.

#### 5.2.1.3 Basic genotype quality control.

Bio-samples collected from the UCLA ATLAS Community Health Initiative in the form of blood samples, were de-identified and then processed for DNA extraction and genotyping. We utilized a “frozen snapshot” of ATLAS data composed of all samples processed up to 6/18/2021. ATLAS participants (N=36,779) were genotyped using a custom genotyping array constructed from the Global Screening Array with the multi-disease drop-in panel [2] under the GRCh38 assembly. Overall, the array measured 700,079 sites for capturing single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels).

We filtered out poor-quality markers by removing unmapped SNPs, SNPs with ¿5% missingness, and strand-ambiguous SNPs (M= 19,313 variants removed). We excluded samples with missingness ¿5% (N=1 individual removed). We identified duplicate individuals (or identical twins/triplets, etc.) using KING 2.2.2 [190] (‘–duplicate’) and removed the individual with the lowest missing rate from each pair (N=42 individuals removed). All quality control steps were conducted using PLINK 1.9. Following sample- and variant-level quality control, M=667,191 genotyped SNPs remained across N=36,736 individuals for downstream analyses. All subsequent genetic analyses in this paper utilize this QC'd set of genotypes. Additional steps of QC may be conducted before running specific analyses, as described in more detail below. We refer to our previous work for a more thorough description of the quality control pipelines constructed for ATLAS[134].

#### 5.2.1.4 Genotype imputation

After performing array-level genotype quality control, the PLINK-formatted genotypes were converted to VCF format and uploaded to the Michigan Imputation Server [81]. On a variant level, the server removed the variant if it was not an A, C, G, T allele, monomorphic, a duplicate, an allele mismatch between the reference panel and provided data, an insertion-deletion, or if the SNP call rate is less than 90%. The server will additionally correct for any necessary strand flips or allele switches needed to match the reference panel. The server additionally phases the data using Eagle v2.4 [181] and imputation is performed against the TOPMed Freeze5 imputation panel [285] using minimac4 [103]. In summary, the explicit parameters used on the server are "TOPMed Freeze5" for the reference panel, "GRCh38/hg38" for the array build, "off" for the rsq filter, "Eagle v2.4" for phasing, no QC frequency check, and "quality control & imputation" for the mode. After we filtered by R2>0.90 and MAF>1%, the final set of variants contained M=8,048,268 sites.

#### 5.2.1.5 Genetic relatedness inference

We computed pairwise kinship coefficients to determine family relationships using King 2.2.2. [190]. We performed inference on the set of genotype data passing quality control (see "Basic genotype quality control") for a total of N=36,736 individuals and M=667,191 SNPs. We identified a set of unrelated individuals (N=35,761) up to degree 2 where individuals with kinship coefficient ¡0.0884 were included ('king –unrelated –degree 2'). This level of relatedness is expected since members of the same family will often be within the same healthcare system.

#### 5.2.1.6 Continental genetic inferred ancestry

We estimated GIA membership using a 2-step clustering procedure. First, we performed prinicpal component analysis (PCA)[136] on all individuals in ATLAS (N=36,736) and samples from 1000 Genomes. Specifically, we first filtered genotypes from ATLAS by Mendel error rate (`plink --me 1 1 {set-me-missing`), founders (`--filter-founders`), minor allele frequency (`{maf 0.15`), genotype missing call rate (`--geno 0.05`), and Hardy-Weinberg equilibrium test p-value (`{hwe 0.001`). The filtered genotypes from ATLAS are then merged with the 1000 Genomes phase 3 dataset. We align reference alleles between the two sets of data and filter out SNPs that are not an A, C, T, or G allele. Next, a 2-step LD pruning is performed on the merged dataset: 1) `--indep 200 5 1.15`, 2) `--indep-pairwise 100 5 0.1`. All filtering steps and LD pruning were performed using PLINK 1.9 [64]. This resulted in a total of 22,589 SNPs across 36,736 individuals in ATLAS. We computed the first 10 principal components using the FlashPCA 2.0 software [8] with all default settings.

For the second step, we perform clustering on the principal components to estimate GIA cluster membership for each individual in ATLAS. We use the K-nearest neighbors (KNN) algorithm where we use the superpopulation name of the samples in 1000 Genomes to define the cluster labels. The superpopulations form 5 clusters: European, African, Admixed American, East Asian, and South Asian genetic ancestry. For each ancestry cluster, we run KNN on the pair of PCs that capture the most variation for each genetic ancestry group: European, East Asian, and African ancestry groups utilize PCs 1 and 2, the Admixed American group uses PCs 2 and 3, and the South Asian group use PCs 4 and 5. For each ancestry group inference, we run KNN separately. In each analysis, we use 10-fold cross-validation to select the 'k' hyper-parameter from k=5, 10, 15, 20. If a sample from ATLAS had >0.50 cluster membership, then the sample is reported as the genetic inferred ancestry represented in that cluster (European genetic ancestry (GA) → European GIA, African GA → African American GIA, Admixed American GA → Hispanic Latino American GIA, East Asian GA → East Asian American GIA, South Asian GA → South Asian American GIA). See "Notes on terminology and naming conventions" for a more in-depth discussion about the naming of GIA clusters. Individuals who did not attain >0.50 membership in any cluster or were matched to multiple clusters were reported as being 'Ambiguous GIA'.

#### 5.2.1.7 Subcontinental genetic inferred ancestry

We estimate subcontinental GIA membership for individuals within the East Asian American GIA group using a 2-step clustering procedure similar to the continental GIA clustering discussed in a prior section ("Continental genetic inferred ancestry"). First, we perform PCA on all individuals in the EAA GIA group in ATLAS (N=3,331) and samples from the East Asian ancestry population in 1000 Genomes. Using only the genotyped SNPs, we perform the same filtering steps as described above, namely filtering ATLAS genotypes by Mendel error rate, founders, MAF > 0.15, genotype missing call rate, Hardy-Weinberg equilibrium test, and LD pruning. Following sample- and variant-level quality control, M=36,504 SNPs remained. We also found that not restricting to only unrelated individuals does not bias our estimates. We then compute the first 10 principal components using FlashPCA with all default settings.

For the second step, we perform clustering on the principal components to estimate subcontinental GIA cluster membership for each individual in the East Asian American GIA group in ATLAS. We use the K-nearest neighbors (KNN) algorithm where we use the population name of the East Asian ancestry samples in 1000 Genomes to define the cluster labels. The populations form 5 clusters: Han Chinese, Southern Han Chinese, Dai Chinese, Japanese, Kinh Vietnamese genetic ancestry. We run KNN using PCs 1-4 with 10-fold cross-validation to select the 'k' hyper-parameter from k=5, 10, 15, 20. If a sample from ATLAS had ¿0.90 cluster membership, then the sample is reported as the genetically inferred ancestry represented in that cluster. Individuals who did not attain ¿0.90 membership in any cluster were reported as being 'Ambiguous EAA.

Alternatively, we can define GIA clusters using self-identified information from the samples in ATLAS. We perform a similar approach as above, except we use ATLAS individuals' self-identified race as the labels to define the clusters. We limit cluster definitions to self-identified race groups with N¿20 for a total of 7 clusters: Chinese, Filipino, Japanese, Korean, Taiwanese, Thai, and Vietnamese. Although we do not utilize label information from 1000 Genomes, we still use the PCs computed on the merged ATLAS and 1000 Genomes datasets to keep PCA projections consistent across the 1000 Genomes-based and self-identified race-based clustering methods. We run KNN using the same procedure and thresholds as above. Again, individuals who did not attain ¿0.90 membership in any cluster were reported as being 'Ambiguous EAA'. Explicit clusters could not be confidently computed for other continental GIA groups.

### 5.2.2 Genetic admixture analysis

We inferred the proportion of genetic ancestry using the ADMIXTURE software [12] under the unsupervised clustering mode with the number of clusters k=4, 5, 6. Specifically, we restrict to only SNPs with only an A, C, G, T allele and with MAF > 0.05 (`--maf 0.05 --snps-only 'just-acgt'`) within ATLAS. We then merge the data from ATLAS with the 1000 Genomes phase 3 data set and limit inference to only the subset of the overlapping SNPs. We then perform LD pruning every 2Kb on the merged data set (`--bp-space 2000`). All filtering steps and LD pruning were performed using PLINK. This resulted in a total of 223,095 SNPs across 36,736 individuals in ATLAS which was then used for ancestry inference using ADMIXTURE.

Finally, we performed the admixture analysis with `./admixture atlas_1kg_bed_file k` with k equal to 4, 5 or 6. We compare the ancestry proportions from each SIRE to estimate the ancestry represented in each mixture component. For k=4, we label the component with the majority of NH-White individuals as European ancestry, the component with the majority of NH-AfAm individuals as African ancestry, the component with the majority of NH-Asian individuals as East Asian ancestry, and the component with the highest number of HL-Other and HL-White individuals as Native American ancestry.

### 5.2.3 Phecodes

Billing codes documented in the medical record were used to generate phenotypes for analysis. The previously described phecode ontology (v1.2) maps the specific ICD-9 and ICD-10 codes from each patient's chart onto a group of ¿1,800 more general and clinically meaningful phenotype terms [85]. Mapping completed with the PheWAS R package[59] (https://github.com/PheWAS/PheWAS) creates binary phenotypes. Patients with one or more instances of a phecode were considered cases while patients without any instance of the corresponding phecode were considered controls. We limited analyses to phecodes with at least N-Case ¿ 50 in each GIA group for a total of 1,568 phecodes meeting this threshold in the EA GIA, 802 in the AA GIA, 1,223 in the HL GIA, and 891 in the EAA GIA group.

### 5.2.4 Role of phecode occurrences for defining cases

We define a phecode occurrence as an encounter containing at least one of the ICD codes specified in the phecode definition. If a corresponding ICD code is found on another separate encounter, we treat this instance as a separate phecode occurrence. We compare two definitions of phecodes. For the first definition, we only require the presence of an ICD code attached to any type of patient encounter (i.e. laboratory tests, hospital, outpatient, medications, telehealth appointments, notes, phone calls, etc.). For the second definition, we require the presence of an ICD code attached to only encounters from appointments or office, hospital, or procedure visits. This stricter definition attempts to avoid capturing encounters that may be less indicative of a diagnosis (e.g. patient-physician telehealth messaging). We refer to these two definitions as all-encounter-derived and visit-derived phecodes. Using these two types of definitions, we then vary the number of phecode occurrences required for defining cases and compute the proportion of retained cases compared to the sample sizes if only 1 occurrence was required.

### 5.2.5 Association between phecodes and genetic ancestry

To test the differential prevalence of phecodes across genetic ancestry groups, we performed a marginal association test for each phecode to compare its prevalence in one of the genetic ancestry groups (EA, AA, HL, and EAA) with the other three groups using the following logistic regression model:

$$logit(phecode) = \beta_0 + \beta_a * GIA + \beta_2 * sex + \beta_3 * age \qquad \text{[over all ATLAS individuals]}$$

To account for the potential confounding effects of SIRE, we performed additional analyses with the model:

$$logit(phecode) = \beta_0 + \beta_a * GIA + \beta_2 * sex + \beta_3 * age + \beta_4 * SIRE$$

Statistical significance was determined after correcting for the number of tested phecodes with the Bonferroni correction procedure (p-value $< 1.12 \times 10^{-5}$). We also applied the method to the East Asian American group to test the phecode prevalence difference across subcontinental ancestry groups including Chinese, Japanese, Filipino, and Korean Americans.

### 5.2.6 Association between genetic admixture proportions and phecodes

Given the substantial variation of admixture proportion within each SIRE group, we test the association of phecode with admixture proportion (k=4) for 600 phecodes within each of the seven ATLAS SIRE groups (NH-White, NH-AfAm, HL-Other, HL-White, NH-Asian, NH-PI, NH-AmIn) with the following model:

$$logit(phecode) = \beta_0 + \beta_a * admixture_proportion + \beta_2 * sex + \beta_3 * age$$

[over individuals within a SIRE]

Each model is limited to individuals of one SIRE instead of all ATLAS individuals. Only traits with $>$ 10 cases per SIRE were tested. Significance is determined after adjusting for the number of tested phenotypes with the Bonferroni correction procedure (p-value $< 2.08 \times 10^{-5}$).

### 5.2.7 GWAS quality control per GIA

When performing GWAS, we stratified individuals by GIA groups and then performed an additional level of QC separately within each GIA group. We limited analyses to the 4 largest GIA groups: European American (N=22,380), African American (N=1,995), Hispanic Latino American (N=6,073), and East Asian American GIA (N=3,331). At this time, we omitted GWAS analyses within the South Asian American GIA group due to the limited sample size (N=625). Individuals who could not be clustered into a specific GIA group (N=2,332) were also omitted from GWAS analyses.

For GWAS, we utilized imputed data consisting of 8,048,268 SNPs across N=36,736 individuals. Within each ancestry group, samples identified as heterozygosity outliers (+/- 3 SDs from the mean) were removed and SNPs that failed the Hardy-Weinberg equilibrium test (p-value $< 1 \times 10^{-12}$) were also removed. Finally, we limited analyses to only SNPs with MAF $>$ 1% within each GIA group, yielding a total of N=22,380 individuals and M=6.0 million SNPs within the European American GIA group, N=1,995 individuals and M=5.9 million SNPs within the African American group, N=6,073 and M=6.3 million SNPs within the Hispanic Latino American group, and N=3,331 individuals and M=4.8 million SNPs within the East Asian American group.

### 5.2.8 Ancestry-specific GWAS

GWAS for all 6 traits were performed separately within each of the 4 continental ancestry groups that met the minimum N¿50 cases. Additional GWAS-specific quality control is performed within each GIA group (see GWAS quality control per GIA). Using marginal logistic regression implemented in PLINK, we computed association statistics at each imputed autosomal SNP (`plink --logistic beta`). We additionally used age, sex, and PCs 1-10 as covariates where age is defined as the individuals' current age within the EHR (as of September 2021). The values used to represent sex in this specific analysis are derived from patients' self-identified sex as reported in the EHR. Within the EHR, this specific field is labeled as 'Sex' and has a list of pre-determined multiple-choice fields where participants select one of the following options: 'Male', 'Female', 'Other', 'Unknown', '*Unspecified', 'X'. We find that 45.1% of individuals self-identify as male and 54.9% self-identify as female.

### 5.2.9 Meta-analyses

We perform meta-analyses for each trait across all GIA groups. First, we run ancestry-specific GWAS (see "Ancestry-specific GWAS") within each GIA group with an adequate sample size (N-cases¿50). We exclude analyses where very few of the SNPs produced a valid (non-NA) p-value which is likely attributed to the small sample size. The meta-analysis for skin cancer consisted of measurements from the EA and HL GIA groups; EA, AA, HL, EAA GIA groups for ischemic heart disease; EA, AA, HL, EAA GIA groups for chronic nonalcoholic liver disease; AA and HL GIA groups for uterine leiomyoma; HL and EAA GIA groups for liver/intrahepatic bile duct cancer; EA, AA, HL GIA groups for chronic kidney disease. We performed each meta-analysis using a fixed effect model as implemented in PLINK (`--meta-analysis + logscale`). Association statistics computed from the meta-analyses are reported for SNPs that occur in at least two of the GIA groups.

### 5.2.10 PheWAS

We perform a PheWAS on the top SNPs from each ancestry-specific GWAS analysis that met genome-wide significance (p-value $< 5 \times 10^{-8}$). Only phecodes with at least N-cases>50 per GIA group were considered, resulting in a total of 1,568 phecodes meeting this threshold in the EA GIA and 1,223 in the HL GIA. Analyses in the AA and EAA GIA groups were excluded since the top SNPs were not significantly associated with these groups. We additionally stratified individuals by sex for the sex-specific phecodes, which are denoted in the definition of each phecode. This resulted in a total of 24 male- and 113 female-specific phecodes within the EA GIA group, and 12 male- and 87 female-specific phecodes within the HL group after limiting to phecodes with at least N-cases $> 50$. We used individuals' self-identified sex as reported in the EHR for this analysis.

We performed an association test between the top SNP and all phecodes in a given GIA group under a logistic regression model. Age, sex, and PCs 1-10 were used as covariates in the regression where age is defined as the individuals' current age within the EHR (as of September 2021) and sex is derived from individuals' EHR. The association test is performed using the logistic regression option implemented in PLINK (`plink --logistic beta`). The PCs used in the regression analysis were re-computed using only individuals from within each respective GIA group. Phenotype significance was determined as p-value $< 0.05/(\#$ phecodes), thus each GIA group has a specific significance threshold due to the different number of tested phecodes. A more stringent threshold also accounting for genome-wide significance is also computed where p-value $< 5 \times 10^{-8}/(\#$ phecodes). Both thresholds are denoted in the PheWAS plots.

### 5.2.11 Effective sample size of associated phecodes

To assess the power of the PheWAS analysis at rs2294915 between the European American (EA) GIA and Hispanic Latino American (HL) GIA groups, we compute the effective sample size ($N_{eff}$) of each associated phecode, where the effective sample size balances the number of cases and controls when measuring the power of an analysis[314]: $N_{eff} = 2/(1/Ncases + 1/Ncontrols)$.

## 5.3 Results

### 5.3.1 ATLAS includes individuals of diverse continental ancestries

The UCLA Health patient population is diverse, with 65.36% self-identifying their race as White, 5.23% as Black or African American, 9.89% as Asian, 0.41% as Native American or

Alaska Native, 0.31% as Pacific Islander, and 18.81% identify as another race. For ethnicity, a separate concept from race and recorded under a different field in the EHR, 15.96% of individuals self-identify as Hispanic or Latino; the remaining individuals self-identify as non-Hispanic/Latino. We define genetic ancestry as the characterization of the population(s) from which an individual is biologically descended and the genetic relationship between an individual and these ancestors. When information describing the origin of individuals' recent biological ancestors is not available, we can instead infer the genetic ancestry using statistical methods. We introduce the term 'genetically inferred ancestry (GIA)' to describe the genetic characterization of individuals within a group who likely share recent biological ancestors as inferred by a method of choice. We emphasize that GIA differs from genetic ancestry in that GIA is highly dependent on the inference method (e.g. PCA, IBD) and choice of reference data. We provide a discussion about the rationale behind the terminology and naming conventions used in this work under the section "Methods: Notes on terminology and naming conventions".

Using data from the 1000 Genomes Project [7], we investigated genetically inferred ancestry in ATLAS through principal component analysis (PCA)[136, 218] and clustering techniques (see Methods). Using the five continental ancestry populations within 1000 Genomes (European, African, Admixed American, East Asian, South Asian ancestry) as a reference, we identify clusters of individuals with European American, African American, Hispanic Latino American, East Asian American, and South Asian American genetically inferred ancestry. Although we broadly find that self-identified race and ethnicity highly correlate with an individual's inferred genetic ancestry, we still observe marked differences between the two (Figure 5.1). For example, we find 10.63% of individuals within the European American GIA cluster do not identify as being within the Non-Hispanic/Latino – White (NH-White) SIRE; 13.33% of individuals within the African American GIA cluster do not self-identify as Non-Hispanic/Latino – Black/African American (NH-AfAm), and 16.58% of the Hispanic Latino American cluster do not identify as Hispanic/Latino – Other Race (HL-Other) or Hispanic/Latino – White (HL-White)). This further emphasizes that SIRE is not equivalent to GIA and that these two concepts form distinct groupings.

Further emphasizing the distinction between GIA and SIRE, we reveal extensive genetic heterogeneity both between and within SIREs, as observed from the orthogonal spectra from PCA (Figure 5.2). For example, most individuals who self-report as NH-AfAm lie along a cline between the AA and EA GIA clusters. However, 102 individuals in this SIRE cannot be clustered into either the AA or EA ancestry cluster. This is likely because many of these individuals in ATLAS self-identify as African American, which suggests genetic admixture between African and European ancestry in this group. We also find that the NH-Asian SIRE has individuals spread along all PC1 and PC2, spanning the entire space between the EAA and EA GIA clusters. However, when looking solely at GIA, we are not able to observe this pattern. Instead, most individuals in between these two clusters were inferred to have ambiguous GIA, where specifically, 221 individuals within the NH-Asian SIRE were not able to be clustered into a specific GIA group. Overall, 6.35% of individuals still have unclassifiable genetic ancestry either because they were clustered into multiple GIA groups or none at all. The latter could be due to extensive admixture in their genomes or the absence of relevant ancestral groups in the chosen reference panels.

Categorizing individuals by self-identified preferred language, we observe trends that are consistent with both SIRE and continental GIA (Figure 5.2C). For example, out of all individuals who report Spanish as their preferred language, 94.47% of these individuals were estimated to have Hispanic Latino American GIA. Additionally, 99.76% of individuals who report Japanese, Korean, Tagalog, Vietnamese, Mandarin Chinese, or Cantonese as their primary languages were inferred to have East Asian American GIA. We also observe clusters of individuals who speak Armenian, Arabic, and Farsi/Persian; we find that 47.13% of the individuals that speak these languages could not be classified into one of the five continental GIA groups. This discrepancy is likely because the 1000 Genomes reference panel does not contain samples from regions where these languages are primarily spoken. These findings showcase the limitation of current reference panels of genetic diversity and demonstrate the value of characterizing individuals using both genetic ancestry and self-identified information.

### 5.3.2 Fine-scale subcontinental ancestry within ATLAS individuals

Next, we assessed genetic ancestry at the subcontinental level. Performing PCA only on individuals from the EAA GIA group from ATLAS and the East Asian ancestry group from 1000 Genomes, we observe distinct clusters of individuals as shown in Figure 5.3A, where the cluster annotations provide a visual reference describing the approximate location and size of GIA clusters (as opposed to the formal cluster membership thresholds). Shading by the subcontinental East Asian genetic ancestry groups present in 1000 Genomes, we observe clusters corresponding to three different subgroups of Chinese ancestry (Han Chinese, Southern Han Chinese, and Dai Chinese). Additionally, we see clusters of both Japanese and Vietnamese ancestry. Using 1000 Genomes as a reference panel, we can use a K-nearest neighbors clustering approach to infer the subcontinental genetic ancestry of individuals in ATLAS where we find N=307 in the Han Chinese American GIA cluster, N=224 in the Southern Han Chinese American GIA cluster, N=483 in the Japanese American GIA cluster, and N=136 in the Vietnamese American GIA cluster (see Methods). There were not any ATLAS individuals assigned to the Dai Chinese American GIA cluster. When projecting ATLAS individuals' preferred language onto the PCs, two distinct clusters are delineated according to the Chinese Mandarin and Chinese Cantonese/Toishanese languages. The Southern Han Chinese American cluster of individuals correlates with individuals speaking Chinese Cantonese/Toishanese, where 37.50% of individuals who speak Chinese Cantonese/Toishanese are within this cluster. The Han Chinese American cluster correlates with Chinese Mandarin where 45.33% of individuals who speak Chinese Mandarin fall within this cluster.

From Figure 5.3A, there are two notable clusters that do not match any of the East Asian subcontinental ancestries represented within 1000 Genomes. Projecting ATLAS individuals' self-identified preferred languages onto the PCs shows that many of these individuals in these two clusters self-identify as speaking Korean and Tagalog. These patterns are similarly reflected by individuals' self-identified race where the majority of these individuals self-identify as Korean and Filipino. Because there is descriptive self-identified demographic information available in the EHR, we can alternatively use this to define subcontinental

GIA clusters in ATLAS. This could be advantageous since a >65.48% (N=2,181) ATLAS individuals within the EA GIA group could not be further clustered into a subcontinental GIA group derived from 1000 Genomes. Using self-identified race groups with N¿20 individuals, we repeat the same clustering process described above using individuals' self-identified race as cluster category labels. Using self-identified race information over the information available in 1000 Genomes, we are able to recover two large clusters consisting of individuals with Korean American (N=533) and Filipino American (N=761) GIA as well as identify additional clusters of individuals corresponding to Thai (N=33) and Taiwanese (N=73) GIA. This clustering not only characterizes the fine-scale genetic and ethnic diversity of ATLAS but also emphasizes how self-reported information such as primary spoken language can be combined with genetic information to identify patterns not otherwise evident.

Next, we looked at individuals with subcontinental genetic ancestry of European descent in ATLAS, but due to limitations in the 1000 Genomes reference panel, we were unable to describe the origins of the majority of the observed genetic variation within the ATLAS European American GIA cluster (Figure 5.3B). Comparing self-identified race and ethnicity information also did not delineate any subgroups since most individuals fell within the NH-White SIRE. Instead, we project individuals' preferred language onto the projected PCs. Aside from English, we observe clusters of individuals whose preferred languages are Arabic, Armenian, and Farsi/Persian. Notably, the primary populations that speak these languages are not present in the current 1000 Genomes reference panel. Although not a definitive determination of ancestral origin, these results suggest that individuals in these clusters may have cultural ties relating to the Middle East. We also observe two distinct clusters of individuals who speak Farsi/Persian, suggesting that although these groups may share cultural and/or ethnic ties, the groups could have multiple ancestral origins. However, due to limited genetic and self-identified information, we did not attempt to formally infer the subcontinental ancestry of these individuals.

We perform a similar analysis for the Hispanic Latino American cluster of individuals where we re-ran PCA only on individuals in the HL GIA cluster within ATLAS and individuals from the Admixed American population in 1000 Genomes. Projecting population labels from 1000 Genomes onto the PCs, we observe relatively sparse clusters of individuals of Mexican, Peruvian, Colombian, and Puerto Rican ancestry from 1000 Genomes (Figure 5.3C). Due to the overlapping and sparse shape of these clusters, we did not attempt to formally infer subcontinental ancestry for these individuals. Overlaying SIRE and language as previously discussed also did not reveal any additional population structure in this group. Since the HL GIA group is inherently an admixed population, we instead shade the PCs by the estimated proportions of European and Native American ancestry (see Methods). We observe a cline between European and Native American ancestries, demonstrating that although we cannot determine discrete clusters within our data, there is still substantial population structure present.

Corresponding analyses were also performed for the African American GIA group in ATLAS, but clear subcontinental clusters could not be constructed from reference panel information. Similarly, SIRE information did not delineate any clusters nor did the preferred language. Since the majority of patients self-identify as African American, an admixed population of African and European ancestry, we project the proportion of European and African ancestry onto the PCs. We observe a cline going from higher proportions of European ancestry to higher proportions of African ancestry. This suggests that for very admixed populations, it would be more advantageous to quantify population substructure continuously rather than within discrete categories. We omitted the subcontinental analysis for the South Asian American GIA group due to the small sample size (N=625).

### 5.3.3 Admixture describes genetic variation within self-identified race/ethnicity groups

As demonstrated in prior sections, many individuals do not fall within a single GIA cluster, but instead lie on the continuum between multiple ancestry groups. We can characterize this variation through genetic admixture, the exchange of genetic information across two or more populations [120]. We estimate genetic ancestry proportions using k=4, 5, or 6 ancestral populations and visualize groups of individuals by SIRE (see Methods). For the following analyses, we use k=4 ancestral populations where the clusters correspond to European, African, Native American, and East Asian ancestry. Among individuals in the NH-AfAm SIRE, the estimated average proportion of European ancestry is 24% and 73% African ancestry. We also observe that the HL-Other and HL-White SIREs have approximately the same admixture profile, where the proportion of European ancestry is 48% and 58% respectively, 6% and 5% African ancestry, and 44% and 35% Native American ancestry. This admixture profile is consistent with individuals of Mexican ancestry where there is mainly European and Native American ancestry [217]. However, there is also a large amount of variation within SIREs, where for example, individuals who identify as Hispanic or Latino ethnicity are estimated to have European ancestry percentages ranging from nearly 0% to almost 100%.

### 5.3.4 Genetic ancestry groups correlate with disease prevalence within ATLAS

Understanding how disease prevalence varies across populations is integral to understanding how the interplay of genetic factors and social determinants of health contribute to disease risk. We investigated over 1,500 EHR-derived phenotypes (phecodes) [85] from across a wide set of disease groups. We define cases as individuals having the presence of at least one occurrence of the specified phecode (see Methods). We find that varying the number of required phecode occurrences and types of encounters when defining cases does not substantially change case and control assignment in this data set. Limiting our analyses to phecodes with a minimum of 50 cases, we identify 1,512 total significant phecode-GIA associations across the 4 largest continental GIA groups after adjusting for age and sex (p-value$< 1.12 \times 10^{-5}$; Bonferroni correction for all phecodes tested across 4 GIA groups). Overall, there are 732 phenotypes that show cross-ancestry differences whose prevalence varies significantly by GIA. From this set of significant associations, the highest number of phecodes are from the circulatory (N=89), endocrine/metabolic (N=84), and digestive (N=90) system-related categories. Specifically, we recapitulate many known associations such as skin cancer (p-value=$3.13 \times 10^{-281}$) in the EA GIA group; chronic nonalcoholic liver disease in the HL GIA group (p-value=$4.83 \times 10^{-97}$); ischemic heart disease (p-value=$6.74 \times 10^{-8}$), chronic kidney disease (p-value=$1.98 \times 10^{-41}$) and uterine leiomyoma (p-value=$2.30 \times 10^{-33}$) in the AA GIA group[215, 11, 315, 139], and liver and intrahepatic bile duct cancer (p-value=$1.85 \times 10^{-38}$) within the EAA GIA group (Figure 5.4).

To further explore the implications of genetic ancestry for a range of diseases, we focus on 6 phenotypes that were found to be significantly associated with genetically inferred ancestry (GIA) in ATLAS. This set represents a wide variety of diseases: skin cancer, ischemic heart disease, chronic nonalcoholic liver disease, uterine leiomyoma, chronic kidney disease, and liver/intrahepatic bile duct cancer. Our goal was to capture variation across each GIA group: ischemic heart disease, chronic kidney disease, and uterine leiomyoma have the strongest association with the African American GIA group, skin cancer with the European GIA, chronic nonalcoholic liver disease with the Hispanic Latino American GIA, and liver/intrahepatic bile duct cancer with the East Asian American GIA group. Additionally, previous literature has already shown that the prevalence of these 6 diseases has some level of variation across racial and ethnic groups, making them ideal candidates for the further analysis of disease variation across GIA groups in ATLAS [217, 67, 270, 253, 161, 256, 110].

The GIA clusters are often correlated with SIRE, as demonstrated in previous sections. To assess whether the observed effect is primarily driven by the role of genetic ancestry, we also add individuals' SIRE as a covariate into the model. After multiple hypothesis testing (Bonferroni correction for all tested phecodes across 4 GIA groups: p-value $< 1.12 \times 10^{-5}$), we replicate 259 out of 1,512 phecode-GIA associations despite the reduced effect magnitude and association significance. Out of the 6 example traits, all but the 2 within the NH-AfAm SIRE maintained significance. This demonstrates that there is some level of disease association attributed to the ancestry component. The incorporation of SIRE should not be interpreted as a formal adjustment for environmental factors. However, SIRE could capture sociocultural and socioeconomic factors that are not explicitly modeled and/or available to use through the EHR.

We also observe substantial disease risk heterogeneity across subgroups of the same continental GIA group. We perform association tests between subcontinental GIA and phecodes within the East Asian American GIA group in ATLAS for phenotypes with N¿20 cases. To maximize sample size, we use the race-based GIA clusters (see Methods) and limit analyses to the Korean (N=552 individuals, 546 phenotypes), Japanese (N=548 individuals, 600 phenotypes), Filipino (N=844 individuals, 700 phenotypes), and Chinese (N=1217 individuals, 812 phenotypes) GIA subgroups in ATLAS. Across subgroups, we observe disease associations to varying degrees. We find 3 significant associations with subcontinental GIA and phenotypes where significance was determined after correcting for 812 tested phecodes, p-value$< 6.16 \times 10^{-5}$ (see Methods). For example, the direction of the association with chronic kidney disease varies across subcontinental GIA groups where the odds ratio for the Chinese American GIA group is 0.54 (p-value=$2.9 \times 10^{-5}$) but the odds ratio for the Filipino American GIA group is 1.83 (p-value=$2.87 \times 10^{-5}$). Additionally, the odds ratio estimated for ischemic heart disease in the Filipino American GIA subgroup is 1.81 (p=$3.33 \times 10^{-7}$), but performing the association at the continental EAA GIA level, a conclusive trend cannot be determined (OR: 0.91, p-value=$7.10 \times 10^{-2}$). These results indicate that genetically grouping individuals across subcontinental GIA groups yields meaningful interpretation of disease risk across groups of individuals that might be missed when only grouping individuals at the continental level.

We also investigated disease prevalence within admixed individuals where variation in genetic ancestry across individuals in the population allows for the correlation of disease risk with the proportion of genetic ancestry from any given continental group. Within each SIRE group, we perform an association test between the proportions of inferred ancestry estimated from ADMIXTURE [12] and each phecode. After correcting for the number of tested phecodes, we find numerous significant phecode-ancestry associations: 210 associations within the HL-Other SIRE, 133 within the NH-White SIRE, and 65 within the NH-Asian SIRE, and 16 associations within the NH-AfAm SIRE. Across SIREs, both the top associated phecode categories, as well as the direction of the associations, greatly vary. Out of the top

3 phecode categories with the most associations in each SIRE group, the most commonly shared group is the endocrine/metabolic category (HL-Other, NH-White, NH-Asian). Even within this category, looking at the statistics quantifying the association of the proportion of European ancestry with endocrine/metabolic phenotypes, there are exclusively 5 negative associations within the NH-White group, 22 negative associations within the HL-Other group (and 2 positive associations), but 5 positive associations and no negative associations within the NH-Asian group. The other top phenotype categories for each SIRE are also unique, where the HL-Other SIRE's top categories include digestive and respiratory phenotypes, the NH-White SIRE's top categories include neoplasms and dermatologic phenotypes, and the NH-Asian SIRE's top categories include psychiatric and infectious diseases. Specifically, we find that within the HL-Other population, the overall proportion of European ancestry is significantly negatively associated (p-value=$8.09 \times 10^{-10}$) with chronic nonalcoholic liver disease, and the proportion of Native American ancestry is significantly positively associated (p=$7.68 \times 10^{-9}$) (Figure 5.5), which is consistent with previous studies [237, 137]. These results suggest that not only are some disease statuses associated with SIRE and continental GIA, but the specific ancestry proportions may also correlate with disease risk.

### 5.3.5 Genome and phenome-wide association scans identify known risk regions and elucidate correlated phenotypes

EHR-linked biobanks also offer the opportunity of investigating genetic associations with traits across the genome. These efforts impose special challenges, such as adjusting for population stratification and cryptic relatedness in health systems that serve entire families as well as extracting phenotypes from EHR, namely due to inconsistencies in mapping diagnosis codes (ICD codes) to phenotypes and difficulties in defining appropriate controls for specific phenotypes. We perform GWAS on each of the 6 phenotypes within each GIA group. After filtering out analyses with small sample sizes (N<50) and analyses where most SNPs failed the regression, we have a total of 17 analyses. Overall, we find associations are well-calibrated with little evidence of test statistic inflation (median lambda-GC: 1.01). We find a total of

212 genome-wide significant SNPs (p-value $< 5 \times 10^{-8}$): 77 associations for skin cancer, 1 for ischemic heart disease, and 58 for chronic nonalcoholic liver disease in the EAA GIA group; 1 association for liver/intrahepatic bile duct cancer and 78 for nonalcoholic liver disease in the HL group; and 1 in the EAA group for heart disease. We did not find any genome-wide significant SNPs within the AA GIA group which could be due to the smaller sample size (N=1,995).

First, we observe ancestry-specific associations, such as a strong association at rs12203592 for skin cancer (p-value=$2.59 \times 10^{-32}$) in the European American (EA) GIA group. When performing the association for this phenotype in the other GIA groups, we do not have an adequate sample size to perform a successful association test at the majority of the SNPs. For the East Asian American (EAA) and the Hispanic Latino (HL) GIA groups, all SNPs resulted in a p-value estimate of NA, denoting that the association test had failed. When performing a meta-analysis between the EA and HL GIA groups, we do not find any significant associations despite the strong association originally reported in the EA GIA group. And an analysis within the African American (AA) GIA group was not performed due to the low sample size (N $<$ 50). We also see significant associations across multiple ancestry groups. For example, in the analyses for nonalcoholic liver disease, we find 58 genome-wide significant SNPs in the EA GIA analysis and 70 in the HL analysis (Figure 5.6). All genome-wide significant SNPs from both studies fall within the 22q13.31 locus, which contains the *PNPLA3* gene. This gene has been extensively studied for its role in the risk of various liver diseases such as nonalcoholic fatty liver disease [296, 308]. Interestingly, we see more associated SNPs within the Admixed American (N-case=1466) ancestry group despite the larger sample size in the European ancestry group (N-case=3177). The lead SNP from both analyses, rs2294915 (p-value(HL)=$2.32 \times 10^{-16}$, p-value(EA)=$6.73 \times 10^{-11}$), is an intronic variant in the PNPLA3 gene and has MAF=0.49 in the HL GIA but only MAF=0.24 in the EA GIA which could contribute to the heightened associations in the HL GIA.

We next perform a meta-analysis across all genetic ancestry groups under a fixed effects model for each trait for a total of 6 meta-analyses. Meta-analyses present a way to increase statistical power through increased sample size. We observe a total of 11 genome-wide significant associations: 28 for ischemic heart disease (27 new), 82 (14 new) for chronic nonalcoholic liver disease, and 1 (new) association for liver/intrahepatic bile duct cancer. Specifically, 42 of these associations were not found in any of the ancestry-specific analyses, such as the two additional significantly associated regions from the meta-analysis of chronic liver disease (Figure 5.6), demonstrating that the added power can identify associations not significant in the ancestry-specific analyses. In the ancestry-specific analyses for heart disease, we only see 1 significant SNP across all ancestry groups that barely reaches genome-wide significance (p-value=$4.11 \times 10^{-8}$). After performing the meta-analysis, this increases to a total of 28 significant SNPs all within a locus on chromosome 9, with the top SNP having p-value=$3.22 \times 10^{-10}$.

We are not able to perform a meta-analysis for skin cancer since were are limited to only the statistics computed from the EA GIA group (N-Case=4,583, N-Control=17,603, M=6,017,984). The analysis in the AA group was omitted to insufficient sample size (N-Case=38, N-Control=1,923), and all of the SNPs in the HL (N-Case=247, N-Control=5720) and EAA (N-Case=83, N-Control=3,205) analyses had failed association tests at all SNPs. Specifically, the association tests resulted in a p-value estimate of NA which is likely due to the small number of cases or difference in allele frequencies across GIA groups. For example, the minor allele frequencies (MAF) of rs12203592, the top SNP for skin cancer in the EA analysis, greatly varies across GIA groups: MAF-EA=0.17, MAF-AA=0.01, MAF-HL=0.07, MAF-EAA=$6.0 \times 10^{-4}$. Thus, populations with a lower MAF of associated variants would require larger sample sizes to have sufficiently powered association tests. This demonstrates that in cases where there are large differences in MAF at associated variations, ancestry-specific analyses would be preferred over a meta-analysis which could actually lead to a reduction in power.

Next, we investigated the top significant association for each phenotype and GIA group through a PheWAS. For rs12203592, an intron variant in the *IRF4* gene, we observe significantly associated phenotypes related to skin cancer such as actinic keratosis and basal cell carcinoma in the EA GIA analysis, both of which have been identified in previous PheWAS studies [84]. At rs1333045, the top SNP associated with ischemic heart disease in the EA GIA analysis, we find related phenotypes such as coronary atherosclerosis and angina pectoris. We also perform a PheWAS at rs2294915, the lead SNP for liver/intrahepatic bile duct cancer in the HL GIA analysis and the lead SNP in the analyses for chronic nonalcoholic liver disease in both the EA and HL GIA analyses (Figure 5.7). We find that multiple neoplastic and neurological phenotypes reach significance exclusively in the HL analysis. These groups of phenotypes are consistent with the known comorbidities with severe liver disease[311, 282, 232]. Performing a power analysis on the effective sample sizes [314] of the associated phenotypes in both GIA groups, we do not find evidence that the observed effects are solely due to sample size. Overall, these findings suggest possible differential genetic architecture across these two populations, as well as variation even at the phenotype level, reflecting possible genetic or environmental modifiers of important comorbidities.

## 5.4 Discussion

As the field moves forward with increased collaboration between the genetics and healthcare communities, it is of utmost importance to also be aware of potential pitfalls that may occur when translating research findings into actual clinical populations. Currently, many clinical protocols implicitly perpetuate racial bias [307, 93, 99, 151, 221]. Many of these flawed policies stemmed from erroneously linking race, a social rather than biological construct, with disease risk despite not presenting any biological justification. Although race and genetic ancestry are correlated [250, 291], our work shows that populations constructed from these two concepts are not analogous. We encourage protocol decisions that are rooted in concrete biological phenomena whenever possible, such as genetic markers, providing transparent, immutable criteria. For example, Benign Ethnic Neutropenia (BEN) is observed predominantly in African Americans but specifically is strongly associated with the variant at rs2814778 [244, 23]. Recent studies have suggested that genotype screening at rs2814778 could aid in the interpretation of neutropenia in African Americans and avoid unnecessary invasive procedures as well as lead to an increase of the inclusion of these individuals to various treatments[302].

However, in practice, genetic information is not easily accessible to patients at all institutions. Additionally, certain disease prediction-based algorithms that leverage SIRE may be favorable to the non-adjusted version. SIRE is correlated with genetic ancestry as well as other disease risk factors (sociocultural, socioeconomic, and geographic), making it straightforward and more easily accessible to add valuable information into models without explicit measurements. We recommend deliberately considering the potential harm versus benefit of using SIRE-adjusted prediction models in each use case. The practice of race/ethnicity-guided algorithms and guidelines inherently reinforces the idea of race-based medicine and embeds the idea that health inequities also stem from biological differences. It is an ongoing discussion about whether or not the inclusion of race/ethnicity information actually re-allocates resources away from racial/ethnic minority patients, causing more harm and an increase in health inequities.

There are various limitations within our study, and we describe a few of these in detail as follows. First, the phecodes are based on ICD codes, and due to the nature of billing codes, this form of labeling does not constitute a formal patient diagnosis and may contain individuals who do not have the specific disease. We also only require the presence of one phecode to define a phenotype which is a significant assumption. Although we present exploratory analyses assessing the role of phecode occurrence when defining phenotypes, we underscore that this imprecise phenotyping limits the power of our study. For further investigation into specific phenotypes, we recommend refining each phenotype definition based on additional disease-specific factors and metrics. For example, one could incorporate additional EHR features, such as those described by the algorithms in the PheKB database [69]. Although ICD codes are an international standard, the accuracy of phecode assignment could differ considerably due to heterogeneity in billing practices across medical centers, hospitals, and clinics both within the UCLA Health System as well as across other institutions. This heterogeneity could present future challenges when replicating studies or porting algorithms to other institutions. Second, due to the de-identified nature of the data, we lack information that could help us better describe the fine-scale population groups. For example, birthplace, zip code, and family history have been shown to be useful descriptors for determining subgroups of genetic ancestry [35]. Geographic information could also be used as a proxy for various environmental exposures such as pollution. Additional socioeconomic information, such as income and availability of health insurance, could likely account for a portion of observed associations as well as provide more insight into the socioeconomic determinants of health. Third, our findings within the African and South Asian ancestry populations are limited due to the smaller sample sizes. As sample sizes increase, we hope to further refine population substructure within these initial continental ancestry groups and have the power to detect novel disease associations that have previously been mired by lack of statistical power.

An open question and potential additional limitation of this work is generalizing these results to broader populations that extend beyond the UCLA Health system. For example, even when assessing self-identified race/ethnicity statistics, there is a discrepancy between the breakdown of SIRE within ATLAS and the city of Los Angeles. For example, it is reported that 48.5% of residents of Los Angeles self-identify as Hispanic or Latino [299], compared to only 15.96% of individuals in ATLAS. This could be due to the specific location of the UCLA Health system which consists of hospitals both in West Los Angeles and Santa Monica which are both located in affluent neighborhoods. When comparing demographic data recorded for the city of Santa Monica, which could be a more accurate representation of the area surrounding the UCLA hospitals as opposed to Los Angeles as a whole, we find that 15.4% of individuals self-identify as Hispanic or Latino [299]. Overall, the distribution of the majority of racial groups in Santa Monica has a high concordance with those reported in ATLAS. Since travel to treatment centers is often a barrier to treatment [284], this might explain why the ATLAS population mostly captures the demographics of the nearby areas. Furthermore, previous work has shown that referral rates for some types of procedures vary disproportionately across race and ethnic groups [169, 281, 107]. As a tertiary and quaternary referral center, this pattern could be reflected in the UCLA patient population. In addition, this discrepancy could specifically reflect the variations in patient participation rates across demographic groups. Previous studies have shown that trust in the health system and the medical community is a large factor when patients consider whether to participate in medical research [264, 38, 257]. Overall, there are a myriad of factors that influence the population and health of a specific region such as socioeconomic status, political geography, immigration, and historical events– the majority of these not being race-neutral. These observations suggest that many of these analyses should be interpreted with respect to the UCLA Health system specifically and extrapolating results to larger geographic areas or groups as a whole should be done with caution.

## 5.5 Conclusion

In this work, we introduce the ATLAS Community Health Initiative, a biobank embedded within the UCLA Health medical system consisting of de-identified EHR-linked genomic data from a diverse patient population. The UCLA Health System serves Los Angeles County, leading to a study population of great demographic, genetic, and phenotypic diversity. We investigate ancestry both on the continental as well as the subcontinental population level and find that genetic ancestry and self-reported demographic information yield distinct subpopulations in the ATLAS biobank. We present a collection of results cataloging the associations between genetically inferred ancestry and EHR-derived phenotypes where we find that disease status is not only associated with continental genetic ancestry but also associated with the specific admixture profile describing an individual. We use multi-ancestry pipelines to recapitulate known associations for chronic nonalcoholic liver disease at the 22q13.31 locus and perform a PheWAS at the lead SNP, where we find associations with neurologic and neoplastic phenotypes exclusively in the HL GIA group. As the sample size increases, the ATLAS Community Health Initiative will enable rigorous genetic and epidemiological studies to further understand the role of genetic ancestry in disease etiology, with a specific aim to accelerate genomic medicine in diverse populations. Already, the ATLAS biobank accounted for 73.4% of the Admixed American samples utilized in the primary analysis from the COVID-19 Host Genetics Initiative [78].

We conclude by discussing directions for future work. Although we investigate admixed populations, such as African American and Hispanic/Latino populations, admixed individuals who do not fall under these groups are excluded from downstream analyses due to concerns over population structure. In the future, we hope to incorporate methods and pipelines that properly control for population structure in all types of admixed populations. Additionally, we plan to compute polygenic risk scores (PRS) across all 5 continental ancestry groups. PRS has already shown modest clinical utility for diseases such as breast cancer [193] and cardiovascular disease [166], but has proven difficult to perform accurate predictions across populations[195]. The genetic diversity within the ATLAS Community Health Initiative biobank partnered with the longitudinal clinical data provides a unique resource to further explore the role of ancestry in PRS prediction. Furthermore, as the size of the biobank grows and more data is collected over time, we hope to explore even more individualized health solutions and interventions.

## 5.6 Figures

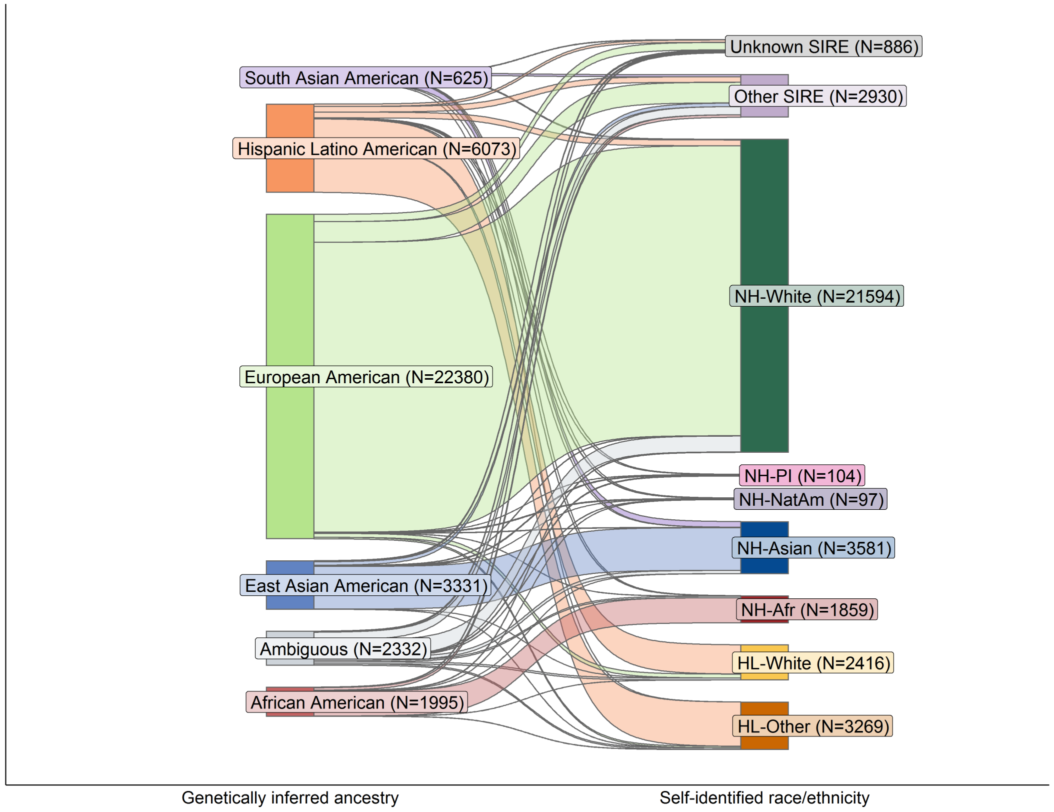

Figure 5.1: Self-identified race/ethnicity (SIRE) and genetically inferred ancestry (GIA) are not analogous. We show a Sankey diagram visualizing the sample size breakdown of individuals in each genetically inferred ancestry groups and SIRE groups for all individuals in ATLAS (N= 36,736).

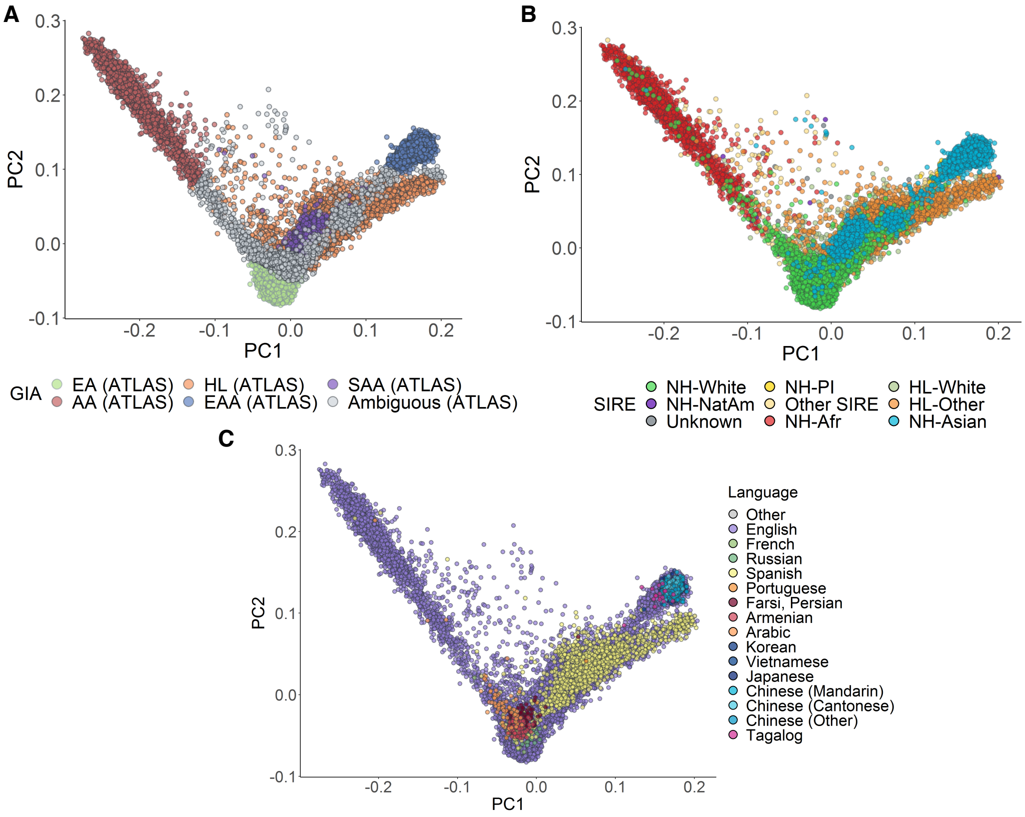

Figure 5.2: Global PCA reflects self-identified race/ethnicity and language of ATLAS participants. (A) Genetic PCs 1 and 2 of individuals in ATLAS (N=36,736) shaded by continental GIA as inferred from 1000 Genomes. (B) and (C) show the first two genetic PCs of the ATLAS participants shaded by SIRE and preferred language, respectively. To improve visualization in (C), only languages with > 10 responses were assigned a color.

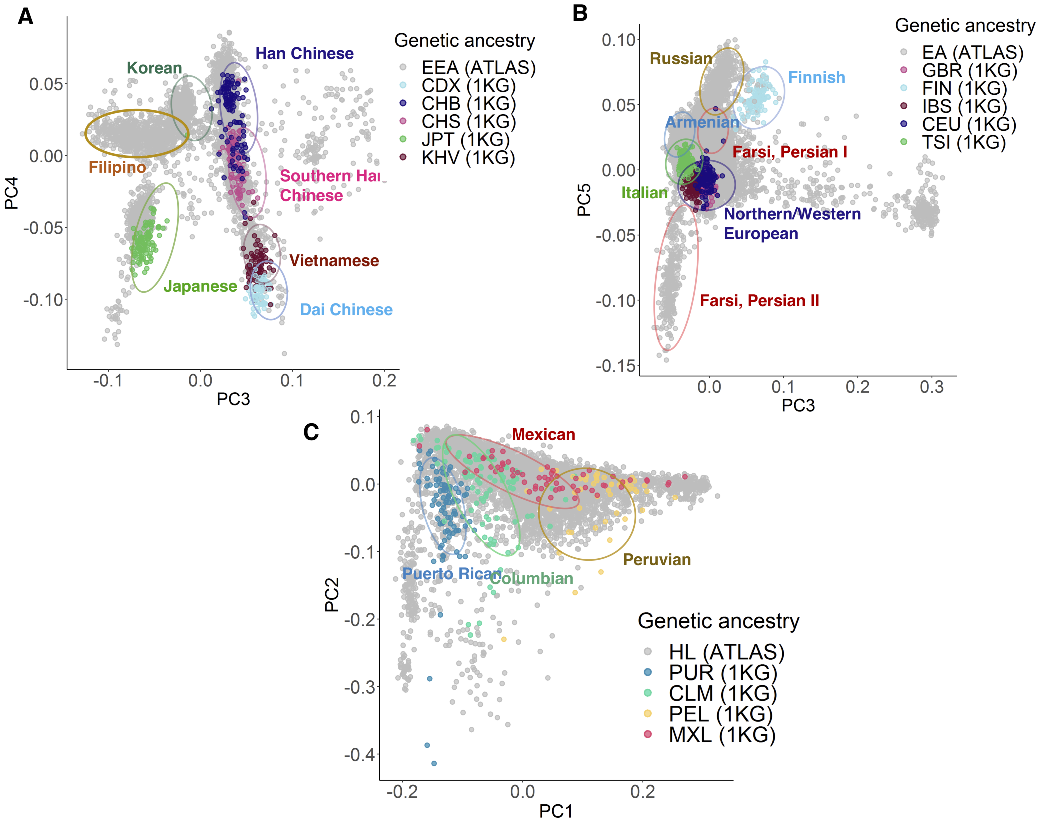

Figure 5.3: PCA of individuals with inferred East Asian American, European American, and Hispanic Latino American genetic ancestry in ATLAS captures fine-scale subcontinental ancestry groupings. PCA was performed separately within each continental GIA in ATLAS with the corresponding subcontinental ancestry samples from 1000 Genomes: (A) East Asian American, (B) European American, (C) Hispanic Latino American. Cluster annotation labels were determined using a combination of samples from 1000 Genomes and self-identified race, ethnicity, and language information from the EHR.

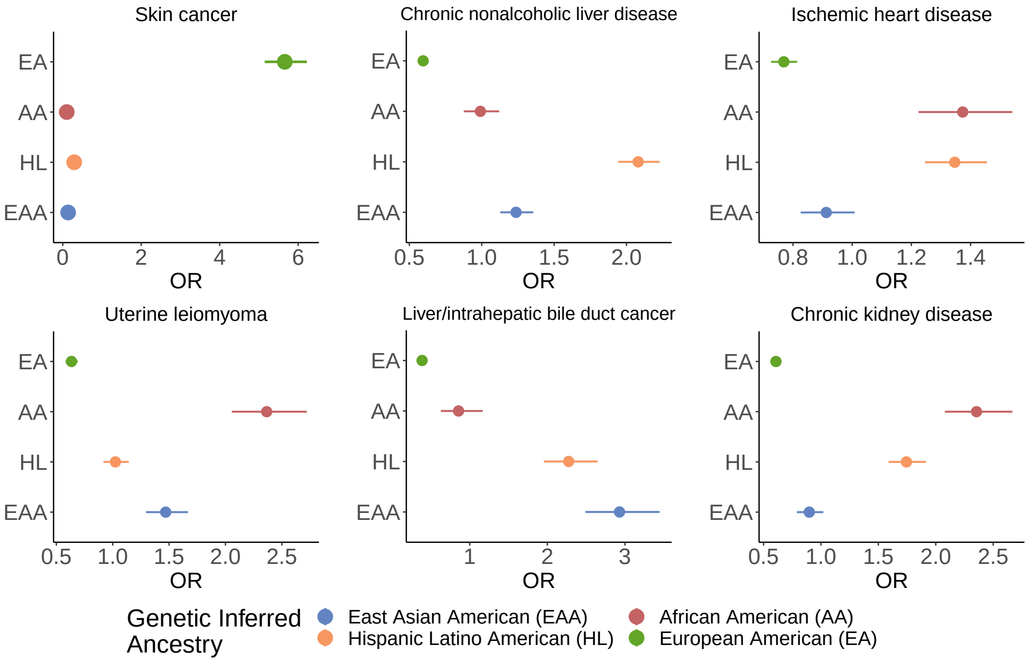

Figure 5.4: Disease associations vary across continental genetically inferred ancestry groups in ATLAS. We show the odds ratio computed from associating each phenotype with individuals' genetically inferred ancestry in ATLAS (N=36,736) under a logistic regression model. Error bars represent 95% confidence intervals.

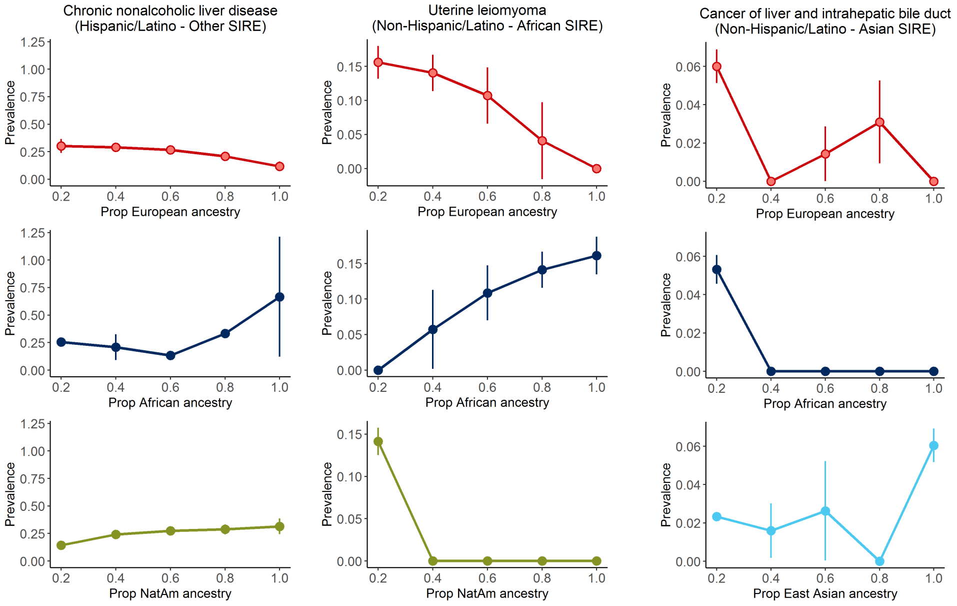

Figure 5.5: Global ancestry correlates with disease prevalence in admixed individuals. Individuals by SIRE who have had a diagnosis of (A) chronic nonalcoholic liver disease, (B) uterine leiomyoma, or (C) liver/intrahepatic bile duct cancer are binned by their proportions of either European, African, Native American, or East Asian ancestry estimated using ADMIXTURE. Within each bin, we plot the prevalence of the diagnoses and provide standard errors (+/-1.96 SE) of the computed frequencies.

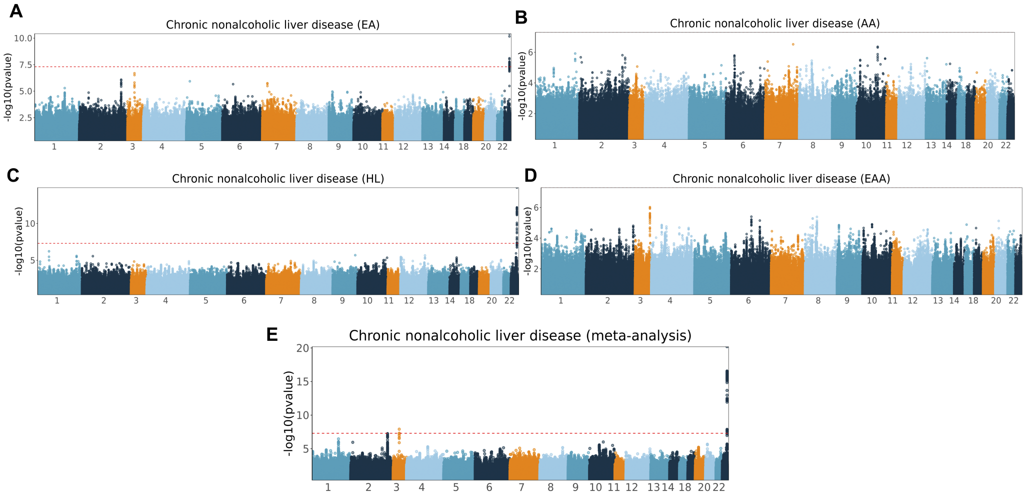

Figure 5.6: Recapitulating known associations for chronic nonalcoholic liver disease in ancestry-specific and multi-ancestry meta-analyses in ATLAS. GWAS Manhattan plots for chronic nonalcoholic liver disease in the (A) European American, (B) Hispanic Latino American, (C) African American, (D) East Asian American GIA groups in ATLAS, and (E) the meta-analysis performed across all 4 GIA groups. The red dashed line denotes genome-wide significance (p-value$< 5 \times 10^{-8}$). We recapitulate a known association at the 22q13.31 locus.

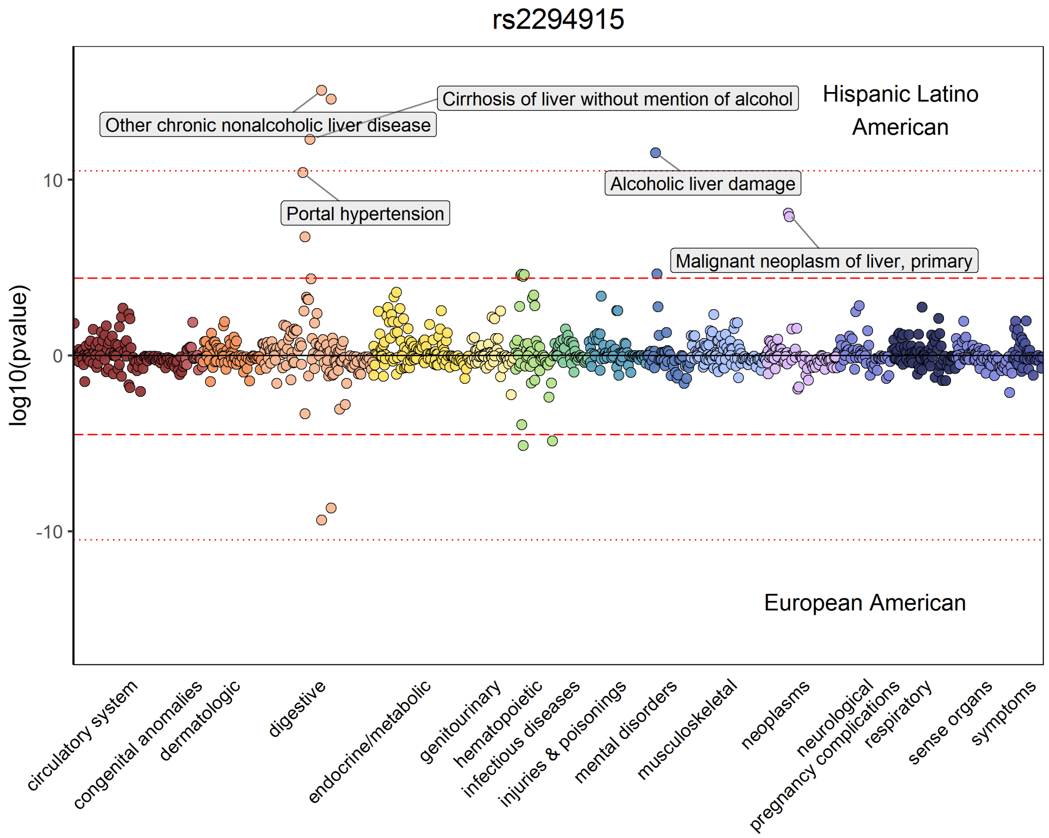

Figure 5.7: Identifying correlated phenotypes at rs2294915 in both the Hispanic Latino American and European American GIA groups in ATLAS. We show a PheWAS plot at rs2294915 for the Hispanic Latino American (top) and European American (bottom) GIA groups. The red dashed line denotes p-value=$4.09 \times 10^{-5}$, the significance threshold after adjusting for the number of tested phenotypes. The red dotted line denotes the significance threshold after correcting for both genome-wide significance and the number of tested phenotypes (p-value=$4.09 \times 10^{-11}$).

# CHAPTER 6

# Race, ethnicity, and genetic ancestry in the age of EHR-linked biobanks

## 6.1 Introduction

With the sharp downturn in the price of genotyping technology and the widespread integration of electronic health records (EHR) in most healthcare systems (¿95% in the US)(Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015, n.d.), millions of samples with accompanying EHR information have been collected across the globe. EHRs provide real-world patient information collected during routine clinical care. EHR-linked biobanks connect this information with human germline DNA samples, creating rich epidemiological resources that offer multidimensional insights into health and disease risk[9, 78, 326]. These resources offer a hypothesis-free approach, allowing the testing of millions of genetic variants against thousands of diseases or phenotypes at scale without the need for costly recruitment studies [33]. In particular, when studying common disease risk, EHR-linked biobanks can provide impressive sample sizes just through the accumulation of diagnoses from routine clinical care.

Global biobanking initiatives have created unparalleled resources for studying genetic risk in ancestrally diverse populations, which we define as groups of genetically heterogeneous individuals from varying ethno-linguistic and geographical backgrounds[117]. Large sample sizes from these populations were not previously available since the majority of prior genetic studies were performed in populations of European descent[239, 274]. Nationwide biobanks, including the BioBank Japan Project[214], the Qatar Biobank[1], the

China Kadoorie Biobank[68] and the UK Biobank[55], have enabled thousands of studies on population health. Single-medical and health-system-wide biobanks, such as the Kaiser Permanente Research Bank[29], Penn Medicine Biobank at the University of Pennsylvania[304], BioMe Biobank at Mount Sinai, and the BioVu Biobank at Vanderbilt University Medical Center, can increase ascertainment of cases for more rare diseases (e.g. cystic fibrosis) or for groups that are not typically not well-represented in general nation-wide ascertainment efforts, such as those with pediatric or geriatric disorders. The proliferation of EHR-linked biobanks could serve as a catalyst to our understanding of how genetic ancestry influences common disease risk and development. A more comprehensive list of current EHR-linked biobank initiatives is outlined in prior work[9, 317].

Many current clinical guidelines recommend race-based “adjustments” for treatments and diagnoses, such as diagnosing chronic kidney disease, treating pain, and assessing lung function, amongst many others[citations]. These stem from the flawed assumption that race is based on physiological differences; this same assumption is also what served as the foundation for the now-defunct idea of “biological race”. It has been suggested to instead use genetic ancestry to better explore how human variation affects disease risk. Completely eliminating the consideration of race and ethnicity in medical research could help prevent the perpetuation of biological race, but it is also too simplistic to overlook the ideas on race that have so deeply shaped healthcare in the United States (US)[61, 321]. Advocates of ‘race-conscious’ medicine express that race and ethnicity can still provide valuable information regarding differences in structural conditions and lived experiences whereas genetic ancestry can be used to describe biological evidence[62]. When using EHR-linked biobanks to interrogate underlying disease etiology, it is important to be aware of how the biases due to race and ethnicity can propagate to downstream analyses and its potential for purporting socially derived patterns as true biological signal. This highlights a critical complexity of EHR-linked biobanks: investigating disease risk due to variation in genetic ancestry in a clinical landscape shaped by race and ethnicity.

This perspective aims to discuss how self-reported race and ethnicity and genetic ancestry are currently utilized in EHR-linked biobanks for studying common disease variation specifically. We highlight considerations where there is no consensus about best practices and provide transparency about the shortcomings of current methodologies. We intend for this resource to support researchers working in the space of EHR-linked biobanks and precision medicine, such as those involved in consortium-wide biobanking efforts[78, 4, 326]. We approach the discussion of these topics with a forward-looking perspective and recognize that the field has used these concepts imprecisely, as well as exploitatively, in the past, but we maintain that this type of information can be beneficial for scientific discovery when used appropriately in specific analyses. Throughout this work, we illustrate specific applications utilizing data from the UCLA ATLAS Community Health Initiative (ATLAS)[134], including inference of genetic ancestry and applications in genome-wide association studies and polygenic risk scores.

## 6.2 Revisiting race, ethnicity, and genetic ancestry

The discussion of how to properly utilize race, ethnicity, and genetic ancestry information in biomedical studies has [27, 101, 131, 132, 148, 273] been an ongoing conversation spanning decades. Recent events and initiatives have spurred renewed conversation regarding what genetic ancestry represents and its utility in genomic medicine. First, the genetics field has come face-to-face with its Euro-centric biases and the impact of these decisions in genomics research[239, 274]. The current lack of ancestral diversity in genome-wide association studies (GWAS) is a harbinger for healthcare disparities in these understudied populations. This concern is not unfounded– polygenic risk scores (PRS) already demonstrate poor predictive utlity when models trained on individuals with European genetic ancestry, which account for the vast majority of currently available PRS models[158], are applied across populations with non-European genetic ancestry[194, 195].

These findings have motivated large-scale government-funded projects that aim to ameliorate these disparities. In 2015, the White House announced the Precision Medicine Initiative with an initial investment of $215 million to fund research in disease prevention and treatments that take into account individual patient needs[5]. From this initiative, the All of Us Research Program was established with the goal of genotyping one million individuals across the United States with a focus on groups that have been historically underrepresented in biomedical research[13]. The National Institutes of Health (NIH) have also separately funded initiatives such as the Population Architecture through Genomics and Environment (PAGE) Study which was designed to characterize the genetic architecture of complex traits among underrepresented minority populations through multi-ethnic omics and multi-ethnic PRS[316].

Second, there has recently been a drastic shift in the appropriation of genetics research by extremists and White nationalists, twisting scientific work to claim biological justification of a racial hierarchy[226]. The exemplary sample sizes of the biobank era have also subsequently increased the number of complex traits that show a degree of genetic signal. This includes numerous studies on phenotypes relating to cognitive behavior and education[147, 157, 162, 248], which are willingly misappropriated to support ideas regarding racial ties to IQ, which has been thoroughly disproven[113, 297]. However, these false ideas can eventually manifest as hatred and violence. Recent racially-motivated attacks (as of February 2023) such as those in Buffalo, NY, Christchurch, New Zealand, and El Paso, Taxes, have brought to light how modern research can be maliciously distorted[226]. The postgenomic era has provided copious amounts of genetics research as well as granted access to genetic datasets to the public (e.g. publicly available data such as 1000 Genomes[7] or direct-to-consumer genetic ancestry tests). These resources have been used by racist pseudo-scientists to create new content that supports their supremacist beliefs which are then disseminated on online platforms, feigning as academic-based studies[255].

Many members of the research community have initiated conversations about potential guardrails in genetics research and scientific communication that can help prevent this misuse[21]. For example, although EHRs and biomedical literature often contain the term “Caucasian” as a racial category, this term has since been denounced due to its erroneous origins and historically racist implications[238]. A recent review has found a precipitous decline in the term's usage in recent biomedical publications[56]. Additionally, an insightful review by Melissa Wills outlines how slight ambiguities in scientific writing and visualizations can lead to the misappropriation of genetic ancestry and population genomics studies[313]. Throughout this work, we utilize the guidelines outlined by the American Psychological Association[17] when describing racial and ethnic identity and the recent recommendations from Khan et al. when describing genetic ancestry[141]. Furthermore, the National Academies of Sciences, Engineering, and Medicine is currently convening to establish best practices for specifically using race, ethnicity, and other population descriptors in genomics research. Organizations within the research community, such as the American Society of Human Genetics, have also formally addressed their fraught historical ties with racism, eugenics, and systematic discrimination and are actively working towards equitable policies to prevent its resurgence[16, 18, 19].

## 6.3 Digitizing race and ethnicity for EHR

EHR was never designed for research purposes, but rather has been designed and optimized for patient clinical care and practice management. Self-identified information such as race and ethnicity (SIRE) is collected to fulfill federal data reporting requirements[118]. Race and ethnicity are studied in a myriad of academic disciplines with no universal set of definitions. In this work, we use the definitions typically referenced in the medical literature. We define race as one's identification with a group or identity typically based on a variety of factors, including physical characteristics, social identity, and geographic history. Ethnicity is one's identification with a cultural group that typically shares traditions, language, and cultural norms[27]. We emphasize that both of these concepts are entirely social constructs and that neither has a direct connection to biology or genetics. Information on SIRE is typically recorded in EHR using structured data fields where selections are limited to a list of controlled vocabularies[288]. The Office of Management and Budget (OMB) outlines the minimum reporting standards and vocabulary for race and ethnicity in the US[118]. The minimum fields for race are American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White.

To prevent the insidious integration of socially constructed biases in precision medicine, it is crucial to acknowledge and account for the biases and complexities associated with the secondary use of EHR for research. Most EHR systems do not allow for the selection of multiple race (or ethnicity) values or the use of freeform text to describe an individual's identity. This inherently constrains how an individual can report this information. For example, representing race as "pan-ethnicities", such as "Asian" and "Black" (both of which are recommended by OMB), can lead to ambiguity among participants[22]. Historically, individuals of mixed race are assigned minority status despite their actual percentage of genetic ancestry[186]. This practice of hypodescent derives from the long-lasting effects of the "one-drop" rule in the United States where individuals with any perceived African lineage were subjected to racial segregation and oppression[121].

It cannot be wholly assumed that information reported in the EHR was provided directly by the patient. Clinical staff and healthcare providers may directly enter patients' race and ethnicity information based on the patient's physical appearance alone. Prior studies have also shown a degree of discordance between the SIRE information reported in the EHR and the SIRE information provided when being asked again through other communications (e.g. phone call, enrollment survey)[34, 149]. Furthermore, an individual might not completely fill out questionnaire fields if the forms are not available in their preferred language[69], resulting in non-random patterns of missingness in the EHR. Although the biomedical field has seen a steady transition away from explicitly using race in genomics research[56], it is unlikely that the field will completely diverge from the utilization of race and ethnicity due to its long-standing integration in the healthcare landscape.

## 6.4 Role of ancestry in the genetic architecture of complex traits

Information about our biological ancestors is encoded through a continuum of complex genetic variation. In this work, genetic ancestry refers to the characterization of the population(s) from which an individual's recent biological ancestors originated and the genetic relationship implied by the shared DNA segments inherited from these ancestors[197]. We use this term to specifically describe individuals where the true origins of their recent biological ancestors is known (e.g. Simons Genome Diversity Project[189]. Genetic ancestry has a significant impact on the genetic landscape underlying human disease risk. Understanding these genetic patterns within and across ancestral populations provides an empirical basis for studying how genetic architecture influences common disease risk. Here, we refer to genetic architecture as the characterization of the genetic contributions influencing a given phenotype, including the number of causal variants, variation in allele frequencies, and effect sizes on phenotype[290]. Throughout this work, population descriptors describing genetic ancestry are explicitly stated as such (i.e. African genetic ancestry). We also attempt to be precise as possible when communicating results from specific studies by utilizing the exact population descriptors reported in each study.

### 6.4.1 Shared common genetic variation

Demographic and evolutionary events have profoundly shaped genetic variation through forces such as genetic drift and adaptive evolution. These factors are reflected in our genomes, leading to distinct patterns of genetic variation across ancestral populations. Most common variants (MAF $> 1\%$) are shared across populations around the globe and association studies have shown that many of these shared variants confer disease risk across populations[184]. This is unsurprising considering that most common genetic variation was already present before the human expansion out of Africa[148]. Significant associations have shown modest replication across multiple genetic ancestry groups[171, 192, 316] and trans-ethnic fine-mapping efforts have further refined these associations to sets of prospective shared causal variants[144, 156, 160].

Furthermore, increased sample sizes from biobanks have enabled the power to capture small genetic effects that do not reach genome-wide significance. Prior studies have found that this extensive sharing of causal variants is not limited to GWAS risk regions, but instead is highly polygenic and extends across the entire genome[133, 265, 50]. Estimates of trans-ethnic correlation ($\rho$), or the correlation of effect sizes between populations of different genetic ancestry, consistently demonstrate $\rho > 0$, meaning that there is some level of shared genetic risk across populations for most complex traits. The strength of this correlation has been shown to greatly vary by trait, ranging from $\rho = 0.98$ for schizophrenia and $\rho = 0.46$ rheumatoid arthritis (both computed between European and East Asian genetic ancestry studies)[156]. However, finding adequate sample sizes of non-European genetic ancestries to test these hypotheses regarding shared genetic architecture is difficult, especially for more rare disease phenotypes[130].

The depth of the clinical phenome available in EHR-linked biobanks provides a unique opportunity to assess the cross-ancestry replication of disease associations at scale. Recent work from Bastarache et al. provides a resource called the Phenotype-Genotype Reference Map (PGRM), a curated set of vetted replicable genome-wide significant associations (p-value $< 5 \times 10^{-8}$) spanning 5 genetic ancestry groups and encompassing over almost 150 disease phenotypes. Each disease phenotype is mapped to standardized ICD-based diagnoses (phecodes[85]) to enable large-scale replication assessments across multiple genetic ancestries within EHR-linked biobanks. International initiatives such as the Global Biobank Meta-analysis Initiative (GBMI) have also aided these efforts by bringing together 24 biobanks across the globe to characterize shared and ancestry-specific genetic variation across standardized phenotype endpoints[326].

### 6.4.2 Effect size heterogeneity

Even if causal variants are shared across populations, differences in disease risk can stem from effect size heterogeneity[220, 316]. Meta-analyses from the first large-scale GBMI study found that 3% of genome-wide significant loci demonstrated significant heterogeneity in effect sizes across ancestry[326]. A separate study stratifying genetic effect sizes by proximity to functional elements (e.g. UTR, histone marks), reveals estimates of stratified trans-ethnic genetic correlation show that effect sizes tend to be more heterogeneous in functionally important regions[266].

Heterogeneity can also occur when the observed effect sizes from GWAS appear to have varying magnitudes (or direction) due to different levels in LD. This is especially relevant when there are ancestry-specific differences in LD between the unobserved causal variant(s) and the observed associated variant(s) inferred from GWAS[152, 265]. Alternatively, this observed heterogeneity could stem from unmodeled environmental factors, such as gene-environment interactions. This is of particular concern given that environmental factors, such as income, insurance, and lived experiences, can greatly vary across individuals and be highly correlated with race and ethnicity. Consequently, these unmodeled factors could lead to observed differences in the estimated genetic effects across study populations even though the differences stem from non-genetic factors.

### 6.4.3 Divergence of allele frequencies

Associations that do not replicate across genetic ancestry groups could be due to differences in allele frequencies. A wide body of evidence has shown that allele frequencies can diverge across populations due to pressures of negative selection which has been estimated to influence up to 85% of the genome. For example, MC1R, which regulates skin pigmentation and is associated with an increased risk of melanoma, exhibits reduced sequence diversity specifically in individuals of African ancestry. It is hypothesized that this is a result of negative selection where variation associated with skin cancer was reduced due to functional constraints of the high ultraviolet radiation environments in parts of Africa[258]. Severe population events, such as bottleneck events, can steeply alter the allele frequency within a surviving population through genetic drift. EHR-linked biobanks can be mined at scale to find previously uncharacterized relationships between patterns of fine-scale genetic ancestry and the clinical phenome. For example, a study in the BioMe Biobank revealed that a genetic mutation in COL27A1, which was previously thought to be rare, is at an appreciable frequency in individuals of recent Puerto Rican descent, a population that has undergone a strong bottleneck event[36].

Biobanks should take caution when performing MAF-based quality control procedures to prevent the filtering of variants that demonstrate highly variable MAF across ancestral populations. Quantity control metrics that are sensitive to variable allele frequencies, such as estimates of individual heterozygosity and tests of Hardy-Weinberg equilibrium, can be performed within each genetic ancestry group individually. The choice of imputation reference panel can also greatly affect the accuracy of low-frequency imputed variants. This is especially relevant for samples of African or Hispanic/Latino ancestry due to the complex linkage disequilibrium (LD) patterns associated with the continuous admixture of ancestral populations that remained in Africa during the out-of-Africa migration[86, 177, 303]. More recent efforts, such as the TOPMed Project[285], have provided large, ancestrally diverse whole genome sequencing panels that substantially improve imputation quality in underrepresented admixed populations compared to prior panels that predominantly used samples of European ancestry[150]. A thorough discussion on recommended workflows and considerations for analyzing samples from ancestrally diverse populations is provided in Peterson et al.[229].

### 6.4.4 Population private genetic variation

Recent whole-genome sequencing efforts have provided insight into ancestry-specific variation, where the majority of this variation is represented by rare variants (MAF ¡ 1%)(1000 Genomes Project Consortium et al., 2015). These low-frequency variants rose to population levels relatively recently in evolutionary history and have remained geographically localized due to this nascency[148, 292]. African populations are expected to have more population-private alleles and haplotypes whereas modern-day non-African populations only contain a fraction of this genetic variation. This genetic divergence is due to the population bottleneck and the resulting loss of genetic diversity associated with the migration out of Africa[57]. Thus, given the expected amount of population-private disease risk variants, there is concern that rare genetic variation has been predominantly explored in individuals of European descent.

This lack of representation leaves substantial knowledge gaps in our understanding of genetic variation for the majority of the global population and consequently creates blind spots when applying findings from genomic medicine research. For example, it was found that individuals with African ancestry received false positive results for hypertrophic cardiomyopathy based on genetic variants classified as “Pathogenic” or “Likely Pathogenic” at the time. However, these were reclassified after it was found that these variants were actually benign in a large portion of individuals with African ancestry. Retrospective analysis shows that this oversight could have been prevented by the inclusion of a modest number of individuals with African ancestry in the primary study[191]. This is just one scenario demonstrating how the extrapolation of genetic findings to multiple populations can unintentionally contribute to inequities in healthcare and how the increased representation of diverse populations can help ameliorate and prevent new healthcare disparities[200]. Recent efforts have shown the potential of mining large-scale data biobanks with exome data for identifying uncharacterized, pathogenic rare variation in diverse populations. A study within the BioMe Biobank systematically re-assessed the connection between the estimated penetrance of pathogenic variants and the actual recorded disease outcomes by scanning through the clinical phenome of the EHR[100]. This revealed that the estimated penetrance of pathogenic/loss-of-function variants was actually relatively low, providing a tangible step towards variant reclassification, which historically has been a slow process[235].

### 6.4.5 Correlation between genetic ancestry and SIRE

Despite being separate concepts, genetic ancestry and race and ethnicity still maintain a level of correlation due to the shared demographic and historical events that shaped subsequent human populations, societies, and cultures. Prior studies have shown that the evolution of language broadly follows human evolution and that some factors influencing genetic variation (e.g. mating, migration) also shape the ethnolinguistic diversity of a population[25, 185]. Variation in skin pigmentation is largely explained by genetic factors, where patterns in the genome were largely shaped by adaptive selection due to differences in UV exposure across. Sociocultural factors such as endogamy associated with religion and other cultural practices lead to assortative mating which limits gene flow between different populations. Extensive endogamy can also lead to an increased disease burden due to the effects of genetic drift within a reduced effective population size. For example, within the Jewish community (specifically those of Ashkenazi heritage), autosomal recessive disorders, such as Tay-Sachs and familial dysautonomia, occur at an increased incidence due to the high frequency of certain founder mutations. This has led to the development of extensive carrier screening panels for these mutations where the American College of Obstetricians and Gynecologists guidelines recommend that couples of Ashkenazi Jewish ancestry receive carrier screening. However, we emphasize that even if natural phenomena, such as evolution, influence traits that are connected with current ideas surrounding race and ethnicity (e.g. skin pigmentation, religion), this does not connote a biological connection between these ideas and genetic ancestry. We encourage the use of genetic ancestry when describing biological concepts, such as genetic disease risk, and the use of SIRE when discussing non-genetic, socio-cultural factors.

## 6.5 Modeling genetically inferred ancestry in disease risk

For EHR-linked biobanks with DNA samples, genetic ancestry is a favored method for describing the biological similarity between individuals due to the strong correlation between patterns of common genetic variation and genetic ancestry. Since true ancestral information is often not available in EHR-linked biobanks, we can instead utilize genetically inferred ancestry (GIA) or the inferred genetic characterization of individuals within a group who likely share recent biological ancestors. Specifically, GIA is dependent on the specific inference processes performed (e.g. PCA and clustering) and the choice of reference data used to compare genetic ancestry. This procedure largely eliminates the dependence on self-reported participant information in genetic studies.

### 6.5.1 Characterizing genetic ancestry through PCA

One commonly used approach to assessing common disease risk in ancestrally diverse biobanks is to first stratify individuals by GIA. A predominantly used technique to characterize genetic ancestry is through principal component analysis (PCA)[136]. Specifically, PCA is performed on the matrix of individuals' genotypes where the largest principal components tend to capture variation due to ancestry-specific differences in allele frequencies[202]. Projecting the data into 2D space shows the continuous spatial variation among samples which is highly correlated with global geography[219]. Clustering algorithms (e.g. K-nearest neighbors) can be applied post-hoc to create groups of individuals based on similar patterns of spatial variation.

GIA is often fitted to ancestral populations characterized by continental boundaries such as the "superpopulations" present in 1000 Genomes (e.g. European). When projecting data from ATLAS and 1000 Genomes into PC space, many individuals that self-identify as White span all along PC 1 and PC 2, but the European GIA cluster only captures a small fraction of this variation (Figure 6.1A). This dichotomy is especially pronounced in admixed groups such as those self-identifying as Hispanic or Latino and the Admixed American GIA group (Figure 6.1C). However, ideas of race are often conflated with a continental-level representation of genetic ancestry since society tends to also delineate race by continental borders. Because EHR already contains information on SIRE and is easier to collect in general, these concepts have been used interchangeably in past biomedical studies. A survey of previously published studies involving human health revealed that 49% of studies used the term "ancestry" when referring to non-genetic data and that "continental ancestry" was the most commonly utilized grouping level[82]. This construct conflation can lead to the inadequate profiling of disease risk and flawed conclusions regarding differences in biological signals, even when populations have identical underlying genetic risk profiles.

Furthermore, because SIRE is not standardized across all healthcare systems, both within the US and internationally, combining results across multiple biobanks that only use SIRE for stratifying individuals can lead to unaccounted population structure since ideas surrounding race and ethnicity often vary across geographic regions[22]. For example, in Brazil, there is not an equivalent concept to race as in the US, but instead, their census uses "race-color" categories which are primarily based on skin pigmentation. However, numerous studies have shown that skin pigmentation is a crude estimate of genetic ancestry[165, 188]. Furthermore, analyses would inherently be limited to the predetermined list of self-identifying fields available in the EHR. PCA plots should be interpreted with caution. "Synthetic maps" created by projecting geography onto PCs have led to false conclusions regarding the determination of a given individual's true biological origins. These PCA and clustering methods do not provide an absolute assignment of an individual's true biological origins, but rather can be viewed as a statistical methodology used to estimate patterns of similarity between genotypes. For this reason, it has been proposed to shift to descriptors of genetic similarity as opposed to names of genetic ancestry groups when describing patterns derived from these types of statistical analyses[76]. A more pernicious assumption is that ideas about an individual's race can be inferred from PC patterns. These erroneous assumptions have resulted in dangerous claims, such as the determination of Jewish ancestry based on genetic information alone(Need et al., 2009). The idea of a distinct genetic signature being used to determine Jewish descent has been widely refuted by numerous studies[91, 94, 293]. We emphasize that extreme caution and consideration be taken when communicating these observations so as to not insinuate connections to biological determinism, especially given the willful appropriation of scientific literature by the White nationalist community[58].

### 6.5.2 Characterizing genetic ancestry through genetic admixture

Describing genetic ancestry through discrete categorizations alone obscures more fine-scale population structure. Much of the observed genetic variation across populations is due to admixture events where individuals from two or more previously diverged or isolated populations (or ancestral populations) mate to form new genetic lineages[269]. The extent of this variation can be measured through global admixture proportions which can be thought of as the average ancestry contribution over an individual's genome. When looking at the estimated ancestry proportions within ATLAS, we see extensive genetic variation within each SIRE category (Figure 6.2). In the African GIA group, individuals are estimated to have between 0-100% African ancestry. Furthermore, when looking at the Hispanic/Latino-Other Race and Hispanic/Latino-White SIRE groups, the admixture profiles look approximately the same despite individuals self-identifying as different races. This not only highlights the impreciseness of SIRE but also demonstrates how collapsing genetic ancestry information into broad categories can obscure extensive levels of variation within a population. For example, Hispanic and Latino populations have a history of extensive admixture between European, African, and Native American ancestral populations. We demonstrate this phenomenon in ATLAS with a ternary plot that visualizes the three-way admixture of individuals that self-identify as an ethnicity other than Non-Hispanic (Figure 6.3). Within ATLAS, individuals self-identifying as Mexican show lower levels of African ancestry when compared to those self-identifying as Puerto Rican which is consistent with prior estimates. Despite the substantial patterns of genetic substructure within the Hispanic and Latino populations, this group is typically represented as a single category in genetic studies.

Modeling GIA through genetic admixture, as opposed to categorical measurements, also provides the opportunity to study disease risk in individuals from highly admixed populations who are typically excluded from GWAS. Specifically, in ATLAS, 7% of individuals are excluded because the samples were unable to be characterized into a single GIA category. As of February 2023, according to the live GWAS Diversity Monitor[205] (https://gwasdiversitymonitor.com/), admixed individuals make up only approximately 2% of GWAS participants which consequently restricts the potential for identifying valuable ancestry-specific information regarding disease risk in these populations. An alternative approach is to perform an association test between a disease phenotype and individuals' genome-wide ancestry instead of individual variants. This can be especially useful for diseases that show different disease burdens across populations but have yet to yield any large effect sizes from GWAS. For example, there is a marked difference between the burden of asthma in Puerto Rican and Mexican ethnic populations even though both are typically grouped under the umbrella ethnicity of “Hispanic or Latino”[6]. Prior studies have shown that asthma risk and lung function is correlated with African ancestry. One hypothesis for the observed difference in disease prevalence is due to the fact that, on average, Puerto Ricans tend to have a larger proportion of African ancestry than Mexican individuals[54, 231].

## 6.6 Bias in phenotype construction from EHR

The effects of racial bias can propagate to association studies when constructing phenotypes from the EHR. Diseases that are consistently underdiagnosed in racial or ethnic groups can translate to a loss of power in GWAS since a considerable number of individuals with a disease will be mistakenly labeled as controls[90]. For example, although Black patients are diagnosed with schizophrenia four times more likely than White patients(Barnes, 2004), one of the largest schizophrenia studies of participants with African ancestry from the Million Veterans Program revealed that the prevalence of schizophrenia for participants was not associated with individuals' PRS risk strata[40]. This discrepancy suggests that common genetic variation may not be the primary driver of the high schizophrenia prevalence in Black populations. Other hypotheses suggest that the increased diagnostic rate could be due to clinician bias or misdiagnosis of mood disorders[30, 261]. This demonstrates the numerous entry points for algorithmic bias and its potential prolific downstream effects in resulting conclusions regarding disease etiology.

Clinical notes and natural language processing (NLP) are among the most common data sources used when constructing phenotypes from the EHR(Kirby et al., 2016), including the phenotyping algorithms utilized in the Phenotype KnowledgeBase, a collection of validated EHR phenotyping algorithms[146]. However, racial biases in clinical notes can also lead to imprecise phenotyping, driven by circumstances of bias instead of true disease presentation. For example, prior studies show that stigmatizing language appears more frequently in admission notes for non-Hispanic Black patients[122]. One alarming example is from a study demonstrating how an NLP model for disease-related automated question answering demonstrates significant racial bias for pain mitigation, where clinical scenarios with Black patients were more likely to be refused pain treatment compared to White patients. Racial disparities in pain treatment stem from the completely speculative, yet widely accepted, belief that Black individuals have higher thresholds for pain[citations]. This emphasizes the need for formal, transparent assessments of bias and fairness on NLP models to screen for patterns and predictions that reinforce health inequities[32, 271, 287].

It is not typically known a priori whether or not algorithmically constructed phenotypes are biased nor the exact source(s) of the bias. A recent work by Dueñas et al. proposed recommendations to control for common healthcare biases in algorithmic phenotype construction for GWAS[88]. Specifically, one recommendation is to reduce sources of heterogeneity in the phenotyping population; for example, due to the known biases in schizophrenia diagnoses, phenotypes should be separately constructed in groups of White, Black, and Hispanic/Latino individuals. Additionally, for many diseases, it is expected that comorbidity profiles among cases should be relatively similar (e.g. medications, symptoms). Assessing phenome-wide differences across cases can help identify biased case/control assignments. Another insidious source of bias is case selection bias due to non-random missingness in the EHR. In particular, patterns of missingness are often correlated to healthcare access and proximity. Recent NLP methods attempt to recover this missing data through information present in clinical notes[143]. Resultantly, it has been shown that mitigating the missing-data bias can lead to improved disease prediction.

## 6.7 Modeling social determinants of health from the EHR

Because most genetic association testing frameworks assume that the environmental noise component (non-genetic factors) is shared across individuals, there is also considerable concern that unmodeled differences in environment can induce spurious associations. A common strategy to account for this stratification is to include SIRE as a covariate in the model, but this relies on the key assumption that SIRE is a suitable surrogate for systematic differences in environment such as social determinants of health (SDoH). However, SIRE is very broad and oftentimes captures a culturally heterogeneous population. This over-adjustment or unnecessary adjustment with covariates from the EHR can cause a reduction in power or false positives[20, 233, 259]. Additionally, many environmental and modifiable risk factors tend to correlate with race and ethnicity, such as diet, socioeconomic status, pollution, and neighborhood[109, 227]. However, if SIRE is used as a proxy for these variables, this prevents the opportunity for the development of healthcare policies that can explicitly address the source of the disparity[104]. Furthermore, there is apprehension that purporting SIRE as an adjustment factor can misleadingly assign race and ethnicity as contributing causes of disease.

Prior EHR studies have demonstrated innovative workflows to extract epidemiological variables or surrogate variables from the EHR[66]. Zip codes can be extracted from the EHR and then augmented with external datasets to construct a myriad of proxy measurements for specific variables of interest. For example, measurements of socioeconomic data per zip code such as median household income, percent unemployment, and percent below the poverty line, can be extracted from demographic databases, then summary statistics can then be used as a surrogate measurement for the individuals residing in that zipcode[254]. Other examples include measurements of "neighborhood walkability" to study BMI[89] and proximity to crop fields to assess agricultural risk in antibiotic resistance[60]. Resident addresses extracted from EHRs in tandem with geographic information system (GIS) technology can provide precise measurements of the human exposome, including measurements of pollution and environmental exposures (e.g. chemicals)[42]. Studies have even shown that birth month captures effects of seasonal variations experienced during infancy (e.g. climate, allergens, flu)[44, 43]. The uptick in health wearables also shows great potential in providing individual-level in-depth data about individuals' physiological factors. The continuous monitoring enabled by many wearables can also provide biometrics measured in everyday conditions (e.g. physical activity).

Ideally, specific SDoH should be measured and modeled explicitly to enable the development of targeted solutions. For example, prior studies have shown that many previously observed healthcare disparities can be attributed to socioeconomic status (SES) as opposed to biological factors once SES is explicitly accounted for. The Institute of Medicine conducted the Capturing Social and Behavioral Domains in Electronic Health Records study in an effort to improve how SDoH are represented in EHR. Specifically, the study recommended a set of social and behavioral domains such as educational attainment, stress, physical activity, and neighborhood median-household income, be integrated as fields within the EHR[70, 71]. Continuing efforts to include measurements of SDoH directly into the EHR are already underway and have shown great promise in increasing EHR-based clinical risk model performance[39, 66, 140, 305].

## 6.8 Applications in EHR-linked biobanks

### 6.8.1 Genome- and phenome-wide association studies in EHR-linked biobanks

EHR-linked biobanks present unique opportunities to conduct association studies across thousands of clinical phenotypes at scale and across multiple clinical populations. A major concern for both genome- and phenome-wide association studies is properly controlling for population stratification. The predominant approach is to use PCA-based GIA categories (see "Characterizing genetic ancestry to study common disease risk in EHR-linked biobanks"). However, at times, other demographic information within the EHR is used to group together individuals, such as language, country of origin, or parents'/grandparents' country of origin. For regions with a history of cross-continental immigration and exogamy, immediate family origin information might not wholly reflect the genetic ancestry of individuals[29, 204]. Prior analyses have also shown that an individual's chosen language is not a consistent identifier in the EHR and may vary due to concerns around stigma or bias[149]. Although the preciseness to which self-reported information surrogates for population structure information likely varies by cohort, prior biobanking studies have found evidence that genetic structure cannot reliably be accounted for by questionnaire data alone[204].

A key disadvantage of this grouped-stratification approach is that many individuals may be excluded because they are not able to be assigned to a specific GIA category. Instead, linear mixed models (LMM) avoid this by jointly modeling all genotypes to account for population structure and cryptic relatedness, bypassing the need for ad-hoc categories[138, 176, 329]. However, jointly modeling all individuals possess great computational concerns for large biobanks. Recent methods, such as BOLT-LMM[183, 182], have enabled scalable association statistical estimation for quantitative traits by modeling SNP effects within a Bayesian framework. Additionally, the SAIGE software employs a saddle-point approximation to enable scalable inference for case-control phenotypes[327]. Another method called SUGEN[172] utilizes generalized estimating equations to account for non-random sampling designs and intricate relatedness between participants, both of which are major concerns when modeling biobank data. Prior studies have shown considerable success in collectively modeling all individuals in large cohorts without an increase in Type I error regardless of genetic ancestry[127, 316], although there is still active concern regarding the use of LMMs as the main procedure for accounting for population structure[72, 280]. A more detailed primer on methods development for EHR-linked biobanks can be found in Wolford et al.[317].

### 6.8.2 Genetic risk prediction in ancestrally diverse populations

The postgenomic era has made way for the translation of genomic medicine into the clinic[112]. As sample sizes have grown, large-scale association studies have achieved the power to capture small genetic effects which have been critical for the increased accuracy of genetic risk prediction methods[65, 87]. These advancements have thrust the study of PRS into a fast-evolving, burgeoning field[168]. Already, PRS has shown promising success in retrospective studies for coronary heart disease[142], breast cancer[198], and type 2 diabetes[295]. Given the rapid pace of advancements in the area, this perspective aims to only highlight a selected set of considerations regarding the application of PRS in ancestrally diverse EHR-linked biobank settings.

Differences in genetic architecture due to population structure can greatly limit the accuracy of PRS when trained and applied to groups with different genetic ancestry[194].

Because PRS weights are often derived from GWAS summary statistics, the substantial Euro-centric bias of these studies means that, in practice, individuals with non-European ancestry or whose ancestry is not represented in large-scale GWAS are likely to be disproportionately plagued by less reliable prediction scores. Population genetics literature posits that the decay in accuracy is proportional to the genetic divergence between the training and discovery cohorts[262]. This pattern is also recapitulated in ATLAS when assessing the accuracy of a PRS for BMI across SIREs. Using weights previously estimated from samples of European ancestry[240], we see that the score achieves the highest performance in the White/Non-Hispanic SIRE and then decays as the proportion of European ancestry within each SIRE decreases (Figure 6.4). However the "Hispanic/Latino" - White and "Hispanic/Latino - Other Race" SIREs achieve similar performance despite the fact that these are separate groups. This is not surprising considering the similar admixture profiles but also underscores the impreciseness of SIRE when performing individual-level genetic prediction for disease risk. Empirical studies of PRS applied to human disease phenotypes also show that biases caused by these differences are unpredictable and the magnitude of this bias greatly varies by trait, indicating that post-hoc statistical corrections are unlikely to fully alleviate these discrepancies[194, 195].

There have since been large efforts to create methods that perform multi-ancestry PRS prediction[105, 289, 324]. Specifically, the PRIMED Consortium was established by the NIH to aid in the development of methods that specifically improve the utility of PRS in diverse populations[312]. However, at this time, there is no universally agreed-upon optimal solution for computing transferable PRS. A prominent concern is how an individual's genetic ancestry will be taken into account when computing patient-level PRS. For example, when performing risk prediction for an individual who is of admixed African and European, would it be more suitable to use a PRS model trained on African or European samples? Currently, there are no guidelines that make formal recommendations for 'ancestry-matching' or 'ancestry-adjustment' in the computation of PRS. A large-scale study being led by the eMERGE Network (Electronic Medical Records and Genomics) has integrated PRS prediction and clinical recommendations through a genome-informed risk assessment that will be returned to the patient[173]. Raw scores are “ancestry calibrated” through the adjustment of an individual’s genetically inferred ancestry information as measured by PCA[142]. In the same study, a survey shows that if given the opportunity to receive a genetic risk assessment, participants would be interested in receiving clinical recommendations despite being made aware that the given PRS could not be fully validated in all ancestral populations[173]. This brings up important ethical issues for patient care and underscores the imminence of these techniques in the clinic as well as its potential for misuse.

## 6.9 Towards a multidimensional approach to genetic ancestry in genomic medicine

The fast-approaching reality of personalized medicine exposes the need to move away from broad, categorical representations of genetic ancestry and instead embrace genetic ancestry as a full continuum[167]. The continuous nature of genetic ancestry is very clear, as seen in clines from PCA, extensive admixture, and spectrum of allele frequencies. Not only does this alleviate concerns regarding how to properly assign individuals to groupings, but this explicit categorization of individuals for clinical algorithms (even if using GIA) is an echo of many now-disputed ‘race-based’ healthcare practices. Instead, genetic ancestry is a multidimensional phenomenon that can be quantified in a myriad of ways and its utility is likely to be highly dependent on the specific phenotype of study and research goal. For example, in some applications, there is no need to explicitly quantify genetic ancestry. Instead, assessing disease risk could be performed by looking directly at the specific genetic mutations of an individual that ultimately impact the biological mechanism of interest. As genetic-based diagnoses and treatments are increasingly integrated into the healthcare community, it is important to be aware of how these concepts have been used in the past and consistently reevaluate the role and impact of these ideas in healthcare as research continues to evolve.

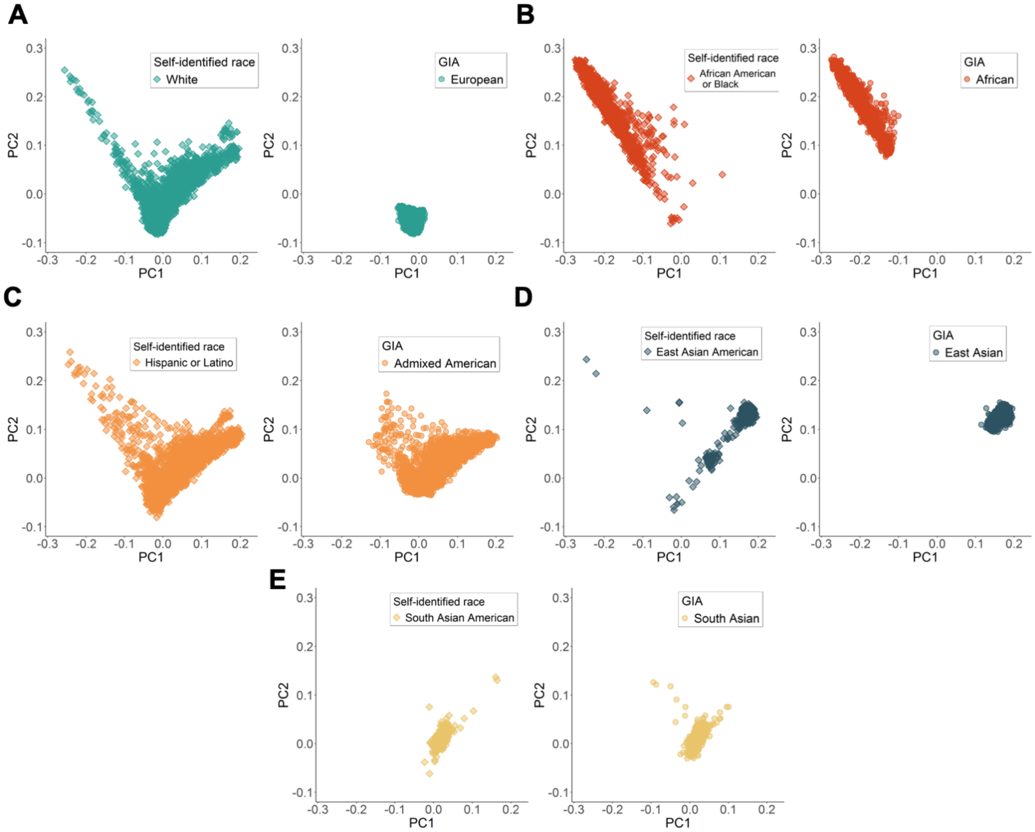

Figure 6.1: GIA and SIRE do not capture the same information. Within each subplot, we show the projection of PC1 and PC2 in ATLAS. Each left-hand panel shows individuals who self-identify as the listed race or ethnicity. Right-hand panels show individuals within each genetically inferred ancestry group.

## 6.10 Figures

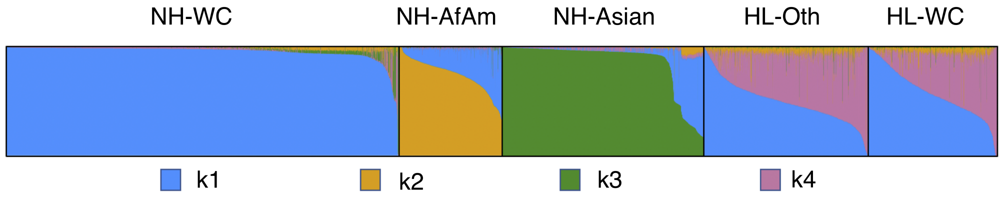

Figure 6.2: Admixture plot showing the ancestry breakdown stratified by SIRE. The ancestral proportions correspond to: European (k1), African (k2), East Asian (k3), Native American (K4).

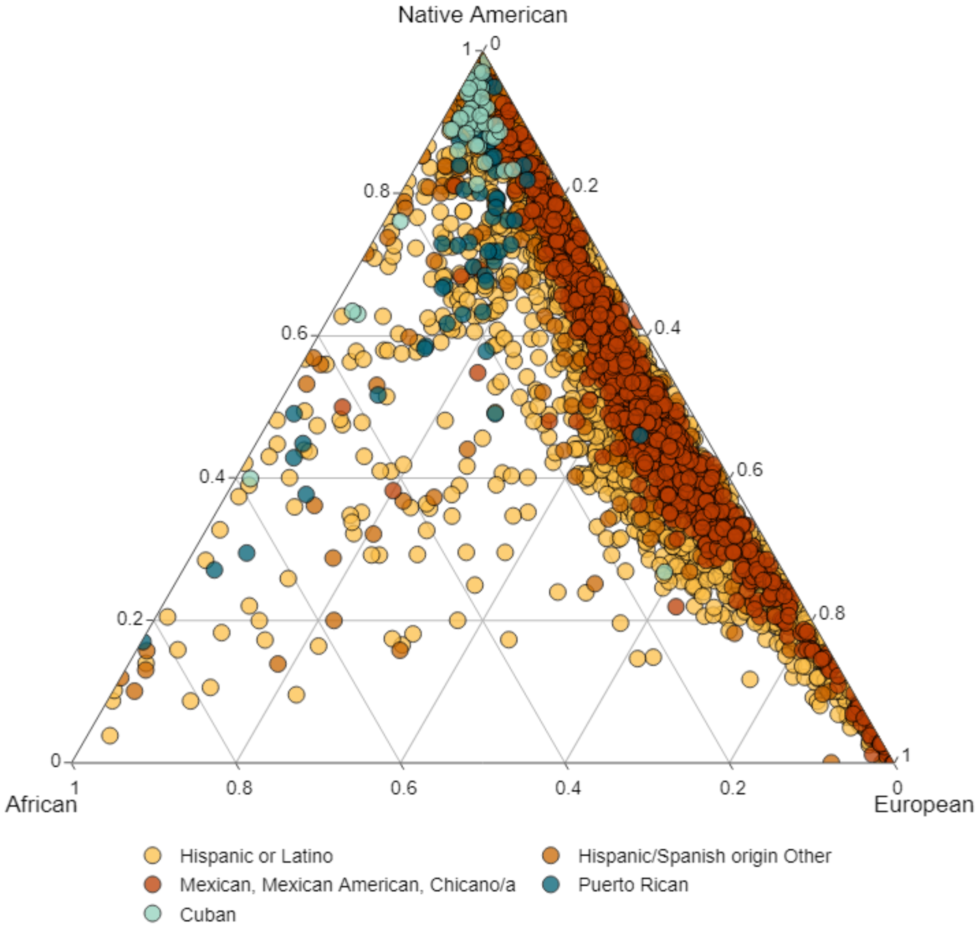

Figure 6.3: Genetic ancestry proportions greatly vary within the Hispanic/Latino SIRE. Ternary diagram shows the proportion of inferred African, Native American, and European genetic ancestry within the self-identified Hispanic and Latino population in ATLAS.

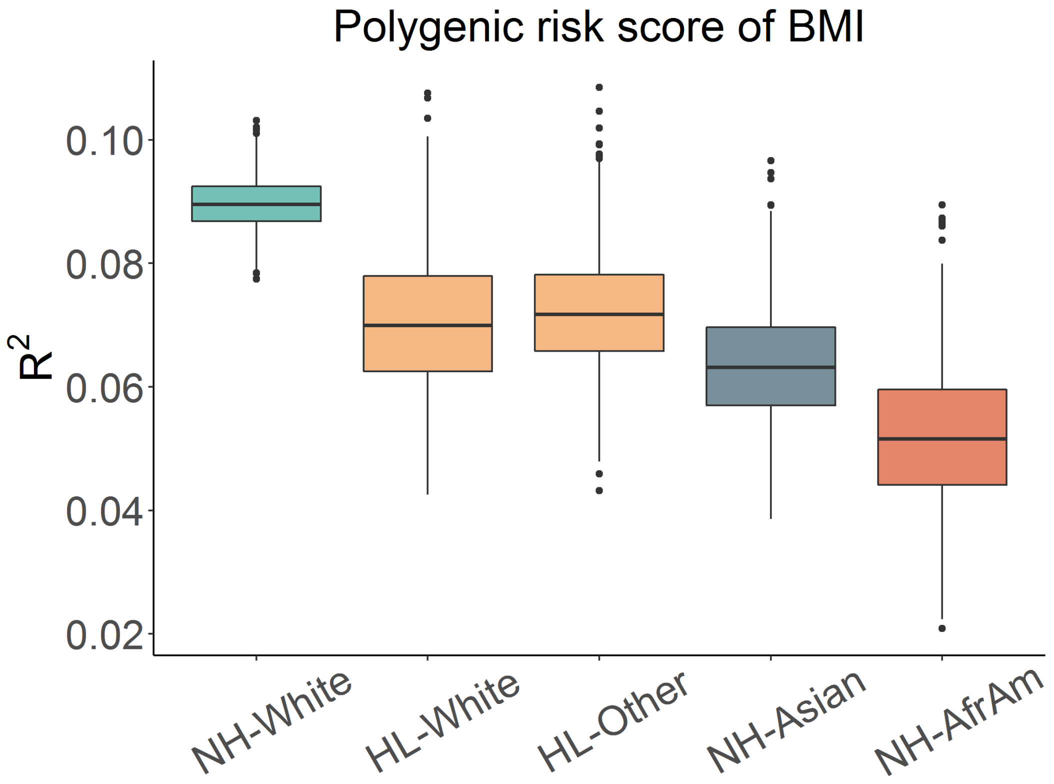

Figure 6.4: Performance of polygenic risk scores have variable performance across SIRE. We show boxplots for the distribution of BMI PRS scores computed within each SIRE in ATLAS. PRS weights were computed from a prior study and are based on samples of European ancestry. Error bars were computed using 1,000 bootstrap samples.

# CHAPTER 7

# Electronic health record signatures identify undiagnosed patients with Common Variable Immunodeficiency Disease

## 7.1 Introduction

Human inborn errors of immunity (IEI), also referred to as primary immunodeficiencies (PIDs), are rare, often monogenic diseases that confer susceptibility to infection, autoimmunity, and auto-inflammation [286]. There are currently over 400 distinct IEIs and dozens more are discovered each year due to the availability of whole exome or genome sequencing[286]. One of the most common IEIs is the Common Variable Immunodeficiency (CVID) phenotype, a heterogeneous group of disorders characterized by a state of functional and/or quantitative antibody deficiency and impaired B cell responses[15, 79]. The most common clinical presentation of CVID includes recurrent sinopulmonary infections, but can also include a variety of symptoms related to autoimmunity (e.g., autoimmune hemolytic anemia) and immune dysregulation (e.g., enteritis, granulomata). The prevalence of the CVID phenotype ranges from 1 in 10,000 to 1 in 50,000 individuals worldwide[45]. Only  2,000 known cases in the United States have been identified as of 2019[310]. suggesting that between 5,000 to 33,000 patients with CVID have yet to be found in the United States alone.

Despite advances in genome sequencing technologies and the increased capacity of diagnosis for IEIs, the spectrum of genetic etiologies of the CVID phenotype is not fully understood. Over 30 genes have been implicated in CVID, but the specific genetic cause is not identified for the majority of individuals. More recently, it has even been proposed that the genetic basis of CVID can be described by a polygenic genetic architecture, where cumulative genetic effects across the genome confer disease risk[272, 224]. Because there is no clear causal genetic mechanism, there is no genetic test available for providing definitive diagnoses. Furthermore, this genetic variability leads to heterogeneous presentations of patients with CVID, making it even more difficult to diagnose. Individuals with CVID present with broad phenotypic patterns of autoimmunity and/or infection susceptibility. Since the immune system is intertwined with nearly all organs and tissues, the clinical presentation of rare immune diseases such as CVID intersects with virtually every medical specialty. This causes the fragmentation of patients across multiple clinical subspecialties, which leads to significant delays in diagnosis and treatment. This consequential delay is one of the major challenges in initiating clinical care for CVID patients, averaging 5 years in children [300] to 15 years in adults. This protracted delay in treatment increases both morbidity and mortality[275, 26, 111]. Thus, there is a highly critical and unmet need to reduce the diagnostic delay for CVID and promptly provide these patients with treatments such as immunoglobulins and immunomodulators.

The recent availability of large-scale electronic health records (EHR) has enabled the computational assessment of patients' phenotypic characteristics solely based on their medical records[31, 209, 28, 241, 164], enabling the systematic and scalable review for millions of individuals. A fundamental difficulty in this approach is having a priori knowledge about how the patterns of CVID are represented solely through EHR. We refer to these patterns describing the manifestations of CVID through EHR as the EHR-signatures of the disease. Because there is not a single clinical presentation for CVID, constructing an EHR-signature for CVID is not straightforward. In this work, we present a computational algorithm, PheNet, that computes a CVID score to prioritize patients likely to have CVID and thus candidates for immune specialist review and formal diagnoses. We leverage a high-quality, clinically curated list of CVID patients (N=197) identified within the UCLA Health System to construct a statistical model to learn the EHR-signature of CVID. Given the low prevalence of CVID and the complexities associated with diagnosing patients, this curated dataset represents one of the largest databases of CVID patients, enabling us to construct models previously not available due to the limited sample sizes of rare disease patients.

We demonstrate that PheNet attains superior accuracy versus state-of-the-art methods. We find that 57% of cases could be detected within the top 10% of all individuals ranked by PheNet in the EHR whereas previous phenotype risk scores specific to CVID[31] only capture 37% of cases, and prior genetic testing risk scores[209] only capture 23% of cases. Using EHR data from UCLA Health medical records, we show in a retrospective analysis that 64% of previously undiagnosed CVID patients at UCLA could have been identified by PheNet over 8 months before they received their initial diagnosis. We further validate our approach with a blinded clinical review from a clinical immunologist for the top 100 patients identified by PheNet out of a total of 880K individuals in the UCLA Health population. We find that 74% of individuals could be confirmed as highly probable as having CVID and specifically 8% of the top 100 could be confirmed as putatively diagnosed with CVID based on an examination of their full medical record. Taken together, EHR-based algorithms such as PheNet will expedite the diagnosis of CVID patients and will help identify novel phenotypic patterns of CVID.

## 7.2 Methods

### 7.2.1 Study population and electronic health record (EHR) data

The data for this study was extracted from the Discovery Data Repository (DDR) at the University of California, Los Angeles (UCLA). This data warehouse contains all UCLA Health patient information since the implementation of the EHR system in March 2013. The data includes various measurements and metrics such as laboratory tests, medications, billing codes, admissions, and others.

To assess the generalizability of using PheNet scores for different facilities that did not participate in model training, we conducted validation on the UC Health Data Warehouse. The clinical data of the five University of California medical centers contain EHR for 4.97

million patients. We computed the PheNet score for all of them. For computing PheNet scores, we excluded the ICD codes for CVID ( '279.06', 'D83.9', 'Z94.2', and 'Z94.3') for consistency during determining the phenotypes of patients.

### 7.2.2 CVID case definition

Central to our approach was establishing a collection of patients with a known CVID status that served as our "ground truth" cohort. These CVID cases were selected by manual chart review through the following process. First, a chart review of patients with the ICD-10 code D80.* (Immunodeficiency with predominantly antibody defects) helped broaden our ability to study immune disorders under an IRB-approved protocol. This search accumulated $3,200$ individuals who all fell under the category of "certain disorders involving the immune mechanism". Medical records were reviewed to determine the significance of those individuals with this recurring diagnosis code. This process helped eliminate patients who received an immunodeficiency code based on acute occurrences of low antibodies or for access to immunomodulatory treatments including Immunoglobulin (IVIG) without a phenotype of an immune disorder. Additionally, many patients with a cancer diagnosis were excluded based on immunosuppressive medications causing immune dysregulation. The resulting list consisted of 197 patients with CVID who can be consented to research.

### 7.2.3 Control cohort construction

We constructed a case-matched (see "CVID case definition") control cohort using the following procedure. Out of the possible N=880K patients in the UCLA Health EHR, we selected individuals based on the self-identified sex, self-identified race/ethnicity, age (closest within a 5-year window), and the number of days recorded in the EHR (closest within a 180-day window) that matched each individual in the case-cohort. For age, we used the age listed on the individuals' most recent encounters. The resulting procedure resulted in a total of N=197 cases and N=1,106 controls.

### 7.2.4 Mapping CVID clinical definition to phecodes

To represent features derived from the EHR in our model, we encoded features as phecodes using the ICD code to phecode mapping v1.2[83]. These codings represent groupings of ICD codes developed to better represent phenotypic and clinical significance from the EHR and were originally used for phenome-wide association studies. To systematically select the set of phecodes describing CVID, we utilized the entries for CVID listed in the Online Mendelian Inheritance in Man (OMIM) catalog [201] which provides clinical descriptions for thousands of rare diseases. Specifically, we selected the following OMIM ids: 607594 (CVID1), 616576 (CVID12), 614700 (CVID8), 240500 (CVID2), 615577 (CVID10), 616873 (CVID13), 613495 (CVID5), 613494 (CVID4), 617765 (CVID14), 613493 (CVID3), 613496 (CVID6), 614699 (CVID7), 615767 (CVID11). We then used a previously defined database annotating syndromes listed in OMIM with Human Phenotype Ontology (HPO) terms, a set of terms used to clinically describe human phenotypic abnormalities[153, 114]. Using this database, we were able to systematically aggregate a list of HPO terms for CVID derived from the clinical descriptions within OMIM. We then used a previously defined mapping between HPO terms and phecodes [31] to translate the list of HPO terms into a list of phecodes which could be constructed using information directly from the EHR. Altogether, this process resulted in a total of 34 unique phecodes describing CVID.

### 7.2.5 Selecting model features derived from training cohorts

In addition to using features derived from OMIM (see "Mapping CVID clinical definition to phecodes"), we also include features learned specifically from the training cohort. Although features derived from OMIM may broadly categorize the disease, leveraging information specific to the training cohort can add additional information not already encoded within OMIM. For example, there is variation in how institutions encode diagnoses within the EHR which may not be captured in all OMIM clinical descriptions. Additionally, OMIM definitions are often derived from a limited number of cases due to the rare nature of the diseases. Thus, some symptoms listed might only actually appear in a small percentage of cases since the definitions were derived from such a limited sample size. Other symptoms not currently listed in the clinical descriptions could also be indicative of the disease, but again were not formally added to the clinical definition because the symptoms did not appear in the original samples used for the OMIM description.

To select cohort-specific features, we considered all phecodes present on the medical records of individuals in the training cohort. From a possible 1,800 phecodes, we limited our selection to phecodes present in at least 2 CVID cases and excluded phecodes already selected from OMIM. We then selected the most highly enriched phecodes within the training cohort. We performed a hypothesis test testing the difference in proportions between the case and control groups for each phecode. Ranking phecodes by p-value, we selected the top K phecodes. In practice, we set K=10 but also explore alternative values.

### 7.2.6 IgG laboratory tests

The final feature included in the model is measurements of Immunoglobulin G (IgG) levels, a common type of antibody. Low IgG is a characteristic of immunocompromised individuals with diseases such as CVID. Instead of using the raw measurement as a feature directly, we convert the values to a categorical scale where the lab value is encoded as '0' if the individual has never received an IgG test, '1' if the individual has had an IgG test $\geq$ 600 mg/dL (normal range), and '2' if the individual has had an IgG test <600 mg/dL (abnormal). If an individual had multiple recorded IgG tests, we selected the lowest recorded value.

### 7.2.7 Model inference

For benchmarking experiments, we performed 5-fold cross-validation within each experiment to quantify the accuracy of various inference frameworks. To address the imbalance of cases in our dataset, we created a more balanced training dataset using random upsampling with an upsampling ratio of 0.50 and downsampling controls to N=10,000. We explored the trade-off of various upsample ratios and downsampling sample sizes. We estimated the weight of each feature using logistic regression (no penalty). We performed additional experiments to quantify performance using a variety of other inference methods, such as ridge regression, random forest, and the inverse-log frequency weighting scheme employed by PheRS. Hyperparameters utilized in the ridge regression and random forest models were selected using an additional 5-fold cross-validation step within the training step.

### 7.2.8 Comparison with previous methods

We compare PheNet with the current state-of-the-art method PheRS[31] and a related phenotype-risk score that identifies patients who would benefit from chromosomal microarray testing (CMA-score)[210]. Both methods also utilize phecodes as features. PheRS selects phecodes that correspond to the OMIM clinical description of a given disease and then computes the log-inverse frequency of the phecode measured in the general patient population. This is then used as the feature weight in the algorithm and the prediction score is a weighted sum of the weights and the presence of a given phecode, making this approach an entirely unsupervised method that does not leverage any labeled case information. To compare methods, we used PheRS weights computed using the UCLA EHR from over N=880K patients. The CMA-score framework utilizes counts of all possible phecodes and a random forest model. This method requires a training dataset composed of individuals with confirmed CMA tests which was not available to us at this time. Instead, for computing the CMA scores, we utilized pre-computed weights from the original CMA-score study conducted at Vanderbilt University. To quantify the performance of PheNet, PheRS, and the CMA-score, we compute the risk scores for all patients using each method with 5-fold cross-validation.

### 7.2.9 ICD-based diagnosis date

Although all individuals in the case-cohort were verified to have CVID, we were not able to directly obtain the exact date of diagnosis from the manually reviewed records. For those individuals for whom we could not discern an exact date of diagnosis, we used a heuristic to estimate the date of diagnosis based on occurrences of the ICD code for CVID (D83.9). We refer to this as the "ICD-based diagnosis date" to clarify that it does not constitute the precise date of a formal clinical diagnosis. However, there were 8 individuals who did not have an ICD code for CVID and thus we could not provide an estimated diagnosis date.

### 7.2.10 Assessing PheNet using retrospective EHR

We first encoded a patient's most recent visit as time point 0. We recorded the time of an encounter in a patient's medical record as the number of days before their most recent visit. This provided us a common metric of time to use when performing analyses across all patients. We computed a patient's PheNet score at 30-day intervals, spanning approximately 6 years (30 days x 12 months x 6 years). At each interval, we only considered features that were recorded up to and including that time point (and time before the given interval). To compute the score percentile for each CVID patient, we used the scores of all other patients taken at time point 0 (i.e. most recent visit) and then added the score of the single CVID patient from the designated time point. Using this distribution, we computed the score percentile for the specific CVID patient at that time point. This is then repeated for all CVID patients across all time points. Because we only had EHR from 2013, we do this to ensure that the overall distribution of scores at earlier time points is not skewed since there are many patients that do not have medical records in the electronic system at earlier time points. We then check to see if any CVID patients reached the top of the score distribution at any time point before an individual's ICD-based diagnosis (see "ICD-based diagnosis date").

### 7.2.11 Clinical validation of individuals identified by PheNet

To validate our approach, we performed a clinical chart review for the top set of individuals prioritized by PheNet. First, we removed all individuals that were deceased or who had phecodes corresponding to solid organ transplants, cystic fibrosis, or human immunodeficiency virus, resulting in the removal of 42,346 individuals. These specific disorders could lead to immunodeficiency and have a similar profile to CVID, but the cause of their immunodeficiency is already explained. We then selected the top 100 individuals identified by PheNet and a control group of 100 randomly selected individuals from the patient population. For external validation in the UC Health data warehouse, we computed the PheNet score for all patients and counted those with ICD codes for CVID ( '279.06', 'D83.9', 'Z94.2', and 'Z94.3') for consistency.

Clinical charts were directly reviewed by a clinical immunologist in a blinded review who had access to each individual's full medical record. The two lists were merged and scrambled, and the clinician was not aware of how the list of individuals was generated. Each individual was ranked according to an ordinal scale from 1 to 5 quantifying the likeliness of having CVID where 1 was defined as "near certainty not CVID" and 5 was "definitive as CVID" meaning that the individual met the criteria of a physician diagnosis.

## 7.3 Results

### 7.3.1 Summary and description of CVID cohort at UCLA

Central to our approach was training and validating our model using a data set of individuals with a known diagnosis of CVID. To first identify a smaller set of medical records to review, we restricted the search to individuals with ICD-10 code D80.* (Immunodeficiency with predominantly antibody defects) which produced approximately N=3,200 individuals within the UCLA Health system. Medical records for each individual were then manually reviewed by a clinical immunologist to determine the significance and recurring patterns associated with the specific diagnosis codes (see Methods). This process helped eliminate individuals who received an immunodeficiency code not directly associated with CVID, for example, based on acute occurrences or individuals with a cancer diagnosis who were receiving immunosuppressive medications that caused immune dysregulation. This procedure identified a cohort of N=197 individuals with a confirmed CVID diagnosis (Figure 7.1). For model training and validation, we constructed a matched control cohort based on self-identified sex, self-identified race/ethnicity, age (closest within a 5-year window), and the number of days recorded in the EHR (closest within a 180-day window) (see Methods), resulting in a total of 197 cases and 1,106 controls (Figure 7.2).

The resulting CVID cohort was 71.6% female, and the average age was 55.4 (SD: 19.5). Previously constructed cohorts had a very similar demographic profile that also showed an increased female predominance[95]. Using the ICD-based diagnosis dates (see Methods), we found that the average age of individuals at diagnosis was 55.0 (SD: 19.0) years old which is consistent with a large portion of individuals being diagnosed with CVID after age 40 [26]. To evaluate the extent of patients' immune dysregulation, we assess patients' immunoglobulin G (IgG) levels, a common antibody that is typically low in CVID patients[45]. Within the case-cohort, we found that 86.3% of individuals had at least one IgG laboratory test and 37.1% had at least one abnormally low IgG laboratory test result (<600 mg/dL). Out of the control cohort, we found that only 3.2% of individuals had an IgG test and only 0.45% with an abnormal result which is consistent with the majority of individuals not likely having any immunodeficiencies. Many individuals had extensive medical history at UCLA Health where the average extent of EHR data was 14.7 years (SD: 7.9), and the average number of unique ICD-10 codes per individual was 95.8 (SD: 95.7) (Table 7.1). However, there were 6 individuals out of the 197 CVID cases who had fewer than 10 encounters within the EHR from UCLA. This could reflect individuals who came to UCLA only to receive a formal diagnosis or those who only came for a second opinion, but largely had their care at a different medical center.

### 7.3.2 Constructing a CVID risk score model from EHR-derived phenotypes

We used the curated set of cases to learn the EHR-signature for CVID as follows. For features, we used phecodes[85](codes derived from ICD codes used to represent clinically meaningful phenotypes from the EHR) and laboratory measurements of immunoglobulin G (IgG) (see Methods). To prevent overfitting, we selected a subset of phecodes (out of possible $1,800$ codes) that best captured the phenotypic patterns of CVID. We first selected phecodes matching the clinical description of CVID listed in the Online Mendelian Inheritance in Man (OMIM)[201] database for a total of 34 phecodes. Then, leveraging the annotated data specifically for this study, we included the top K significant phecodes with a significantly higher frequency in the cases as compared to the controls (see Methods). To prevent biases, we excluded the actual phecode for CVID itself (279.11) from the set of features. Varying the value of K controlled the tradeoff between adding more information to the model and overfitting due to the increased data dimensionality. We found that $K = 10$ provided a high level of performance while also preventing overfitting, and we used this parameter in all subsequent analyses. In addition to the set of selected phecodes, we included the laboratory test for immunoglobulin G (IgG) levels as that gives a proxy for the immune state of the patient. Because we are only interested in capturing whether patients have had a test and then if this result was abnormally high, we discretized laboratory measurements as a categorical variable where patients' lab result was either normal (IgG >600 mg/dL), abnormal (IgG <600 mg/dL), or no IgG test was recorded in their medical record.

We next compared a variety of prediction methods to learn a function that best mapped the feature set to each individual's CVID status. We evaluated various methods that varied in model complexity, including linear methods such as marginal logistic regression of each feature, penalized joint models like ridge regression, as well as non-linear methods such as random forest regression. We found that marginal regression and ridge regression achieved similarly high performance (AUC-ROC/PR(marginal): 0.95/0.83, AUC-ROC/PR(ridge): 0.96/0.88). We opted to use marginal regression to maintain the most straightforward interpretability of the regression coefficients. We additionally assessed the performance of the model when including IgG information and found that the inclusion of this single feature added a substantial increase in performance (AUC-ROC/PR(IgG): 0.95/0.83, AUC-ROC/PR(no-IgG): 0.89/0.73). To account for the severe case imbalance associated with predicting rare diseases, we performed random upsampling of the cases to achieve a more balanced dataset. Comparing various upsampled ratios, we found that a ratio of 0.50 provided optimal performance. Our final prediction model included the 34 phecodes selected from OMIM, the K=10 phecodes learned from the case-cohort, and the IgG laboratory test with an upsampling ratio of 0.50. Using 5-fold cross-validation, we showed that the average PheNet scores for individuals with CVID had a significantly higher risk score than the matched controls (Cochran-Armitage Test test: p-value $< 2.2 \times 10^{-16}$). We emphasize that this risk score does not quantify the risk of an individual developing CVID in the future, but instead assesses whether or not the patient likely already has CVID at the present time (but has just not yet been diagnosed).

### 7.3.3 PheNet is more accurate than existing phenotype risk scores for predicting CVID

Next, we compared PheNet against PheRS, an unsupervised risk score designed for identifying undiagnosed patients with rare diseases [31], and the chromosomal microarray (CMA) risk score, a method designed to predict patients who would benefit from CMA tests for diagnoses[209]. The model for the CMA risk score was pre-computed from the Vanderbilt EHR since re-training the model would require constructing a training dataset of individuals with validated CMA tests which we did not have at this time. However, because PheRS is an entirely unsupervised method, we were able to re-train the model for CVID prediction within the UCLA EHR. We found that PheNet performed 17% better than PheRS when comparing AUC-ROC and 42% better when comparing AUC-PR. In comparison to the CMA test, PheNet performed 30% and 66% better in terms of AUC-ROC and AUC-PR (Figure 7.3A, B). In practice, only individuals with very high-risk scores would be candidates for follow-up. Setting a threshold score of 0.90, we found that 57% of cases could be detected within the top 10% of individuals ranked by PheNet score (Figure 7.3C). In contrast, PheRS and the CMA-score only captured 37% and 23% of cases at the same threshold.

The improvement of PheNet over the CMA-score is most likely because the CMA-score was developed for a highly-related but broader problem such as the identification of individuals that could benefit from CMA testing for diagnosis. The genetic cause for CVID is not found for over half of the patients, suggesting that even performing a CMA test would not likely identify CVID patients in the majority of cases. Additionally, deviations in performance could also be due to the fact that the CMA model was trained in EHRs from a separate institution. However, the CMA model was previously tested in out-of-hospital populations and showed little decrease in performance. To further investigate the potential bias due to EHR from different institutions, we compared the accuracy of PheRS using a model that was trained in the Vanderbilt EHR. Because the PheRS method is entirely unsupervised, we could do a systematic performance comparison when using models trained at UCLA and Vanderbilt (VU). We found that the pre-trained feature weight for the phecode 561.1 (Diarrhea) was not available in the Vanderbilt EHR, thus we excluded this feature from both models for this set of analyses. We found that the AUC-ROC and AUC-PR were almost exactly the same (PheRS-UCLA: $0.79, 0.48$; PheRS-VU: $0.79, 0.49$), suggesting that the EHR-signature for CVID is very similar between the institutions and not likely a major source of bias.

We performed additional analyses to assess whether the performance of PheNet was artificially biased because scores were computed for individuals based on EHR information obtained both before and after their diagnoses. To test whether a more temporally restricted set of EHR data could still have appropriate predictive power, we created a "censored" testing dataset that limited each individual to only information present in the medical record prior to an individual's "ICD-based diagnosis" for CVID. Because we do not have access to the exact date of patients' formal diagnoses (due to a date shift in the EHR not present in the manually reviewed medical records), we estimate the date based on the occurrences of the ICD-10 code for CVID (D83.9) within the EHR. We clarify that the cohort of CVID patients was formally identified through manual chart review and that this ICD-based procedure was only used to identify the approximate date of diagnosis within the EHR (see Methods). The training dataset was still trained on all data points up to the present regardless of the diagnosis date as this does not affect test performance. Using this more restricted test set, we found a modest reduction in performance, but we were still able to capture a large percentage of CVID patients. Specifically, we found 46% of cases compared to 5% of cases within the top 10% of patients ranked by PheNet. When comparing AUC-ROC and AUC-PR, we see a 17.7% and 51.7% decrease in performance. This drop in performance could be because some patients do not have substantial medical history at UCLA preceding their diagnosis which would drastically limit the prediction power. When limiting our assessment to CVID patients with at least 1 year of EHR data before their diagnosis (N=58), we find only an 8.1% and 44.6% drop in performance for AUC-ROC and AUC-PR respectively, suggesting that with adequate medical history, there is limited performance bias when using all EHR information up to the present.

### 7.3.4 PheNet identifies new patients with CVID in real data

We next sought to quantify the utility of PheNet as a predictive tool for identifying whether patients could be diagnosed earlier by conducting an analysis using the UCLA EHR data from over 880K individuals at UCLA. The dataset comprised all individuals within UCLA Health with at least one encounter and at least one ICD code, for a total of 880K individuals (as of 2019), including our previously established case-cohort of 197 CVID cases. Using 5-fold cross-validation, we divided the data into 80% training set and 20% testing set folds and ran PheNet on each fold of the data (see Methods). To mirror how PheNet would be used in practice, we limited the testing dataset to only features that appeared in the EHR prior to an individual's ICD-based diagnosis. For different scoring thresholds, we captured CVID patients at various times both before and after their diagnoses. In practice, the score threshold could be chosen according to a specific goal and the amount of resources available. For example, one would recommend using a high stringency score threshold, thus capturing fewer individuals, if patients were to be followed up individually which is a resource-expensive undertaking.

To ensure individuals had an adequate amount of medical history prior to their diagnosis of CVID, we restricted the analysis to individuals with at least one year of EHR data prior to their ICD-based diagnosis ($N = 58$). We set a threshold PheNet score of 0.9 for this analysis and found that PheNet identified 64% of individuals with CVID before their ICD-based diagnosis (Figure 7.4). The median number of days between the date identified by PheNet and the date of diagnosis was 244 days (SD: 374). For example, the individual shown in Figure 7.5 reached the top percentile of the PheNet score distribution 41 days before their record reflected any immunodeficiency diagnosis codes. This patient had accumulated 7 phecodes that influenced their PheNet score in the years prior to diagnosis. Then the patient's record revealed measurement of a modestly low IgG level, which further increased their risk score. This example demonstrates the advantage of aggregating information from both phenotypes and laboratory tests to identify individuals as high-risk. These results show that PheNet has substantial utility for not only identifying undiagnosed individuals with CVID but also bringing them to a diagnosis earlier than they might have in a usual clinical scenario.

### 7.3.5 Clinical validation of identified undiagnosed individuals with CVID

To validate the utility of PheNet for identifying new patients with CVID, we conducted an analysis using the UCLA EHR data from over 880K individuals at UCLA as the discovery cohort. We removed from consideration all individuals who were deceased or who had phecodes corresponding to solid organ transplants, cystic fibrosis, or infection with the human immunodeficiency virus, resulting in the removal of 42,346 individuals. Individuals with these disorders may exhibit similar clinical profiles as those with CVID, but their phenotypes are likely due to their immunocompromised conditions, not a primary genetic disease. We then selected the top 100 individuals identified by PheNet and a control group of 100 randomly selected individuals from the patient population. On average, the group of top 100 individuals had an average of 15.5 years of medical history and the randomly selected group had 7.1 years.

We scrambled these two sets of patients and performed a clinical chart review for these individuals (see Methods). Medical records were directly examined by a clinical immunologist who was blinded to the groups and not informed that they were validating a risk score algorithm for CVID. The clinician had access to each individual's full medical record including notes, images, and scanned documents, which were not available to the PheNet algorithm. Each individual was ranked according to an ordinal scale from 1 to 5 quantifying the likeliness of having CVID where 1 was defined as "near certainty not CVID" and 5 was “definitive as CVID” meaning that the patient meets the criteria of a physician diagnosis. From the list of top 100 ranked individuals, 74% of individuals were assigned a score of 3, 4, or 5, indicating that they were highly probable as having CVID (Figure 7.6). Specifically, 8% of individuals were assigned a score of 5, meaning that they were positively diagnosed with CVID by having low immunoglobulin levels and poor humoral responses to vaccine antigens or having a prior outside physician diagnosis of CVID. In contrast, the individuals who were randomly chosen exclusively had scores of 1, 2, or 3, and 90% of individuals had a score of 1 or 2, indicating that they likely did not have CVID. Overall, these results validate that our approach is useful to identify patients with CVID and overcome the major challenge of initiating care in a timely manner. The reduction of delays in diagnosis will enable patients to seek appropriate medical care to reduce morbidity and mortality.

In addition to prediction performance, it is also important to understand the symptoms that contributed to each individual's increased risk status. In practice, it would not be sufficient to only identify individuals to refer to an immunology clinic, but it is also necessary to explain exactly which factors contributed to their identification. Examining the regression coefficients from the model in the form of odds ratios, we can identify phenotypes that were most predictive (Table 7.2). We find that some of the most predictive features (e.g. primary thrombocytopenia) were not provided from the OMIM clinical description but were from the set of enriched phecodes identified from the case-cohort, further emphasizing the benefit of including a well-curated case-cohort in the prediction model. The signs and symptoms that contributed to each of the top 100 individuals' risk scores are shown in Figure 7.7. Overall, there are wide variations in the symptoms of each individual, demonstrating the utility of methods that aggregate both numerous symptoms and laboratory results to identify patients at risk. There is no single feature present in all 100 individuals with the highest PheNet scores, underscoring the lack of any single clinical manifestation as being pathognomonic of CVID. We also observed that the majority of individuals had a mixture of both autoimmune and infection-related phecodes, further demonstrating the heterogeneity of the CVID phenotype. These patterns were consistent with those observed in the cohort of formally diagnosed

CVID patients. In contrast, the majority of randomly selected individuals did not have any major symptoms matching the patterns of CVID estimated by PheNet, and the signs and symptoms present within this group were those that were among the most common in the general population such as upper respiratory infections and asthma.

### 7.3.6 Validation on the University of California wide data warehouse

For general applicability, we next tested whether PheNet could be applied to new databases. We validated the generalizability of PheNet using de-identified clinical data collected from the University of California medical centers that include (a) University of California Los Angeles; (b) University of California San Francisco; (c) University of California Davis; (d) University of California San Diego (but not Rady Children's Hospital); and (e) University of California Irvine ($>$ 4.9 million patient records, https://www.ucbraid.org/cords (Table 7.3). We scored and ranked each individual in the UC-wide data set using the PheNet weights calculated from UCLA data as above (i.e., no training was performed on the UC-wide data). To assess the utility of the scores, we asked whether PheNet could identify patients who had at least one encounter with a diagnosis code of CVID (ICD code 'D83.9') (N$= 1,838$ out of $>$ 4.9M patient records). When ranking patients by PheNet scores, we found a striking enrichment of patients with a diagnosis code of CVID in the top-ranked patients. For example, among the top $10,000$ patients ranked by PheNet score, we found 44% for UCLA to 64% for UCD of all patients with a CVID diagnosis code among more than 2 million patient records. A random ranking of patients would find less than 6 patients from each clinical site in the top 10,000 patients. This result demonstrates enrichment of CVID cases among those with high PheNet scores and showcases the power of this approach to prioritize patients suspected of CVID for follow-up analyses. Taken together, these results confirmed that PheNet maintains robust interoperability with new databases and new data formats after training PheNet with UCLA's CVID patients.

## 7.4 Discussion

In this work, we identified phenotypic patterns of CVID, or EHR-signatures, encoded in patients' medical records and trained an algorithm to identify patients who likely have CVID but who have been otherwise "hiding" in the medical system. Due to the heterogeneity of clinical presentations for IEI phenotypes, CVID patients can initially present to a wide range of clinical specialists who focus on the specific organ system involved (e.g., the lung) rather than directly to an immunologist for the underlying immune defect. This organ-based approach of our current healthcare system can result in tunnel vision and hinder a formal diagnosis in IEI, particularly for those patients who have multi-system manifestations that fluctuate over time. As a result, these patients face a diagnostic delay of 10 or more years. Each year of delay in the diagnosis of CVID results in an increase in infections, antibiotic use, emergency room visits, hospitalizations, and missed days of school and work totaling over USD $108,000 compared to the year after the diagnosis of CVID is made (in 2011 dollars, which is approximately $145,000 in 2022 dollars) [207]. The diagnostic delay for adults with CVID ranges between 10-15 years(Slade et al. 2018b), suggesting that $1M or more per CVID patient is being misdirected by the US healthcare system because of diagnostic delays. Considering that there may be $\sim 10,000$ or more individuals currently waiting to be diagnosed with CVID, the aggregate impact on the US health system is billions of dollars due to a failure to diagnose CVID in a timely fashion. Beyond the economic impact, the non-quantifiable impacts on patients' lives due to diagnostic delays are even more significant. For example, previous studies have shown that undiagnosed patients suffer from anxiety and depression as they undergo costly tests and specialty visits[207].

Prompt identification of patients who may have IEIs by primary care providers is paramount to reduce the risk of irrevocable sequelae of invasive infections, such as bronchiectasis, encephalitis, or kidney failure. A number of efforts have attempted to codify a set of "warning signs" that offer guidance to primary care doctors. Most recognizable are the "10 Warning Signs" that have been widely disseminated by the Jeffrey Modell Foundation for two decades. Before EHRs, the broad phenotyping necessary to assemble a proper picture of heterogeneous IEIs like CVID was not possible, and so guidelines had to be developed by committees and expert opinions. These warning signs largely emphasized infections as a core feature of IEIs. Our results suggest that phenotypes of inflammation, autoimmunity, malignancy, and atopy should also be included. Indeed, two analyses found these 10 warning signs were unable to identify many subjects with known IEIs[187, 225], possibly because phenotypes aside from infections were missing. When the warning signs were applied to adults versus children with known IEIs, adult patients were often missed (45% sensitivity for adults versus 64% for children)[41], suggesting the need to modify assessments based on age. In other studies, the need for intravenous antibiotics, failure to thrive, or relevant family history was found to be the only strong predictors of IEIs[243, 279]. An algorithm developed by the Modell Foundation improved IEI diagnoses by using a summation of diagnostic codes[208] and another recent algorithm that summed weighted ICD codes further improved diagnoses[247]; however, these approaches did not include laboratory values. Retrospective gathering of features as we performed here has been useful in aggregating the phenotypic features of patients with IEIs into a score that can discern those with IEIs from those with secondary immunodeficiencies [294]. Recent work used a Bayesian network model to score "risk" in a framework that categorized individuals into either high, medium, or low-risk categories of having any IEI[247]. Their approach also classified each patient into a likely IEI categorization (e.g., combined immunodeficiency, antibody deficiency, etc.). One limitation in Bayesian analyses is in the assessment of probabilities (and conditional probabilities) for rare events; this concern was partially alleviated by employing a large cohort of children with known IEIs. But as a result, that work suggested that 1% of all patients were at medium-to-high risk of IEI, potentially overestimating the true prevalence by 100-200 fold. That work highlights one of our limitations, too, that ascertaining a proper threshold for risk scores is fraught. Regardless, these efforts showcase both the potential and the unmet need for identifying previously undiagnosed patients in large healthcare systems.

There are several inherent limitations to our study. The prediction algorithm is derived primarily from ICD codes within the EHR. Although ICD codes represent an international standard, the specific patterns of assigning ICD codes can often vary across physicians and institutions[223, 125]. We overcame this concern by employing phecodes, a generalization of phenotypes derived from ICD codes and better suited for EHR research [318]. However, even using phecodes requires a careful examination of their level of descriptive granularity. For example, one clinical description for CVID in OMIM includes hypothyroidism as a potential phenotypic feature. Accordingly, we utilized the phecode for "Hypothyroidism" (phecode 244.2) in the prediction model. However, no individuals within the CVID cohort at UCLA had this phecode within their medical records. Upon further inspection, we found that this symptom was instead attached as the phecode under "Hypothyroidism NOS" (phecode 244.4). The lesson was that many phecodes under 244.X could equally apply, and that small deviations in diagnosis coding practices could have a large impact on algorithmic outputs. We ameliorated these deviations by not only utilizing symptoms provided in OMIM, but also by learning important model features directly from the training data.

Another limitation of our work was the amount of longitudinal information available in the EHR. Patients move frequently and obtain care from a variety of settings (private practices, urgent care clinics, in addition to large health systems); consequently, only a subset of their data are contained in the health system's database. Because EHR vendors change with regular occurrence, many EHRs hold only a maximum of $5 - 15$ years worth of data, which may not be enough to fully glean the necessary details of a patient's health trajectory. We also did not consider the number of times a specific diagnosis appeared nor the order that the phecodes appeared on the medical record. Since CVID is characterized by recurrent infections, we believe that longitudinal information of multiple occurrences would increase the specificity of the model by disregarding individuals with single acute diagnoses. We also did not restrict the types of encounters when collecting the diagnosis codes (e.g., hospitalization, emergency department, or outpatient clinic). Annotations of past and present ICD codes vary considerably across these settings. Instead, we wanted to use as much information as possible to increase the power of our model. However, limiting diagnoses that occur specifically during appointments or hospital visits (as opposed to laboratory tests) could also increase specificity and better differentiate individuals with other immunocompromising conditions (e.g. cancer). We hope to develop these extensions in future work.

Genetic sequencing in patients with immunodeficiency can alter disease management, treatments, and clinical diagnosis[278]. Our approach can be used to expedite the referral of patients to immunologists and to support the need for genome sequencing. By broadening the base of patients studied and their phenotypes, such efforts should expand our understanding of the immunogenetic basis of antibody deficiencies like CVID. In the future, we want to investigate the variants associated with the highest risk scores. The impact of our work will greatly benefit the IEI community as there is an urgent need for more systematic, resource-efficient ways to identify and categorize patients with IEIs.

## 7.5 Tables

| **feature** | **OMIM** | **phenotype** | **freq(all)** | **freq(cases)** | **OR** |
|---|---|---|---|---|---|
| IGG | | Igg <600 | NA | NA | 9.64 |
| 279.1 | | Immunity deficiency | 41.62% | 0.48% | 42.84 |
| 475 | x | Chronic sinusitis | 47.72% | 4.45% | 5.52 |
| 495 | x | Asthma | 42.13% | 9.56% | 2.14 |
| 136 | | Other infectious and parasitic diseases | 34.52% | 3.13% | 5.48 |
| 496.3 | x | Bronchiectasis | 23.35% | 0.56% | 20.00 |
| 472 | | Chronic pharyngitis and nasopharyngitis | 24.37% | 1.96% | 6.62 |
| 709.7 | | Unspecified diffuse connective tissue disease | 14.21% | 0.51% | 13.09 |

| | | | | | |
|---|---|---|---|---|---|
| 283 | x | Acquired hemolytic anemias | 4.57% | 0.16% | 28.01 |
| 287.31 | | Primary thrombocytopenia | 9.64% | 0.24% | 19.16 |
| 579.2 | x | Splenomegaly | 8.12% | 1.02% | 3.91 |
| 502 | | Postinflammatory pulmonary fibrosis | 15.23% | 0.79% | 10.34 |
| 31 | | Diseases due to other mycobacteria | 4.57% | 0.13% | 17.92 |
| 289.4 | x | Lymphadenitis | 17.77% | 3.24% | 2.85 |
| 41 | x | Bacterial infection NOS | 12.18% | 2.86% | 2.29 |
| 264.2 | x | Failure to thrive (childhood) | 2.54% | 0.39% | 2.95 |
| 696.4 | x | Psoriasis | 5.08% | 1.33% | 2.18 |
| 504 | | Other alveolar and parietoalveolar pneumonopathy | 9.64% | 0.52% | 7.21 |
| 381.1 | x | Otitis media | 4.06% | 1.92% | 1.13 |
| 561.1 | x | Diarrhea | 23.86% | 6.08% | 1.80 |
| 496 | x | Chronic airway obstruction | 13.71% | 2.16% | 3.19 |
| 480 | x | Pneumonia | 25.38% | 4.67% | 2.63 |
| 283.1 | x | Autoimmune hemolytic anemias | 3.55% | 0.06% | 18.09 |
| 279.8 | | Other specified disorders involving the immune mechanism | 4.06% | 0.07% | 11.98 |
| 202.2 | x | Non-Hodgkins lymphoma | 4.57% | 0.74% | 2.88 |
| 497 | x | Bronchitis | 12.18% | 2.60% | 2.32 |
| 255.2 | x | Adrenal hypofunction | 0.51% | 0.01% | 26.00 |

| | | | | | |
|---|---|---|---|---|---|
| 279.2 | | Autoimmune disease NEC | 6.09% | 0.15% | 15.57 |
| 555 | x | Inflammatory bowel disease and other gastroenteritis and colitis | 0.00% | 0.00% | 1.00 |
| 535 | x | Gastritis and duodenitis | 0.00% | 0.00% | 1.00 |
| 244 | x | Hypothyroidism | 0.00% | 0.00% | 1.00 |
| 686 | x | Other local infections of skin and subcutaneous tissue | 8.63% | 2.52% | 1.88 |
| 264 | x | Lack of normal physiological development | 0.00% | 0.00% | 1.00 |
| 716 | x | Other arthropathies | 0.00% | 0.02% | 0.00 |
| 253.5 | x | Pituitary dwarfism | 1.52% | 0.10% | 7.37 |
| 555.2 | x | Ulcerative colitis | 3.55% | 0.56% | 2.55 |
| 320 | x | Meningitis | 1.52% | 0.32% | 1.77 |
| 251.1 | x | Hypoglycemia | 2.54% | 0.72% | 1.98 |
| 369.5 | x | Conjunctivitis, infectious | 6.60% | 3.08% | 1.14 |
| 573.3 | x | Hepatomegaly | 6.09% | 1.43% | 2.20 |
| 284.1 | x | Pancytopenia | 4.57% | 0.81% | 3.23 |
| 704.1 | x | Alopecia | 4.57% | 2.11% | 1.13 |
| 287.3 | x | Thrombocytopenia | 11.17% | 2.37% | 2.74 |
| 315 | x | Develomental delays and disorders | 1.02% | 1.14% | 0.47 |

| | | | | | |
|---|---|---|---|---|---|
| 465 | x | Acute upper respiratory infections of multiple or unspecified sites | 23.86% | 12.86% | 0.92 |

Table 7.2: PheNet algorithm features with the frequency in the UCLA patient population and the CVID cohort. Odds ratios are reported in the last column.

| | Case | Control |
|---|---|---|
| Sample size | 197 | 1,106 |
| Age (years) | | |
| Mean (s.d.) | 55.35 (19.53) | 54.20 (20.26) |
| Median | 60.00 | 57.00 |
| Sex (%) | | |
| Male | 28.43 | 30.81 |
| Female | 71.57 | 69.19 |
| Mean number of unique ICD codes | 95.76 (95.67) | 69.00 (95.67) |
| Mean medical record length years | 14.72 (7.93) | 14.27 (7.89) |
| IgG laboratory test | | |
| No test | 12.7% | 96.8% |
| IgG >600 mg/dL (normal) | 49.2% | 2.7% |
| IgG <600 mg/dL (abnormal) | 37.1% | 0.45% |

Table 7.1: Demographics of CVID case/control cohorts. We show a summary of the individuals in both the CVID case (N=197) and control cohorts (N=1,106) including age, sex, number of unique ICD codes, number of years recorded in the EHR, and immunoglobulin G (IgG) laboratory tests. If patients had more than one IgG test, the lowest value was used.

| | **UCLA** | **UCSF** | **UCD** | **UCSD** | **UCI** |
|---|---|---|---|---|---|
| **Total patients** | 1,642,284 | 1,222,352 | 794,585 | 733,128 | 579,360 |
| **Total patients with CVID ICD code** | 1,018 | 580 | 355 | 98 | 372 |
| **PheNet top 100** | 10<br>(1.72%) | 10<br>(2.82%) | 6<br>(6.12%) | 15<br>(4.03%) | 17<br>(3.93%) |
| **PheNet top 1,000** | 63<br>(10.9%) | 57<br>(16.1%) | 32<br>(32.7%) | 53<br>(14.3%) | 80<br>(18.5%) |
| **PheNet top 5,000** | 178<br>(30.7%) | 131<br>(36.9%) | 50<br>(51.0%) | 139<br>(37.4%) | 206<br>(47.6%) |
| **PheNet top 10,000** | 255<br>(44.0%) | 184<br>(51.8%) | 63<br>(64.3%) | 188<br>(50.5%) | 272<br>(62.8%) |

Table 7.3: Enrichment of patients with a CVID diagnosis code identified within the UC-Wide Data Warehouse. The number of patients with the CVID ICD-10 code (D83.9) in each top PheNet scored cohort across the UC-Wide Data Warehouse.

## 7.6 Figures

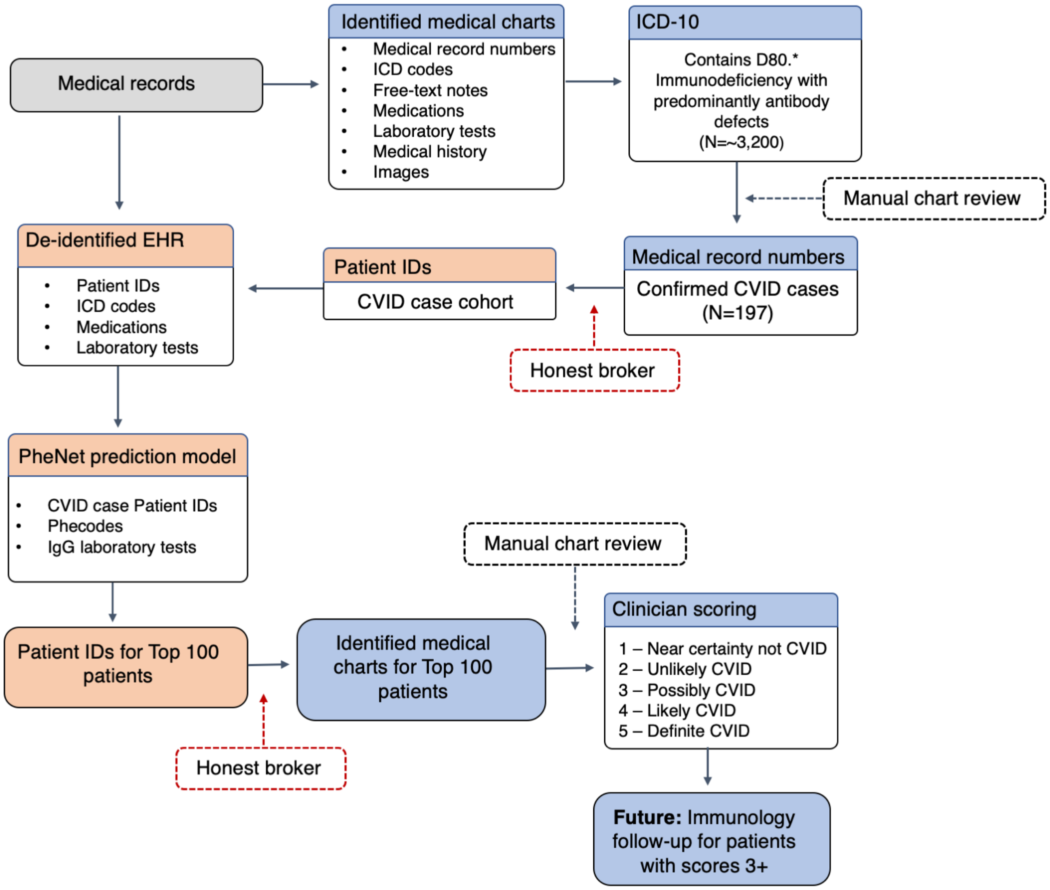

Figure 7.1: Overview of CVID cohort curation and new CVID patient identification. We provide a flowchart describing the EHR review process for constructing a well-curated list of clinically diagnosed patients with CVID. We then demonstrate how this cohort is used for training a prediction model which is then used to identify undiagnosed CVID patients in a discovery cohort. A manual chart review is performed on the patients with the highest risk score with the future goal of highly probable CVID patients being referred to an immunologist.

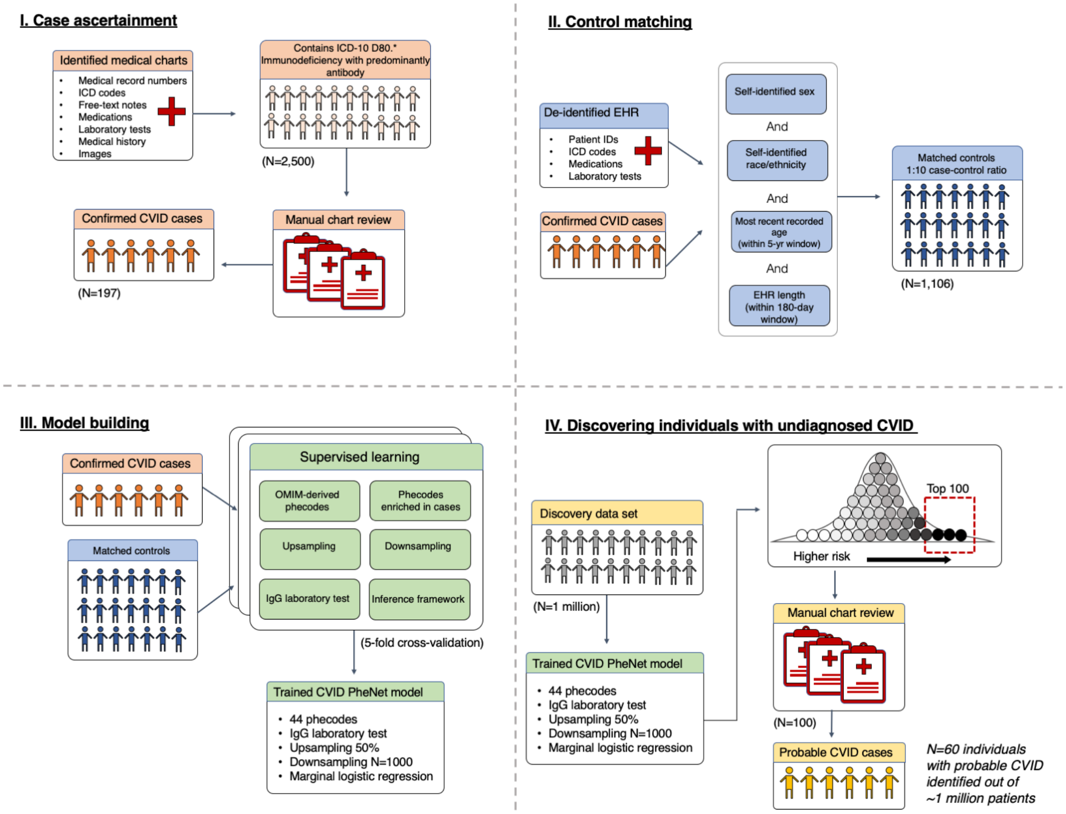

Figure 7.2: Overview of PheNet model training and application within a discovery cohort. We present a visual summary of case/control cohort construction, PheNet model training, and application within a discovery dataset at UCLA Health. In (I) we show the workflow for constructing a case-cohort of clinically diagnosed patients with CVID from medical charts (N=197). (II) shows the criteria used to create a matched control cohort from the EHR (N=1,106). (III) visually summarizes the construction of a prediction model, including feature selection from phecodes, the inclusion of laboratory values, a variety of inference frameworks, and data balancing techniques. Finally, (IV) demonstrates how the PheNet model can be applied within a discovery cohort to identify patients with a high likelihood of CVID who can then be further assessed by manual chart review to confirm diagnosis.

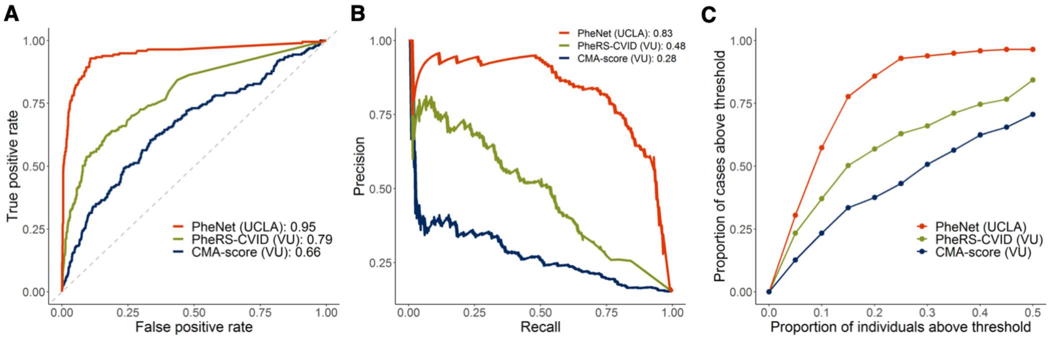

Figure 7.3: PheNet is more accurate than existing phenotype risk scores for predicting CVID. Performance metrics comparing the performance of PheNet, PheRS-CVID, and the CMA-score within the case (N=197) and control (N=1,106) cohorts from the UCLA Health population. The CMA-score was computed using weights pre-trained from data from Vanderbilt (VU); PheNet and PheRS-CVID were trained using weights trained from EHR data at UCLA. (A) and (B) show the receiver operating characteristic and precision-recall curves across the different prediction models. AUC is provided in parentheses in the legend. In (C), we display a curve showing individuals with a PheNet score ¿ 0.90 and the proportion of CVID cases captured within the varying percentiles of PheNet scores.

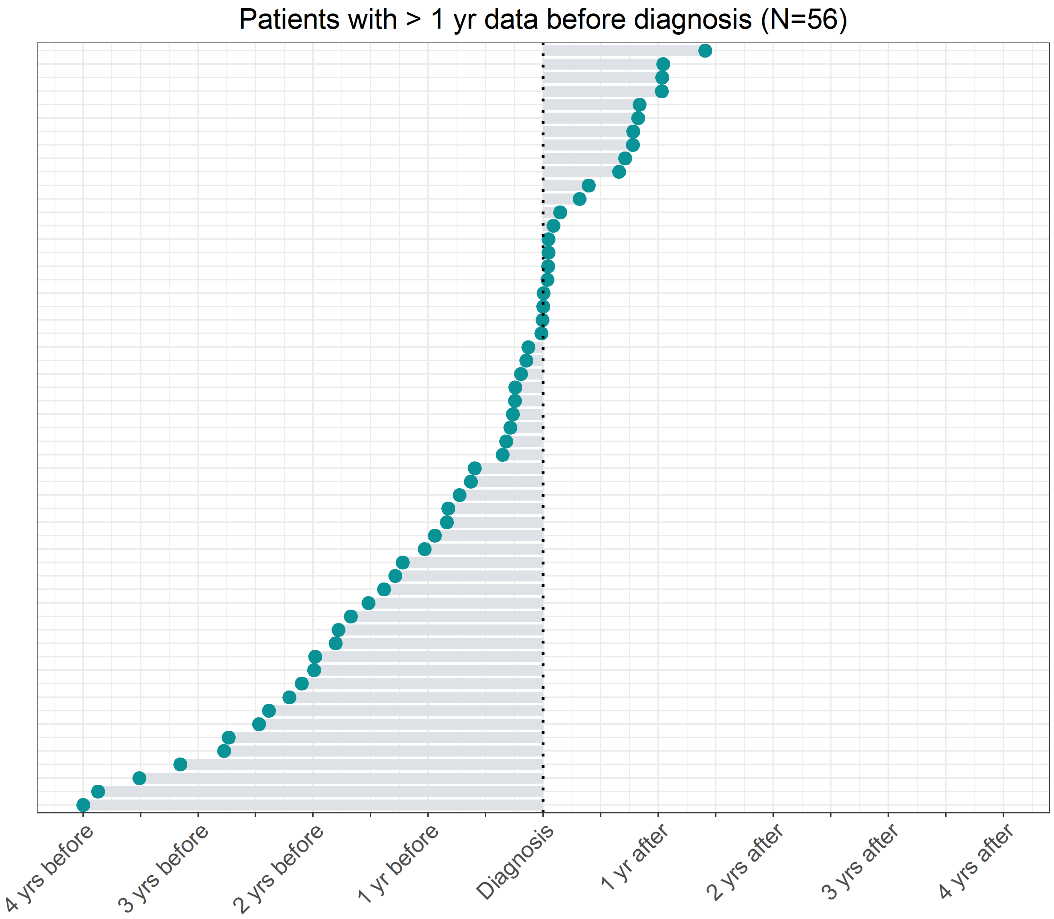

Figure 7.4: PheNet identifies CVID patients before their original diagnosis dates. Distribution of the time between individuals' ICD-based diagnoses for CVID and the time point at which individuals' risk score ¿ 0.90 (denoted at the blue circles). Only individuals with at least 1 year of EHR data prior to their ICD-based diagnosis were included. Two individuals were excluded from the graph because they did not meet the score threshold at any point in time for a total of N=56 individuals shown. ICD-based diagnoses were determined as the time point when individuals first accumulate at least two CVID ICD codes (D83.9) within a year.

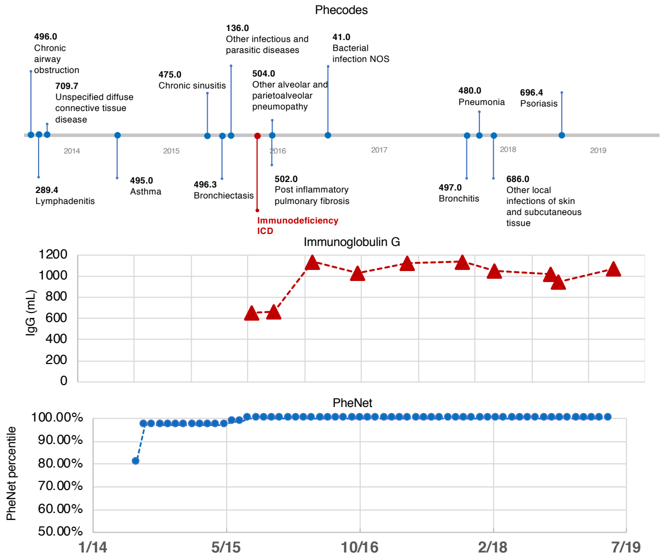

Figure 7.5: Sample patient's CVID timeline. The top panel lists all CVID-relevant phecodes on a sample patient's record. The point when the patient received their first immunodeficiency billing code is denoted by the red star. The middle panel shows the patient's immunoglobulin G (IgG) laboratory results over time, where a value ¡ 600 mg/dL is considered abnormal. The bottom panel shows the percentile of the patient's risk score computed over time. Specifically, we show that the patient reached the 99th percentile of the PheNet score distribution 41 days before their medical record showed evidence of specific immunodeficiency care. Note that the patient's timeline has been date-shifted.

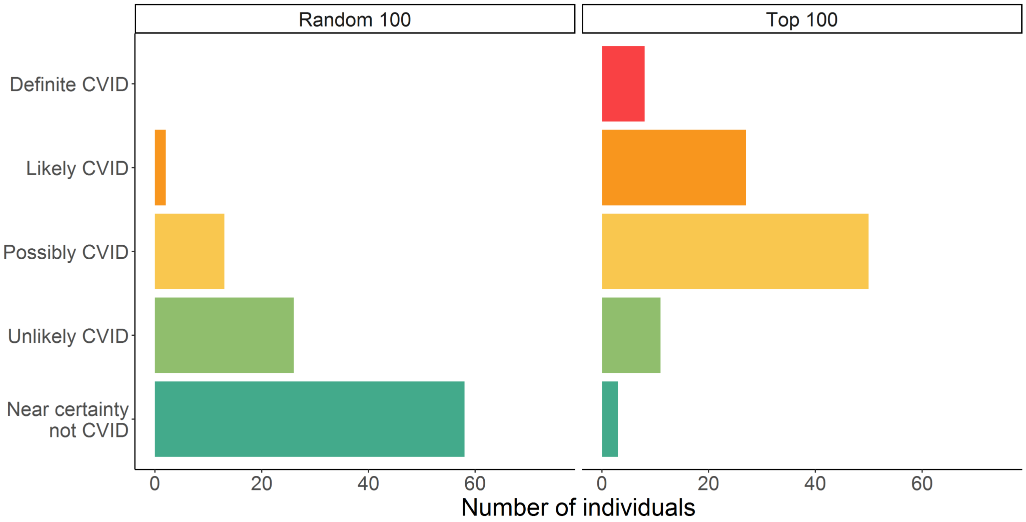

Figure 7.6: PheNet identifies undiagnosed individuals with CVID. We show the CVID clinical validation scores for the top 100 individuals with the highest PheNet score and 100 randomly sampled individuals. Each individual was ranked according to an ordinal scale from 1 to 5 quantifying the likeliness of having CVID where 1 was defined as "near certainty not CVID" and 5 was "definitive as CVID".

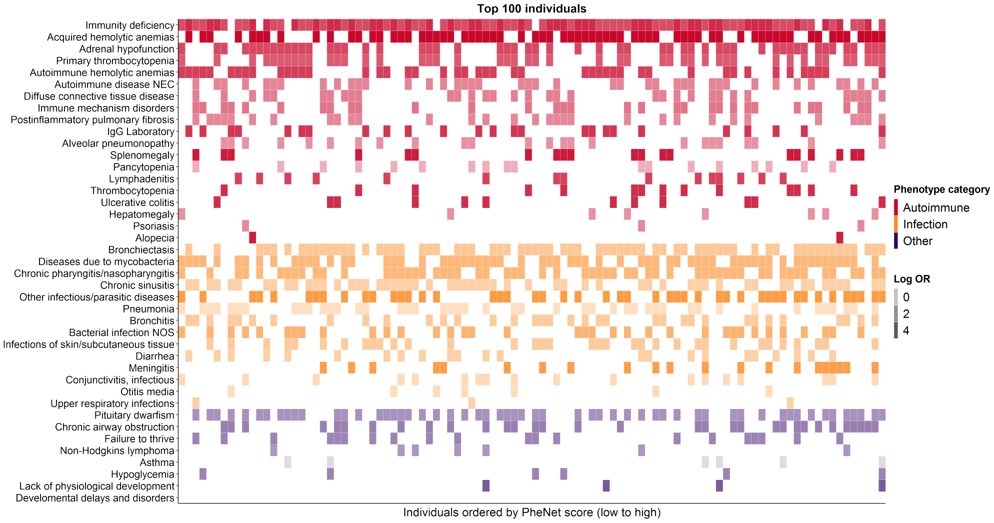

Figure 7.7: PheNet identifies undiagnosed individuals with CVID. We show the CVID clinical validation scores for the top 100 individuals with the highest PheNet score and 100 randomly sampled individuals. Each individual was ranked according to an ordinal scale from 1 to 5 quantifying the likeliness of having CVID where 1 was defined as "near certainty not CVID" and 5 was "definitive as CVID".

# References

[1] Home. https://www.qatarbiobank.org.qa/. Accessed: 2023-3-3.

[2] Infinium Global Screening Array-24 Kit.

[3] Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. Publication Title: The White House.

[4] welcome to emerge ¿ collaborate. https://emerge-network.org/. Accessed: 2023-2-26.

[5] White house precision medicine initiative. https://obamawhitehouse.archives.gov/node/333101. Accessed: 2023-3-3.

[6] *Ethnic Disparities in the Burden and Treatment of Asthma*. 2005.

[7] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[8] Gad Abraham, Yixuan Qiu, and Michael Inouye. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics*, 33(17):2776–2778, September 2017.

[9] Noura S Abul-Husn and Eimear E Kenny. Personalized medicine and the power of electronic health records, 2019.

[10] Adewole S Adamson and Avery Smith. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol.*, 154(11):1247–1248, November 2018.

[11] Aya J Alame, Sonia Garg, Julia Kozlitina, Colby Ayers, Ronald M Peshock, Susan A Matulevicius, and Mark H Drazner. Association of African Ancestry With Electrocardiographic Voltage and Concentric Left Ventricular Hypertrophy: The Dallas Heart Study. *JAMA Cardiol*, 3(12):1167–1173, December 2018.

[12] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, 19(9):1655–1664, September 2009.

[13] All of Us Research Program Investigators, Joshua C Denny, Joni L Rutter, David B Goldstein, Anthony Philippakis, Jordan W Smoller, Gwynne Jenkins, and Eric Dishman. The "all of us" research program. *N. Engl. J. Med.*, 381(7):668–676, August 2019.

[14] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, Soumya Raychaudhuri, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832, 2010.

[15] R Ameratunga, S-T Woon, D Gillis, W Koopmans, and R Steele. New diagnostic criteria for common variable immune deficiency (CVID), which may assist with decisions to treat with intravenous or subcutaneous immunoglobulin. *Clin. Exp. Immunol.*, 174(2):203–211, November 2013.

[16] American Academy of Pediatrics Board of Directors and Executive Committee. AAP perspective: Race-Based medicine. *Pediatrics*, 148(4), October 2021.

[17] American Psychological Association. Guidelines for psychological practice with transgender and gender nonconforming people. *American Psychologist*, 70(9):832–864, December 2015.

[18] Christina Amutah, Kaliya Greenidge, Adjoa Mante, Michelle Munyikwa, Sanjna L Surya, Eve Higginbotham, David S Jones, Risa Lavizzo-Mourey, Dorothy Roberts, Jennifer Tsai, and Jaya Aysola. Misrepresenting race — the role of medical schools in propagating physician bias, 2021.

[19] William Anderson. ASHG documents and apologizes for past harms of human genetics research, commits to building an equitable future. https://www.ashg.org/publications-news/press-releases/ashg-documents-and-apologizes-for-past-harms-of-human-genetics-research-commits-to January 2023. Accessed: 2023-2-26.

[20] Hugues Aschard, Bjarni J Vilhjálmsson, Amit D Joshi, Alkes L Price, and Peter Kraft. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The American Journal of Human Genetics*, 96(2):329–339, 2015.

[21] Undark Ashley Smart. A field at a crossroads: Genetics and racial mythmaking. *Scientific American*, 2022.

[22] Peter J Aspinall. Ethnic/Racial terminology as a form of representation: A critical review of the lexicon of collective and specific terms in use in britain. *Genealogy*, 4(3):87, August 2020.

[23] Suheil Albert Atallah-Yunes, Audrey Ready, and Peter E Newburger. Benign ethnic neutropenia. *Blood Rev.*, 37:100586, September 2019.

[24] Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, Richard JL Anney, Stephan Ripke, Verneri Anttila, Jakob Grove, Peter Holmans, Hailiang Huang, Lambertus Klei, Phil H Lee, Sarah E Medland, et al. Meta-analysis of gwas of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Molecular autism*, 8:1–17, 2017.

[25] Jennifer L Baker, Charles N Rotimi, and Daniel Shriner. Human ancestry correlates with language and reveals that race is not an objective genomic classifier. *Sci. Rep.*, 7(1):1–10, May 2017.

[26] Carolyn Baloh, Anupama Reddy, Michele Henson, Katherine Prince, Rebecca Buckley, and Patricia Lugar. 30-year review of pediatric- and Adult-Onset CVID: Clinical correlates and prognostic indicators. *J. Clin. Immunol.*, 39(7):678–687, October 2019.

[27] Michael Bamshad, Stephen Wooding, Benjamin A Salisbury, and J Claiborne Stephens. Deconstructing the relationship between genetics and race. *Nat. Rev. Genet.*, 5(8):598–609, August 2004.

[28] Juan M Banda, Ashish Sarraju, Fahim Abbasi, Justin Parizo, Mitchel Pariani, Hannah Ison, Elinor Briskin, Hannah Wand, Sebastien Dubois, Kenneth Jung, Seth A Myers, Daniel J Rader, Joseph B Leader, Michael F Murray, Kelly D Myers, Katherine Wilemon, Nigam H Shah, and Joshua W Knowles. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. *NPJ Digit Med*, 2:23, April 2019.

[29] Yambazi Banda, Mark N Kvale, Thomas J Hoffmann, Stephanie E Hesselson, Dilrini Ranatunga, Hua Tang, Chiara Sabatti, Lisa A Croen, Brad P Dispensa, Mary Henderson, Carlos Iribarren, Eric Jorgenson, Lawrence H Kushi, Dana Ludwig, Diane Olberg, Charles P Quesenberry, Sarah Rowell, Marianne Sadler, Lori C Sakoda, Stanley Sciortino, Ling Shen, David Smethurst, Carol P Somkin, Stephen K Van Den Eeden, Lawrence Walter, Rachel A Whitmer, Pui-Yan Kwok, Catherine Schaefer, and Neil Risch. Characterizing Race/Ethnicity and genetic ancestry for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. *Genetics*, 200(4):1285–1295, June 2015.

[30] Arnold Barnes. Race and hospital diagnoses of schizophrenia and mood disorders. *Soc. Work.*, 53(1):77–83, January 2008.

[31] Lisa Bastarache, Jacob J Hughey, Scott Hebbring, Joy Marlo, Wanke Zhao, Wanting T Ho, Sara L Van Driest, Tracy L McGregor, Jonathan D Mosley, Quinn S Wells, Michael Temple, Andrea H Ramirez, Robert Carroll, Travis Osterman, Todd Edwards, Douglas Ruderfer, Digna R Velez Edwards, Rizwan Hamid, Joy Cogan, Andrew Glazer, Wei-Qi Wei, Qiping Feng, Murray Brilliant, Zhizhuang J Zhao, Nancy J Cox, Dan M Roden, and Joshua C Denny. Phenotype risk scores identify patients with unrecognized mendelian disease patterns. *Science*, 359(6381):1233–1239, March 2018.

[32] Oliver J Bear Don't Walk, 4th, Harry Reyes Nieva, Sandra Soo-Jin Lee, and Noémie Elhadad. A scoping review of ethics considerations in clinical natural language processing. *JAMIA Open*, 5(2):ooac039, July 2022.

[33] Lauren J Beesley, Maxwell Salvatore, Lars G Fritsche, Anita Pandit, Arvind Rao, Chad Brummett, Cristen J Willer, Lynda D Lisabeth, and Bhramar Mukherjee. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Stat. Med.*, 39(6):773–800, March 2020.

[34] Gillian M Belbin, Sinead Cullina, Stephane Wenric, Emily R Soper, Benjamin S Glicksberg, Denis Torre, Arden Moscati, Genevieve L Wojcik, Ruhollah Shemirani, Noam D Beckmann, Ariella Cohain, Elena P Sorokin, Danny S Park, Jose-Luis Ambite, Steve Ellis, Adam Auton, Erwin P Bottinger, Judy H Cho, Ruth J F Loos, Noura S Abul-Husn, Noah A Zaitlen, Christopher R Gignoux, and Eimear E Kenny. Toward a fine-scale population health monitoring system, 2021.

[35] Gillian M Belbin, Sinead Cullina, Stephane Wenric, Emily R Soper, Benjamin S Glicksberg, Denis Torre, Arden Moscati, Genevieve L Wojcik, Ruhollah Shemirani, Noam D Beckmann, Ariella Cohain, Elena P Sorokin, Danny S Park, Jose-Luis Ambite, Steve Ellis, Adam Auton, Erwin P Bottinger, Judy H Cho, Ruth J F Loos, Noura S Abul-Husn, Noah A Zaitlen, Christopher R Gignoux, and Eimear E Kenny. Toward a fine-scale population health monitoring system. *Cell*, 184(8):2068–2083.e11, April 2021.

[36] Gillian Morven Belbin, Jacqueline Odgis, Elena P Sorokin, Muh-Ching Yee, Sumita Kohli, Benjamin S Glicksberg, Christopher R Gignoux, Genevieve L Wojcik, Tielman Van Vleck, Janina M Jeff, Michael Linderman, Claudia Schurmann, Douglas Ruderfer, Xiaoqiang Cai, Amanda Merkelson, Anne E Justice, Kristin L Young, Misa Graff, Kari E North, Ulrike Peters, Regina James, Lucia Hindorff, Ruth Kornreich, Lisa Edelmann, Omri Gottesman, Eli E A Stahl, Judy H Cho, Ruth J F Loos, Erwin P Bottinger, Girish N Nadkarni, Noura S Abul-Husn, and Eimear E Kenny. Genetic identification of a common collagen disease in puerto ricans via identity-by-descent mapping in a health system, 2017.

[37] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283, 2016.

[38] Laura M Beskow and Kevin P Weinfurt. Exploring Understanding of "Understanding": The Paradigm Case of Biobank Consent Comprehension. *Am. J. Bioeth.*, 19(5):6–18, May 2019. Publisher: Taylor & Francis.

[39] Nrupen A Bhavsar, Aijing Gao, Matthew Phelan, Neha J Pagidipati, and Benjamin A Goldstein. Value of neighborhood socioeconomic status in predicting risk of outcomes in studies that use electronic health record data. *JAMA Netw Open*, 1(5):e182716, September 2018.

[40] Tim B Bigdeli, Georgios Voloudakis, Peter B Barr, Bryan R Gorman, Giulio Genovese, Roseann E Peterson, David E Burstein, Vlad I Velicu, Yuli Li, Rishab Gupta, Manuel Mattheisen, Simone Tomasi, Nallakkandi Rajeevan, Frederick Sayward, Krishnan Radhakrishnan, Sundar Natarajan, Anil K Malhotra, Yunling Shi, Hongyu Zhao, Thomas R Kosten, John Concato, Timothy J O'Leary, Ronald Przygodzki, Theresa Gleason, Saiju Pyarajan, Mary Brophy, Grant D Huang, Sumitra Muralidhar, J Michael Gaziano, Mihaela Aslan, Ayman H Fanous, Philip D Harvey, Panos Roussos, M Antonelli, M de Asis, M S Bauer, John Concato, F Cunningham, R Freedman, Michael Gaziano, Philip Harvey, Grant Huang, J Kelsoe, Thomas Kosten, T Lehner, J B Lohr, S R Marder, P Miller, Timothy O Leary, T Patterson, P Peduzzi, Ronald

Przygodski, Larry Siever, P Sklar, S Strakowski, Ayman Fanous, W Farwell, A Malhorta, S Mane, P Palacios, Tim Bigdeli, M Corsey, L Zaluda, Juanita Johnson, Melyssa Sueiro, D Cavaliere, V Jeanpaul, Alysia Maffucci, L Mancini, J Deen, G Muldoon, Stacey Whitbourne, J Canive, L Adamson, L Calais, G Fuldauer, R Kushner, G Toney, M Lackey, A Mank, N Mahdavi, G Villarreal, E C Muly, F Amin, M Dent, J Wold, B Fischer, A Elliott, C Felix, G Gill, P E Parker, C Logan, J McAlpine, L E DeLisi, S G Reece, M B Hammer, D Agbor-Tabie, W Goodson, M Aslam, M Grainger, Neil Richtand, Alexander Rybalsky, R Al Jurdi, E Boeckman, T Natividad, D Smith, M Stewart, S Torres, Z Zhao, A Mayeda, A Green, J Hofstetter, S Ngombu, M K Scott, A Strasburger, J Sumner, G Paschall, J Mucciarelli, R Owen, S Theus, D Tompkins, S G Potkin, C Reist, M Novin, S Khalaghizadeh, Richard Douyon, Nita Kumar, Becky Martinez, S R Sponheim, T L Bender, H L Lucas, A M Lyon, M P Marggraf, L H Sorensen, C R Surerus, C Sison, J Amato, D R Johnson, N Pagan-Howard, L A Adler, S Alerpin, T Leon, K M Mattocks, N Araeva, J C Sullivan, T Suppes, K Bratcher, L Drag, E G Fischer, L Fujitani, S Gill, D Grimm, J Hoblyn, T Nguyen, E Nikolaev, L Shere, R Relova, A Vicencio, M Yip, I Hurford, S Acheampong, G Carfagno, G L Haas, C Appelt, E Brown, B Chakraborty, E Kelly, G Klima, S Steinhauer, R A Hurley, R Belle, D Eknoyan, K Johnson, J Lamotte, E Granholm, K Bradshaw, J Holden, R H Jones, T Le, I G Molina, M Peyton, I Ruiz, L Sally, A Tapp, S Devroy, V Jain, N Kilzieh, L Maus, K Miller, H Pope, A Wood, E Meyer, P Givens, P B Hicks, S Justice, K McNair, J L Pena, D F Tharp, L Davis, M Ban, L Cheatum, P Darr, W Grayson, J Munford, B Whitfield, E Wilson, S E Melnikoff, B L Schwartz, M A Tureson, D D Souza, K Forselius, M Ranganathan, L Rispoli, M Sather, C Colling, C Haakenson, D Kruegar, Rachel Ramoni, Jim Breeling, Kyong-Mi Chang, Christopher O Donnell, Philip Tsao, Jennifer Moser, Jessica Brewer, Stuart Warren, Dean Argyres, Brady Stevens, Donald Humphries, Nhan Do, Shahpoor Shayan, Xuan-Mai Nguyen, Kelly Cho, Elizabeth Hauser, Yan Sun, Peter Wilson, Rachel McArdle, Louis Dellitalia, John Harley, and Jeffrey Whittle. Penetrance and pleiotropy of polygenic risk scores for schizophrenia, bipolar disorder, and depression among adults in the US veterans affairs health care system. *JAMA Psychiatry*, 79(11):1092–1101, November 2022.

[41] Jaclyn A Bjelac, Jennifer R Yonkof, and James Fernandez. Differing performance of the warning signs for immunodeficiency in the diagnosis of pediatric versus adult patients in a Two-Center tertiary referral population. *J. Clin. Immunol.*, 39(1):90–98, January 2019.

[42] Mary Regina Boland, Lena M Davidson, Silvia P Canelón, Jessica Meeker, Trevor Penning, John H Holmes, and Jason H Moore. Harnessing electronic health records to study emerging environmental disasters: a proof of concept with perfluoroalkyl substances (PFAS). *NPJ Digit Med*, 4(1):122, August 2021.

[43] Mary Regina Boland, Pradipta Parhi, Li Li, Riccardo Miotto, Robert Carroll, Usman Iqbal, Phung-Anh (alex) Nguyen, Martijn Schuemie, Seng Chan You, Donahue Smith, Sean Mooney, Patrick Ryan, Yu-Chuan (jack) Li, Rae Woong Park, Josh Denny, Joel T

Dudley, George Hripcsak, Pierre Gentine, and Nicholas P Tatonetti. Uncovering exposures responsible for birth season – disease effects: a global study. *J. Am. Med. Inform. Assoc.*, 25(3):275–288, September 2017.

[44] Mary Regina Boland, Zachary Shahn, David Madigan, George Hripcsak, and Nicholas P Tatonetti. Birth month affects lifetime disease risk: a phenome-wide method. *J. Am. Med. Inform. Assoc.*, 22(5):1042–1053, June 2015.

[45] Francisco A Bonilla, Isil Barlan, Helen Chapel, Beatriz T Costa-Carvalho, Charlotte Cunningham-Rundles, M Teresa de la Morena, Francisco J Espinosa-Rosales, Lennart Hammarström, Shigeaki Nonoyama, Isabella Quinti, John M Routes, Mimi L K Tang, and Klaus Warnatz. International consensus document (ICON): Common variable immunodeficiency disorders. *J. Allergy Clin. Immunol. Pract.*, 4(1):38–59, January 2016.

[46] Luisa N Borrell. Racial identity among hispanics: implications for health and well-being. *Am. J. Public Health*, 95(3):379–381, March 2005.

[47] Luisa N Borrell, Jennifer R Elhawary, Elena Fuentes-Afflick, Jonathan Witonsky, Nirav Bhakta, Alan H B Wu, Kirsten Bibbins-Domingo, José R Rodríguez-Santana, Michael A Lenoir, James R Gavin, Rick A Kittles, Noah A Zaitlen, David S Wilkes, Neil R Powe, Elad Ziv, and Esteban G Burchard. Race and genetic ancestry in medicine — a time for reckoning with racism, 2021.

[48] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[49] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[50] Brielin C Brown, Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Chun Jimmie Ye, Alkes L Price, and Noah Zaitlen. Transethnic Genetic-Correlation Estimates from Summary Statistics. *Am. J. Hum. Genet.*, 99(1):76–88, July 2016.

[51] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236, 2015.

[52] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[53] Annalisa Buniello, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L. Whetzel, Ridwan Amode,

Jose A. Guillen, Harpreet S. Riat, Stephen J. Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A. Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, January 2019.

[54] Esteban González Burchard, Luisa N Borrell, Shweta Choudhry, Mariam Naqvi, Hui-Ju Tsai, Jose R Rodriguez-Santana, Rocio Chapela, Scott D Rogers, Rui Mei, William Rodriguez-Cintron, Jose F Arena, Rick Kittles, Eliseo J Perez-Stable, Elad Ziv, and Neil Risch. Latino populations: A unique opportunity for the study of race, genetics, and social environment in epidemiological research, 2005.

[55] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, 2018.

[56] Yen Ji Julia Byeon, Rezarta Islamaj, Lana Yeganova, W John Wilbur, Zhiyong Lu, Lawrence C Brody, and Vence L Bonham. Evolving use of ancestry, ethnicity, and race in genetics research-a survey spanning seven decades. *Am. J. Hum. Genet.*, 108(12):2215–2223, December 2021.

[57] Michael C Campbell and Sarah A Tishkoff. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.*, 9:403–433, 2008.

[58] Jedidiah Carlson and Kelley Harris. Quantifying and contextualizing the impact of biorxiv preprints through automated social media audience segmentation. *PLoS Biol.*, 18(9):e3000860, September 2020.

[59] Robert J Carroll, Lisa Bastarache, and Joshua C Denny. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*, 30(16):2375–2376, August 2014.

[60] Joan A Casey, Frank C Curriero, Sara E Cosgrove, Keeve E Nachman, and Brian S Schwartz. High-density livestock operations, crop field application of manure, and risk of community-associated methicillin-resistant staphylococcus aureus infection in pennsylvania. *JAMA Intern. Med.*, 173(21):1980–1990, November 2013.

[61] CDC. Racism and health. https://www.cdc.gov/minorityhealth/racism-disparities/index.html, May 2022. Accessed: 2023-3-6.

[62] Jessica P Cerdeña, Marie V Plaisime, and Jennifer Tsai. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. *The Lancet*, 396(10257):1125–1128, 2020.

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

[64] Christopher C Chang, Carson C Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4:7, February 2015.

[65] Nilanjan Chatterjee, Bill Wheeler, Joshua Sampson, Patricia Hartge, Stephen J Chanock, and Ju-Hyun Park. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature genetics*, 45(4):400, 2013.

[66] Min Chen, Xuan Tan, and Rema Padman. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *J. Am. Med. Inform. Assoc.*, 27(11):1764–1773, November 2020.

[67] Moon S Chen, Jr. Cancer health disparities among Asian Americans: what we do and what we need to do. *Cancer*, 104(12 Suppl):2895–2902, December 2005.

[68] Z Chen, J Chen, R Collins, Y Guo, R Peto, F Wu, and L Li. China kadoorie biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.*, 40(6), December 2011.

[69] Elizabeth Gross Cohn, Nalo Hamilton, Elaine L Larson, and Janet K Williams. Self-reported race and ethnicity of US biobank participants compared to the US census. *J. Community Genet.*, 8(3):229–238, June 2017.

[70] Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, and Institute of Medicine. Summary. In *Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1*. National Academies Press (US), June 2014.

[71] Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, and Institute of Medicine. Recommended core domains and measures. In *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. National Academies Press (US), January 2015.

[72] Matthew P Conomos, Alex P Reiner, Mary Sara McPeek, and Timothy A Thornton. Genome-wide control of population structure and relatedness in genetic association studies via linear mixed models with orthogonally partitioned structure. *bioRxiv*, page 409953, 2018.

[73] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

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

[75] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[76] Graham Coop. Genetic similarity and genetic ancestry groups. *arXiv preprint arXiv:2207.11595*, 2022.

[77] Chris Cotsapas, Benjamin F Voight, Elizabeth Rossin, Kasper Lage, Benjamin M Neale, Chris Wallace, Gonçalo R Abecasis, Jeffrey C Barrett, Timothy Behrens, Judy Cho, et al. Pervasive sharing of genetic effects in autoimmune disease. *PLoS genetics*, 7(8):e1002254, 2011.

[78] COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature*, July 2021.

[79] C Cunningham-Rundles. Autoimmune manifestations in common variable immunodeficiency. *J. Clin. Immunol.*, 28 Suppl 1:S42–5, May 2008.

[80] David Curtis. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.*, 28(5):85–89, October 2018.

[81] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, David Schlessinger, Dwight Stambolian, Po-Ru Loh, William G Iacono, Anand Swaroop, Laura J Scott, Francesco Cucca, Florian Kronenberg, Michael Boehnke, Gonçalo R Abecasis, and Christian Fuchsberger. Next-generation genotype imputation service and methods. *Nat. Genet.*, 48(10):1284–1287, October 2016.

[82] Bege Dauda, Santiago J Molina, Danielle S Allen, Agustin Fuentes, Nayanika Ghosh, Madelyn Mauro, Benjamin M Neale, Aaron Panofsky, Mashaal Sohail, Sarah R Zhang, and Anna C F Lewis. Ancestry: How researchers use it and what they mean by it. *Front. Genet.*, 14, January 2023.

[83] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, Melissa A Basford, David S Carrell, Peggy L Peissig, Abel N Kho, Jennifer A Pacheco, Luke V Rasmussen, David R Crosslin, Paul K Crane, Jyotishman Pathak, Suzette J Bielinski, Sarah A Pendergrass, Hua Xu, Lucia A Hindorff, Rongling Li, Teri A Manolio, Christopher G Chute, Rex L Chisholm, Eric B Larson, Gail P Jarvik, Murray H Brilliant, Catherine A McCarty, Iftikhar J Kullo, Jonathan L Haines, Dana C Crawford, Daniel R Masys, and Dan M Roden. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.*, 31(12):1102–1110, December 2013.

[84] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, Melissa A Basford, David S Carrell, Peggy L Peissig, Abel N Kho, Jennifer A Pacheco, Luke V Rasmussen, David R Crosslin, Paul K Crane, Jyotishman Pathak, Suzette J Bielinski, Sarah A Pendergrass, Hua Xu, Lucia A Hindorff, Rongling Li, Teri A Manolio, Christopher G Chute, Rex L Chisholm, Eric B Larson, Gail P Jarvik, Murray H Brilliant, Catherine A McCarty, Iftikhar J Kullo, Jonathan L Haines, Dana C Crawford, Daniel R Masys, and Dan M Roden. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.*, 31(12):1102–1110, December 2013.

[85] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, 26(9):1205–1210, May 2010.

[86] Qing Duan, Eric Yi Liu, Paul L Auer, Guosheng Zhang, Ethan M Lange, Goo Jun, Chris Bizon, Shuo Jiao, Steven Buyske, Nora Franceschini, Chris S Carlson, Li Hsu, Alex P Reiner, Ulrike Peters, Jeffrey Haessler, Keith Curtis, Christina L Wassel, Jennifer G Robinson, Lisa W Martin, Christopher A Haiman, Loic Le Marchand, Tara C Matise, Lucia A Hindorff, Dana C Crawford, Themistocles L Assimes, Hyun Min Kang, Gerardo Heiss, Rebecca D Jackson, Charles Kooperberg, James G Wilson, Gonçalo R Abecasis, Kari E North, Deborah A Nickerson, Leslie A Lange, and Yun Li. Imputation of coding variants in african americans: better performance using data from the exome sequencing project. *Bioinformatics*, 29(21):2744–2749, August 2013.

[87] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3):e1003348, 2013.

[88] Hillary R Dueñas, Carina Seah, Jessica S Johnson, and Laura M Huckins. Implicit bias of encoded variables: frameworks for addressing structured bias in EHR–GWAS data. *Hum. Mol. Genet.*, 29(R1):R33–R41, September 2020.

[89] Dustin T Duncan, Mona Sharifi, Steven J Melly, Richard Marshall, Thomas D Sequist, Sheryl L Rifas-Shiman, and Elsie M Taveras. Characteristics of walkable built environments and BMI z-scores in children: evidence from a large electronic health record database. *Environ. Health Perspect.*, 122(12):1359–1365, December 2014.

[90] Brian J Edwards, Chad Haynes, Mark A Levenstien, Stephen J Finch, and Derek Gordon. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genet.*, 6:18, 2005.

[91] Eran Elhaik. In search of the jüdische typus: A proposed benchmark to test the genetic basis of jewishness challenges notions of "jewish biomarkers". *Front. Genet.*, 7, August 2016.

[92] David Ellinghaus, Eva Ellinghaus, Rajan P Nair, Philip E Stuart, Tõnu Esko, Andres Metspalu, Sophie Debrus, John V Raelson, Trilokraj Tejasvi, Majid Belouchi, et al. Combined analysis of genome-wide association studies for crohn disease and psoriasis identifies seven shared susceptibility loci. *The American Journal of Human Genetics*, 90(4):636–647, 2012.

[93] Nwamaka Denise Eneanya, Wei Yang, and Peter Philip Reese. Reconsidering the Consequences of Using Race to Estimate Kidney Function. *JAMA*, 322(2):113–114, July 2019. Publisher: American Medical Association.

[94] Raphael Falk. Genetic markers cannot determine jewish descent. *Front. Genet.*, 5, January 2015.

[95] Jocelyn R Farmer, Mei-Sing Ong, Sara Barmettler, Lael M Yonker, Ramsay Fuleihan, Kathleen E Sullivan, Charlotte Cunningham-Rundles, USIDNET Consortium, and Jolan E Walter. Common Variable Immunodeficiency Non-Infectious Disease Endotypes Redefined Using Unbiased Network Clustering in Large Electronic Datasets. *Front. Immunol.*, 8:1740, 2017.

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

[97] Annette Flanagin, Tracy Frey, Stacy L Christiansen, and Howard Bauchner. The Reporting of Race and Ethnicity in Medical and Science Journals: Comments Invited. *JAMA*, 325(11):1049–1052, March 2021.

[98] Laura E Flores, Walter R Frontera, Michele P Andrasik, Carlos Del Rio, Antonio Mondríguez-González, Stephanie A Price, Elizabeth M Krantz, Steven A Pergam, and Julie K Silver. Assessment of the Inclusion of Racial/Ethnic Minority, Female, and Older Individuals in Vaccine Clinical Trials. *JAMA Netw Open*, 4(2):e2037640, February 2021.

[99] Phil B Fontanarosa and Howard Bauchner. Race, Ancestry, and Medical Research. *JAMA*, 320(15):1539–1540, October 2018. Publisher: American Medical Association.

[100] Iain S Forrest, Kumardeep Chaudhary, Ha My T Vy, Ben O Petrazzini, Shantanu Bafna, Daniel M Jordan, Ghislain Rocheleau, Ruth J F Loos, Girish N Nadkarni, Judy H Cho, and Ron Do. Population-Based penetrance of deleterious clinical variants. *JAMA*, 327(4):350–359, January 2022.

[101] Morris W Foster and Richard R Sharp. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome Res.*, 12(6):844–850, June 2002.

[102] Malika Kumar Freund, Kathryn S Burch, Huwenbo Shi, Nicholas Mancuso, Gleb Kichaev, Kristina M Garske, David Z Pan, Zong Miao, Karen L Mohlke, Markku Laakso, et al. Phenotype-specific enrichment of mendelian disorder genes near gwas regions across 62 complex traits. *The American Journal of Human Genetics*, 103(4):535–552, 2018.

[103] Christian Fuchsberger, Gonçalo R Abecasis, and David A Hinds. minimac2: faster genotype imputation. *Bioinformatics*, 31(5):782–784, March 2015.

[104] Suzanne H Gage, George Davey Smith, Jennifer J Ware, Jonathan Flint, and Marcus R Munafò. G = e: What GWAS can tell us about the environment. *PLoS Genet.*, 12(2):e1005765, February 2016.

[105] Tian Ge, Marguerite R Irvin, Amit Patki, Vinodh Srinivasasainagendra, Yen-Feng Lin, Hemant K Tiwari, Nicole D Armstrong, Barbara Benoit, Chia-Yen Chen, Karmel W Choi, James J Cimino, Brittney H Davis, Ozan Dikilitas, Bethany Etheridge, Yen-Chen Anne Feng, Vivian Gainer, Hailiang Huang, Gail P Jarvik, Christopher Kachulis, Eimear E Kenny, Atlas Khan, Krzysztof Kiryluk, Leah Kottyan, Iftikhar J Kullo, Christoph Lange, Niall Lennon, Aaron Leong, Edyta Malolepsza, Ayme D Miles, Shawn Murphy, Bahram Namjou, Renuka Narayan, Mark J O'Connor, Jennifer A Pacheco, Emma Perez, Laura J Rasmussen-Torvik, Elisabeth A Rosenthal, Daniel Schaid, Maria Stamou, Miriam S Udler, Wei-Qi Wei, Scott T Weiss, Maggie C Y Ng, Jordan W Smoller, Matthew S Lebo, James B Meigs, Nita A Limdi, and Elizabeth W Karlson. Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. *Genome Med.*, 14(1):70, June 2022.

[106] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[107] Hassan M K Ghomrawi, Russell J Funk, Michael L Parks, Jason Owen-Smith, and John M Hollingsworth. Physician referral patterns and racial disparities in total hip replacement: A network analysis approach. *PLoS One*, 13(2):e0193014, February 2018.

[108] Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics*, 10(5):e1004383, 2014.

[109] Ellen Goldstein, James Topitzes, Julie Miller-Cribbs, and Roger L Brown. Influence of race/ethnicity and income on the link between adverse childhood experiences and child flourishing. *Pediatr. Res.*, 89(7):1861–1869, October 2020.

[110] Michael I Goran and Emily E Ventura. Genetic predisposition and increasing dietary fructose exposure: the perfect storm for fatty liver disease in Hispanics in the U.S. *Dig. Liver Dis.*, 44(9):711–713, September 2012.

[111] Vincenzo Graziano, Antonio Pecoraro, Ilaria Mormile, Giuseppe Quaremba, Arturo Genovese, Claudio Buccelli, Mariano Paternoster, and Giuseppe Spadaro. Delay in diagnosis affects the clinical outcome in a cohort of cvid patients with marked reduction of iga serum levels, 2017.

[112] Eric D Green and Mark S Guyer. Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333):204–213, February 2011.

[113] Neil S Greenspan. Genes, heritability, 'race', and intelligence: Misapprehensions and implications. *Genes*, 13(2), February 2022.

[114] Tudor Groza, Sebastian Köhler, Dawid Moldenhauer, Nicole Vasilevsky, Gareth Baynam, Tomasz Zemojtel, Lynn Marie Schriml, Warren Alden Kibbe, Paul N Schofield, Tim Beck, Drashtti Vasant, Anthony J Brookes, Andreas Zankl, Nicole L Washington, Christopher J Mungall, Suzanna E Lewis, Melissa A Haendel, Helen Parkinson, and Peter N Robinson. The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. *Am. J. Hum. Genet.*, 97(1):111–124, July 2015.

[115] Yongtao Guan, Matthew Stephens, et al. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815, 2011.

[116] Xiaofan Guo, Eric Vittinghoff, Jeffrey E Olgin, Gregory M Marcus, and Mark J Pletcher. Volunteer Participation in the Health eHeart Study: A Comparison with the US Population. *Sci. Rep.*, 7(1):1956, May 2017.

[117] Deepti Gurdasani, Inês Barroso, Eleftheria Zeggini, and Manjinder S Sandhu. Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.*, 20(9):520–535, June 2019.

[118] Romana Hasnain-Wynia and David W Baker. Obtaining data on patient race, ethnicity, and primary language in health care organizations: Current challenges and proposed solutions. *Health Serv. Res.*, 41(4 Pt 1):1501, August 2006.

[119] J. A. Heit, S. M. Armasu, Y. W. Asmann, J. M. Cunningham, M. E. Matsumoto, T. M. Petterson, and M. De Andrade. A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. *Journal of thrombosis and haemostasis: JTH*, 10(8):1521–1531, August 2012.

[120] Garrett Hellenthal, George B J Busby, Gavin Band, James F Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, February 2014.

[121] Christine B Hickman. The devil and the one drop rule: Racial categories, african americans, and the U.S. census. *Mich. Law Rev.*, 95(5):1161–1265, 1997.

[122] Gracie Himmelstein, David Bates, and Li Zhou. Examination of stigmatizing language in the electronic health record. *JAMA Netw Open*, 5(1):e2144967, January 2022.

[123] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, pages genetics–114, 2014.

[124] Farhad Hormozdiari, Martijn van de Bunt, Ayellet V Segrè, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

[125] Jan Horsky, Elizabeth A Drucker, and Harley Z Ramelson. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. *AMIA Annu. Symp. Proc.*, 2017:912–920, 2017.

[126] Kangcheng Hou, Kathryn S Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nature genetics*, 51(8):1244–1251, 2019.

[127] Yao Hu, Stephanie A Bien, Katherine K Nishimura, Jeffrey Haessler, Chani J Hodonsky, Antoine R Baldassari, Heather M Highland, Zhe Wang, Michael Preuss, Colleen M Sitlani, Genevieve L Wojcik, Ran Tao, Mariaelisa Graff, Laura M Huckins, Quan Sun, Ming-Huei Chen, Abdou Mousas, Paul L Auer, Guillaume Lettre, the Blood Cell Consortium, and Charles Kooperberg. Multi-ethnic genome-wide association analyses of white blood cell and platelet traits in the population architecture using genomics and epidemiology (PAGE) study. *BMC Genomics*, 22, 2021.

[128] Yiming Hu, Qiongshi Lu, Wei Liu, Yuhua Zhang, Mo Li, and Hongyu Zhao. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS genetics*, 13(6):e1006836, 2017.

[129] Yiming Hu, Qiongshi Lu, Ryan Powles, Xinwei Yao, Can Yang, Fang Fang, Xinran Xu, and Hongyu Zhao. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS computational biology*, 13(6):e1005589, 2017.

[130] J E Huffman. Examining the current standards for genetic discovery and replication in the era of mega-biobanks. *Nat. Commun.*, 9(1):1–4, November 2018.

[131] Linda M Hunt and Mary S Megyesi. The ambiguous meanings of the racial/ethnic categories routinely used in human genetics research. *Soc. Sci. Med.*, 66(2):349–361, January 2008.

[132] John P A Ioannidis, Neil R Powe, and Clyde Yancy. Recalibrating the use of race in medical research. *JAMA*, 325(7):623–624, February 2021.

[133] Ruth Johnson, Huwenbo Shi, Bogdan Pasaniuc, and Sriram Sankararaman. A unifying framework for joint trait analysis under a non-infinitesimal model. *Bioinformatics*, 34(13):i195–i201, July 2018.

[134] Ruth D Johnson, Yi Ding, Arjun Bhattacharya, Alec Chiu, Clara Lajonchere, Daniel H Geschwind, and Bogdan Pasaniuc. The UCLA ATLAS Community Health Initiative: promoting precision health research in a diverse biobank. Publication Title: bioRxiv, February 2022.

[135] Ruth Dolly Johnson, Yi Ding, Vidhya Venkateswaran, Arjun Bhattacharya, Alec Chiu, Tommer Schwarz, Malika Freund, Lingyu Zhan, Kathryn S Burch, Christa Caggiano, Brian Hill, Nadav Rakocz, Brunilda Balliu, Jae Hoon Sul, Noah Zaitlen, Valerie A Arboleda, Eran Halperin, Sriram Sankararaman, Manish J Butte, Clara Lajonchere, Daniel H Geschwind, Bogdan Pasaniuc, UCLA Precision Health Data Discovery Repository Working Group, and UCLA Precision Health ATLAS Working Group. Leveraging genomic diversity for discovery in an EHR-linked biobank: the UCLA ATLAS Community Health Initiative. September 2021.

[136] I T Jolliffe. Principal Component Analysis and Factor Analysis. In I T Jolliffe, editor, *Principal Component Analysis*, pages 115–128. Springer New York, New York, NY, 1986.

[137] Eric R Kallwitz, Bamidele O Tayo, Mark H Kuniholm, Jianwen Cai, Martha Daviglus, Richard S Cooper, and Scott J Cotler. American Ancestry Is a Risk Factor for Suspected Nonalcoholic Fatty Liver Disease in Hispanic/Latino Adults. *Clin. Gastroenterol. Hepatol.*, 17(11):2301–2309, October 2019.

[138] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-Yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, 42(4):348–354, April 2010.

[139] Alex N Kasembeli, Raquel Duarte, Michèle Ramsay, and Saraladevi Naicker. African origins and chronic kidney disease susceptibility in the human immunodeficiency virus era. *World J Nephrol*, 4(2):295–306, May 2015.

[140] Suranga N Kasthurirathne, Joshua R Vest, Nir Menachemi, Paul K Halverson, and Shaun J Grannis. Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services. *J. Am. Med. Inform. Assoc.*, 25(1):47–53, January 2018.

[141] Alyna T Khan, Stephanie M Gogarten, Caitlin P McHugh, Adrienne M Stilp, Tamar Sofer, Michael L Bowers, Quenna Wong, L Adrienne Cupples, Bertha Hidalgo, Andrew D Johnson, Merry-Lynn N McDonald, Stephen T McGarvey, Matthew R G Taylor, Stephanie M Fullerton, Matthew P Conomos, and Sarah C Nelson. Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: Experiences from the NHLBI TOPMed program. *Cell Genom*, 2(8), August 2022.

[142] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor,

and Sekar Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.*, 50(9):1219–1224, August 2018.

[143] Shaan Khurshid, Christopher Reeder, Lia X Harrington, Pulkit Singh, Gopal Sarma, Samuel F Friedman, Paolo Di Achille, Nathaniel Diamant, Jonathan W Cunningham, Ashby C Turner, Emily S Lau, Julian S Haimovich, Mostafa A Al-Alusi, Xin Wang, Marcus D R Klarqvist, Jeffrey M Ashburner, Christian Diedrich, Mercedeh Ghadessi, Johanna Mielke, Hanna M Eilken, Alice McElhinney, Andrea Derix, Atlas, Steven J, Patrick T Ellinor, Anthony A Philippakis, Christopher D Anderson, Jennifer E Ho, Puneet Batra, and Steven A Lubitz. Cohort design and natural language processing to reduce bias in electronic health records research. *npj Digital Medicine*, 5(1):1–14, April 2022.

[144] Gleb Kichaev and Bogdan Pasaniuc. Leveraging Functional-Annotation data in transethnic Fine-Mapping studies. *Am. J. Hum. Genet.*, 97(2):260, August 2015.

[145] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, 10(10):e1004722, 2014.

[146] Jacqueline C Kirby, Peter Speltz, Luke V Rasmussen, Melissa Basford, Omri Gottesman, Peggy L Peissig, Jennifer A Pacheco, Gerard Tromp, Jyotishman Pathak, David S Carrell, Stephen B Ellis, Todd Lingren, Will K Thompson, Guergana Savova, Jonathan Haines, Dan M Roden, Paul A Harris, and Joshua C Denny. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Inform. Assoc.*, 23(6):1046–1052, November 2016.

[147] Robert M Kirkpatrick, Matt McGue, William G Iacono, Michael B Miller, and Saonli Basu. Results of a "GWAS plus:" general cognitive ability is substantially heritable and massively polygenic. *PLoS One*, 9(11):e112390, November 2014.

[148] Rick A Kittles and Kenneth M Weiss. Race, ancestry, and genes: implications for defining disease risk. *Annu. Rev. Genomics Hum. Genet.*, 4:33–67, 2003.

[149] Elissa V Klinger, Sara V Carlini, Irina Gonzalez, Stella St Hubert, Jeffrey A Linder, Nancy A Rigotti, Emily Z Kontos, Elyse R Park, Lucas X Marinacci, and Jennifer S Haas. Accuracy of race, ethnicity, and language preference in an electronic health record. *J. Gen. Intern. Med.*, 30(6):719–723, December 2014.

[150] Madeline H Kowalski, Huijun Qian, Ziyi Hou, Jonathan D Rosen, Amanda L Tapia, Yue Shan, Deepti Jain, Maria Argos, Donna K Arnett, Christy Avery, Kathleen C Barnes, Lewis C Becker, Stephanie A Bien, Joshua C Bis, John Blangero, Eric Boerwinkle, Donald W Bowden, Steve Buyske, Jianwen Cai, Michael H Cho, Seung Hoan Choi, Hélène Choquet, L Adrienne Cupples, Mary Cushman, Michelle Daya, Paul S de Vries, Patrick T Ellinor, Nauder Faraday, Myriam Fornage, Stacey Gabriel, Santhi K Ganesh,

Misa Graff, Namrata Gupta, Jiang He, Susan R Heckbert, Bertha Hidalgo, Chani J Hodonsky, Marguerite R Irvin, Andrew D Johnson, Eric Jorgenson, Robert Kaplan, Sharon L R Kardia, Tanika N Kelly, Charles Kooperberg, Jessica A Lasky-Su, Ruth J F Loos, Steven A Lubitz, Rasika A Mathias, Caitlin P McHugh, Courtney Montgomery, Jee-Young Moon, Alanna C Morrison, Nicholette D Palmer, Nathan Pankratz, George J Papanicolaou, Juan M Peralta, Patricia A Peyser, Stephen S Rich, Jerome I Rotter, Edwin K Silverman, Jennifer A Smith, Nicholas L Smith, Kent D Taylor, Timothy A Thornton, Hemant K Tiwari, Russell P Tracy, Tao Wang, Scott T Weiss, Lu-Chen Weng, Kerri L Wiggins, James G Wilson, Lisa R Yanek, Sebastian Zöllner, Kari E North, Paul L Auer, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Hematology & Hemostasis Working Group, Laura M Raffield, Alexander P Reiner, and Yun Li. Use of ¿100,000 NHLBI Trans-Omics for precision medicine (TOPMed) consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed african and Hispanic/Latino populations. *PLoS Genet.*, 15(12):e1008500, December 2019.

[151] Rachel H Kowalsky, Ashley C Rondini, and Shari L Platt. The Case for Removing Race From the American Academy of Pediatrics Clinical Practice Guideline for Urinary Tract Infection in Infants and Young Children With Fever. *JAMA Pediatr.*, 174(3):229–230, March 2020. Publisher: American Medical Association.

[152] Peter Kraft, Eleftheria Zeggini, and John P A Ioannidis. Replication in genome-wide association studies. *Stat. Sci.*, 24(4):561, November 2009.

[153] Sebastian Köhler, Sandra C Doelken, Christopher J Mungall, Sebastian Bauer, Helen V Firth, Isabelle Bailleul-Forestier, Graeme C M Black, Danielle L Brown, Michael Brudno, Jennifer Campbell, David R FitzPatrick, Janan T Eppig, Andrew P Jackson, Kathleen Freson, Marta Girdea, Ingo Helbig, Jane A Hurst, Johanna Jähn, Laird G Jackson, Anne M Kelly, David H Ledbetter, Sahar Mansour, Christa L Martin, Celia Moss, Andrew Mumford, Willem H Ouwehand, Soo-Mi Park, Erin Rooney Riggs, Richard H Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J Wapner, Andrew O M Wilkie, Caroline F Wright, Anneke T Vulto-van Silfhout, Nicole de Leeuw, Bert B A de Vries, Nicole L Washingthon, Cynthia L Smith, Monte Westerfield, Paul Schofield, Barbara J Ruef, Georgios V Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E Lewis, and Peter N Robinson. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, 42(Database issue):D966–74, January 2014.

[154] Clara Lajonchere, Arash Naeim, Sarah Dry, Neil Wenger, David Elashoff, Sitaram Vangala, Antonia Petruse, Maryam Ariannejad, Clara Magyar, Liliana Johansen, Gabriela Werre, Maxwell Kroloff, and Daniel Geschwind. An Integrated, Scalable, Electronic Video Consent Process to Power Precision Health Research: Large, Population-Based, Cohort Implementation and Scalability Study. *J. Med. Internet Res.*, 23(12):e31121, December 2021.

[155] Clara Lajonchere, Arash Naeim, Sarah Dry, Neil Wenger, David Elashoff, Sitaram Vangala, Antonia Petruse, Maryam Ariannejad, Clara Magyar, Liliana Johansen, Gabriela Werre, Maxwell Kroloff, and Daniel Geschwind. An integrated, scalable, electronic video consent process to power precision health research: Large, Population-Based, cohort implementation and scalability study. *J. Med. Internet Res.*, 23(12):e31121, December 2021.

[156] Max Lam, Chia-Yen Chen, Zhiqiang Li, Alicia R Martin, Julien Bryois, Xixian Ma, Helena Gaspar, Masashi Ikeda, Beben Benyamin, Brielin C Brown, Ruize Liu, Wei Zhou, Lili Guan, Yoichiro Kamatani, Sung-Wan Kim, Michiaki Kubo, Agung A A A Kusumawardhani, Chih-Min Liu, Hong Ma, Sathish Periyasamy, Atsushi Takahashi, Zhida Xu, Hao Yu, Feng Zhu, Wei J Chen, Stephen Faraone, Stephen J Glatt, Lin He, Steven E Hyman, Hai-Gwo Hwu, Steven A McCarroll, Benjamin M Neale, Pamela Sklar, Dieter B Wildenauer, Xin Yu, Dai Zhang, Bryan J Mowry, Jimmy Lee, Peter Holmans, Shuhua Xu, Patrick F Sullivan, Stephan Ripke, Michael C O'Donovan, Mark J Daly, Shengying Qin, Pak Sham, Nakao Iwata, Kyung S Hong, Sibylle G Schwab, Weihua Yue, Ming Tsuang, Jianjun Liu, Xiancang Ma, René S Kahn, Yongyong Shi, and Hailiang Huang. Comparative genetic architectures of schizophrenia in east asian and european populations. *Nat. Genet.*, 51(12):1670–1678, November 2019.

[157] Max Lam, Joey W Trampush, Jin Yu, Emma Knowles, Gail Davies, David C Liewald, John M Starr, Srdjan Djurovic, Ingrid Melle, Kjetil Sundet, Andrea Christoforou, Ivar Reinvang, Pamela DeRosse, Astri J Lundervold, Vidar M Steen, Thomas Espeseth, Katri Räikkönen, Elisabeth Widen, Aarno Palotie, Johan G Eriksson, Ina Giegling, Bettina Konte, Panos Roussos, Stella Giakoumaki, Katherine E Burdick, Antony Payton, William Ollier, Ornit Chiba-Falek, Deborah K Attix, Anna C Need, Elizabeth T Cirulli, Aristotle N Voineskos, Nikos C Stefanis, Dimitrios Avramopoulos, Alex Hatzimanolis, Dan E Arking, Nikolaos Smyrnis, Robert M Bilder, Nelson A Freimer, Tyrone D Cannon, Edythe London, Russell A Poldrack, Fred W Sabb, Eliza Congdon, Emily Drabant Conley, Matthew A Scult, Dwight Dickinson, Richard E Straub, Gary Donohoe, Derek Morris, Aiden Corvin, Michael Gill, Ahmad R Hariri, Daniel R Weinberger, Neil Pendleton, Panos Bitsios, Dan Rujescu, Jari Lahti, Stephanie Le Hellard, Matthew C Keller, Ole A Andreassen, Ian J Deary, David C Glahn, Anil K Malhotra, and Todd Lencz. Large-Scale cognitive GWAS Meta-Analysis reveals Tissue-Specific neural expression and potential nootropic drug targets. *Cell Rep.*, 21(9):2597–2613, November 2017.

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

[159] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen

Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, 46(D1):D1062–D1067, January 2018.

[160] Nathan LaPierre, Kodi Taraszka, Helen Huang, Rosemary He, Farhad Hormozdiari, and Eleazar Eskin. Identifying causal variants by fine mapping across multiple studies. *PLoS Genet.*, 17(9):e1009733, September 2021.

[161] Mariana Lazo, Ruben Hernaez, Mark S Eberhardt, Susanne Bonekamp, Ihab Kamel, Eliseo Guallar, Ayman Koteish, Frederick L Brancati, and Jeanne M Clark. Prevalence of nonalcoholic fatty liver disease in the United States: the Third National Health and Nutrition Examination Survey, 1988-1994. *Am. J. Epidemiol.*, 178(1):38–45, July 2013.

[162] James J Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, Mark Alan Fontana, Tushar Kundu, Chanwook Lee, Hui Li, Ruoxi Li, Rebecca Royer, Pascal N Timshel, Raymond K Walters, Emily A Willoughby, Loïc Yengo, 23andMe Research Team, COGENT (Cognitive Genomics Consortium), Social Science Genetic Association Consortium, Maris Alver, Yanchun Bao, David W Clark, Felix R Day, Nicholas A Furlotte, Peter K Joshi, Kathryn E Kemper, Aaron Kleinman, Claudia Langenberg, Reedik Mägi, Joey W Trampush, Shefali Setia Verma, Yang Wu, Max Lam, Jing Hua Zhao, Zhili Zheng, Jason D Boardman, Harry Campbell, Jeremy Freese, Kathleen Mullan Harris, Caroline Hayward, Pamela Herd, Meena Kumari, Todd Lencz, Jian'an Luan, Anil K Malhotra, Andres Metspalu, Lili Milani, Ken K Ong, John R B Perry, David J Porteous, Marylyn D Ritchie, Melissa C Smart, Blair H Smith, Joyce Y Tung, Nicholas J Wareham, James F Wilson, Jonathan P Beauchamp, Dalton C Conley, Tõnu Esko, Steven F Lehrer, Patrik K E Magnusson, Sven Oskarsson, Tune H Pers, Matthew R Robinson, Kevin Thom, Chelsea Watson, Christopher F Chabris, Michelle N Meyer, David I Laibson, Jian Yang, Magnus Johannesson, Philipp D Koellinger, Patrick Turley, Peter M Visscher, Daniel J Benjamin, and David Cesarini. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.*, 50(8):1112–1121, July 2018.

[163] James J Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, et al. Gene discovery and polygenic prediction from a 1.1-million-person gwas of educational attainment. *Nature genetics*, 50(8):1112, 2018.

[164] Su-In Lee, Safiye Celik, Benjamin A Logsdon, Scott M Lundberg, Timothy J Martins, Vivian G Oehler, Elihu H Estey, Chris P Miller, Sylvia Chien, Jin Dai, Akanksha Saxena, C Anthony Blau, and Pamela S Becker. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat. Commun.*, 9(1):42, January 2018.

[165] Tailce K M Leite, Rômulo M C Fonseca, Nanci M de França, Esteban J Parra, and Rinaldo W Pereira. Genomic ancestry, Self-Reported "color" and quantitative measures of skin pigmentation in brazilian admixed siblings. *PLoS One*, 6(11):e27162, November 2011.

[166] Michael G Levin and Daniel J Rader. Polygenic Risk Scores and Coronary Artery Disease: Ready for Prime Time? *Circulation*, 141(8):637–640, February 2020.

[167] Anna C F Lewis, Santiago J Molina, Paul S Appelbaum, Bege Dauda, Anna Di Rienzo, Agustin Fuentes, Stephanie M Fullerton, Nanibaa' A Garrison, Nayanika Ghosh, Evelynn M Hammonds, David S Jones, Eimear E Kenny, Peter Kraft, Sandra S-J Lee, Madelyn Mauro, John Novembre, Aaron Panofsky, Mashaal Sohail, Benjamin M Neale, and Danielle S Allen. Getting genetic ancestry right for science and society. *Science*, 376(6590):250–252, April 2022.

[168] Ruowang Li, Yong Chen, Marylyn D Ritchie, and Jason H Moore. Electronic health records and polygenic risk scores for predicting disease risk. *Nat. Rev. Genet.*, 21(8):493–502, August 2020.

[169] Shanshan Li, Gregg C Fonarow, Kenneth Mukamal, Haolin Xu, Roland A Matsouaka, Adam D Devore, and Deepak L Bhatt. Sex and Racial Disparities in Cardiac Rehabilitation Referral at Hospital Discharge and Gaps in Long-Term Mortality. *J. Am. Heart Assoc.*, 7(8), April 2018.

[170] Yue Li and Manolis Kellis. Joint bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Research*, 44(18):e144–e144, 2016.

[171] Yun R Li and Brendan J Keating. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.*, 6(10):1–14, October 2014.

[172] Dan-Yu Lin, Ran Tao, William D Kalsbeek, Donglin Zeng, Franklyn Gonzalez, Lindsay Fernández-Rhodes, Mariaelisa Graff, Gary G Koch, Kari E North, and Gerardo Heiss. Genetic association analysis under complex survey sampling: The hispanic community health Study/Study of latinos. *Am. J. Hum. Genet.*, 95(6):675–688, December 2014.

[173] Jodell E Linder, Aimee Allworth, Sarah T Bland, Pedro J Caraballo, Rex L Chisholm, Ellen Wright Clayton, David R Crosslin, Ozan Dikilitas, Alanna DiVietro, Edward D Esplin, Sophie Forman, Robert R Freimuth, Adam S Gordon, Richard Green, Maegan V Harden, Ingrid A Holm, Gail P Jarvik, Elizabeth W Karlson, Sofia Labrecque, Niall J Lennon, Nita A Limdi, Kathleen F Mittendorf, Shawn N Murphy, Lori Orlando, Cynthia A Prows, Luke V Rasmussen, Laura Rasmussen-Torvik, Robb Rowley, Konrad Teodor Sawicki, Tara Schmidlen, Shannon Terek, David Veenstra, Digna R Velez Edwards, Devin Absher, Noura S Abul-Husn, Jorge Alsip, Hana Bangash, Mark Beasley, Jennifer E Below, Eta S Berner, James Booth, Wendy K Chung, James J Cimino, John Connolly, Patrick Davis, Beth Devine, Stephanie M Fullerton, Candace

Guiducci, Melissa L Habrat, Heather Hain, Hakon Hakonarson, Margaret Harr, Eden Haverfield, Valentina Hernandez, Christin Hoell, Martha Horike-Pyne, George Hripcsak, Marguerite R Irvin, Christopher Kachulis, Dean Karavite, Eimear E Kenny, Atlas Khan, Krzysztof Kiryluk, Bruce Korf, Leah Kottyan, Iftikhar J Kullo, Katie Larkin, Cong Liu, Edyta Malolepsza, Teri A Manolio, Thomas May, Elizabeth M McNally, Frank Mentch, Alexandra Miller, Sean D Mooney, Priyanka Murali, Brenda Mutai, Naveen Muthu, Bahram Namjou, Emma F Perez, Megan J Puckelwartz, Tejinder Rakhra-Burris, Dan M Roden, Elisabeth A Rosenthal, Seyedmohammad Saadatagah, Maya Sabatello, Dan J Schaid, Baergen Schultz, Lynn Seabolt, Gabriel Q Shaibi, Richard R Sharp, Brian Shirts, Maureen E Smith, Jordan W Smoller, Rene Sterling, Sabrina A Suckiel, Jeritt Thayer, Hemant K Tiwari, Susan B Trinidad, Theresa Walunas, Wei-Qi Wei, Quinn S Wells, Chunhua Weng, Georgia L Wiesner, Ken Wiley, eMERGE Consortium, and Josh F Peterson. Returning integrated genomic risk and clinical recommendations: The eMERGE study. *Genet. Med.*, 25(4):100006, January 2023.

[174] Sara Lindström, Lu Wang, Erin N. Smith, William Gordon, Astrid van Hylckama Vlieg, Mariza de Andrade, Jennifer A. Brody, Jack W. Pattee, Jeffrey Haessler, Ben M. Brumpton, Daniel I. Chasman, Pierre Suchon, Ming-Huei Chen, Constance Turman, Marine Germain, Kerri L. Wiggins, James MacDonald, Sigrid K. Braekkan, Sebastian M. Armasu, Nathan Pankratz, Rebecca D. Jackson, Jonas B. Nielsen, Franco Giulianini, Marja K. Puurunen, Manal Ibrahim, Susan R. Heckbert, Scott M. Damrauer, Pradeep Natarajan, Derek Klarin, Million Veteran Program, Paul S. de Vries, Maria Sabater-Lleal, Jennifer E. Huffman, CHARGE Hemostasis Working Group, Theo K. Bammler, Kelly A. Frazer, Bryan M. McCauley, Kent Taylor, James S. Pankow, Alexander P. Reiner, Maiken E. Gabrielsen, Jean-François Deleuze, Chris J. O'Donnell, Jihye Kim, Barbara McKnight, Peter Kraft, John-Bjarne Hansen, Frits R. Rosendaal, John A. Heit, Bruce M. Psaty, Weihong Tang, Charles Kooperberg, Kristian Hveem, Paul M. Ridker, Pierre-Emmanuel Morange, Andrew D. Johnson, Christopher Kabrhel, David-Alexandre Trégouët, and Nicholas L. Smith. Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. *Blood*, 134(19):1645–1657, November 2019.

[175] Richard Karlsson Linnér, Pietro Biroli, Edward Kong, S Fleur W Meddens, Robbee Wedow, Mark Alan Fontana, Maël Lebreton, Stephen P Tino, Abdel Abdellaoui, Anke R Hammerschlag, et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature genetics*, 51(2):245, 2019.

[176] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. Improved linear mixed models for genome-wide association studies. *Nat. Methods*, 9(6):525–526, May 2012.

[177] Eric Yi Liu, Mingyao Li, Wei Wang, and Yun Li. MaCH-admix: genotype imputation for admixed populations. *Genet. Epidemiol.*, 37(1):25–37, January 2013.

[178] Xuanyao Liu, Yang I Li, and Jonathan K Pritchard. Trans effects on gene expression can drive omnigenic inheritance. *Cell*, 177(4):1022–1034, 2019.

[179] Luke R Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tonu Esko, et al. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature communications*, 10(1):1–11, 2019.

[180] Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics*, 47(12):1385, 2015.

[181] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, Richard Durbin, and Alkes L Price. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.*, 48(11):1443–1448, November 2016.

[182] Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P Schoech, and Alkes L Price. Mixed-model association for biobank-scale datasets. *Nat. Genet.*, 50(7):906–908, June 2018.

[183] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, and Alkes L Price. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, 47(3):284–290, February 2015.

[184] Kirk E Lohmueller, Matthew M Mauney, David Reich, and John M Braverman. Variants associated with common disease are not unusually differentiated in frequency across populations. *The American Journal of Human Genetics*, 78(1):130–136, 2006.

[185] Saioa López, Ayele Tarekegn, Gavin Band, Lucy van Dorp, Nancy Bird, Sam Morris, Tamiru Oljira, Ephrem Mekonnen, Endashaw Bekele, Roger Blench, Mark G Thomas, Neil Bradman, and Garrett Hellenthal. Evidence of the interplay of genetics and culture in ethiopia. *Nat. Commun.*, 12(1):1–15, June 2021.

[186] H L Lujan and S E DiCarlo. The racist “one drop rule” influencing science: it is time to stop teaching “race corrections” in medicine. *Adv. Physiol. Educ.*, 45(3), September 2021.

[187] Andrew MacGinnitie, Frank Aloi, and Seema Mishra. Clinical characteristics of pediatric patients evaluated for primary immunodeficiency. *Pediatr. Allergy Immunol.*, 22(7):671–675, November 2011.

[188] Thiago Magalhães da Silva, M R Sandhya Rani, Gustavo Nunes de Oliveira Costa, Maria A Figueiredo, Paulo S Melo, João F Nascimento, Neil D Molyneaux, Maurício L Barreto, Mitermayer G Reis, M Glória Teixeira, and Ronald E Blanton. The correlation between ancestry and color in two cities of northeast brazil with contrasting ethnic compositions. *Eur. J. Hum. Genet.*, 23(7):984–989, October 2014.

[189] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, Pontus Skoglund, Iosif Lazaridis, Sriram Sankararaman, Qiaomei Fu, Nadin Rohland, Gabriel Renaud, Yaniv Erlich, Thomas Willems, Carla Gallo, Jeffrey P Spence, Yun S Song, Giovanni Poletti, Francois Balloux, George van Driem, Peter de Knijff, Irene Gallego Romero, Aashish R Jha, Doron M Behar, Claudio M Bravi, Cristian Capelli, Tor Hervig, Andres Moreno-Estrada, Olga L Posukh, Elena Balanovska, Oleg Balanovsky, Sena Karachanak-Yankova, Hovhannes Sahakyan, Draga Toncheva, Levon Yepiskoposyan, Chris Tyler-Smith, Yali Xue, M Syafiq Abdullah, Andres Ruiz-Linares, Cynthia M Beall, Anna Di Rienzo, Choongwon Jeong, Elena B Starikovskaya, Ene Metspalu, Jüri Parik, Richard Villems, Brenna M Henn, Ugur Hodoglugil, Robert Mahley, Antti Sajantila, George Stamatoyannopoulos, Joseph T S Wee, Rita Khusainova, Elza Khusnutdinova, Sergey Litvinov, George Ayodo, David Comas, Michael F Hammer, Toomas Kivisild, William Klitz, Cheryl A Winkler, Damian Labuda, Michael Bamshad, Lynn B Jorde, Sarah A Tishkoff, W Scott Watkins, Mait Metspalu, Stanislav Dryomov, Rem Sukernik, Lalji Singh, Kumarasamy Thangaraj, Svante Pääbo, Janet Kelso, Nick Patterson, and David Reich. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, October 2016.

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

[191] Arjun K Manrai, Birgit H Funke, Heidi L Rehm, Morten S Olesen, Bradley A Maron, Peter Szolovits, David M Margulies, Joseph Loscalzo, and Isaac S Kohane. Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.*, 375(7):655–665, August 2016.

[192] Urko M Marigorta and Arcadi Navarro. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.*, 9(6):e1003566, June 2013.

[193] Nina Mars, Elisabeth Widén, Sini Kerminen, Tuomo Meretoja, Matti Pirinen, Pietro Della Briotta Parolo, Priit Palta, FinnGen, Aarno Palotie, Jaakko Kaprio, Heikki Joensuu, Mark Daly, and Samuli Ripatti. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat. Commun.*, 11(1):6383, December 2020.

[194] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.*, 100(4):635–649, April 2017.

[195] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.*, 51(4):584–591, April 2019.

[196] Iain Mathieson and Aylwyn Scally. What is ancestry? *PLoS Genet.*, 16(3):e1008624, March 2020.

[197] Iain Mathieson and Aylwyn Scally. What is ancestry? *PLoS Genet.*, 16(3):e1008624, March 2020.

[198] Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P Tyrer, Ting-Huei Chen, Qin Wang, Manjeet K Bolla, Xin Yang, Muriel A Adank, Thomas Ahearn, Kristiina Aittomäki, Jamie Allen, Irene L Andrulis, Hoda Anton-Culver, Natalia N Antonenkova, Volker Arndt, Kristan J Aronson, Paul L Auer, Päivi Auvinen, Myrto Barrdahl, Laura E Beane Freeman, Matthias W Beckmann, Sabine Behrens, Javier Benitez, Marina Bermisheva, Leslie Bernstein, Carl Blomqvist, Natalia V Bogdanova, Stig E Bojesen, Bernardo Bonanni, Anne-Lise Børresen-Dale, Hiltrud Brauch, Michael Bremer, Hermann Brenner, Adam Brentnall, Ian W Brock, Angela Brooks-Wilson, Sara Y Brucker, Thomas Brüning, Barbara Burwinkel, Daniele Campa, Brian D Carter, Jose E Castelao, Stephen J Chanock, Rowan Chlebowski, Hans Christiansen, Christine L Clarke, J Margriet Collée, Emilie Cordina-Duverger, Sten Cornelissen, Fergus J Couch, Angela Cox, Simon S Cross, Kamila Czene, Mary B Daly, Peter Devilee, Thilo Dörk, Isabel Dos-Santos-Silva, Martine Dumont, Lorraine Durcan, Miriam Dwek, Diana M Eccles, Arif B Ekici, A Heather Eliassen, Carolina Ellberg, Christoph Engel, Mikael Eriksson, D Gareth Evans, Peter A Fasching, Jonine Figueroa, Olivia Fletcher, Henrik Flyger, Asta Försti, Lin Fritschi, Marike Gabrielson, Manuela Gago-Dominguez, Susan M Gapstur, José A García-Sáenz, Mia M Gaudet, Vassilios Georgoulias, Graham G Giles, Irina R Gilyazova, Gord Glendon, Mark S Goldberg, David E Goldgar, Anna González-Neira, Grethe I Grenaker Alnæs, Mervi Grip, Jacek Gronwald, Anne Grundy, Pascal Guénel, Lothar Haeberle, Eric Hahnen, Christopher A Haiman, Niclas Håkansson, Ute Hamann, Susan E Hankinson, Elaine F Harkness, Steven N Hart, Wei He, Alexander Hein, Jane Heyworth, Peter Hillemanns, Antoinette Hollestelle, Maartje J Hooning, Robert N Hoover, John L Hopper, Anthony Howell, Guanmengqian Huang, Keith Humphreys, David J Hunter, Milena Jakimovska, Anna Jakubowska, Wolfgang Janni, Esther M John, Nichola Johnson, Michael E Jones, Arja Jukkola-Vuorinen, Audrey Jung, Rudolf Kaaks, Katarzyna Kaczmarek, Vesa Kataja, Renske Keeman, Michael J Kerin, Elza Khusnutdinova, Johanna I Kiiski, Julia A Knight, Yon-Dschun Ko, Veli-Matti Kosma, Stella Koutros, Vessela N Kristensen, Ute Krüger, Tabea Kühl, Diether Lambrechts, Loic Le Marchand, Eunjung Lee, Flavio Lejbkowicz, Jenna Lilyquist, Annika Lindblom, Sara Lindström, Jolanta Lissowska, Wing-Yee Lo, Sibylle Loibl, Jirong Long, Jan Lubiński, Michael P Lux, Robert J MacInnis, Tom Maishman, Enes Makalic, Ivana Maleva Kostovska, Arto Mannermaa, Siranoush Manoukian, Sara Margolin, John W M Martens, Maria Elena Martinez, Dimitrios Mavroudis, Catriona McLean, Alfons Meindl, Usha Menon, Pooja Middha, Nicola Miller, Fernando Moreno, Anna Marie

Mulligan, Claire Mulot, Victor M Muñoz-Garzon, Susan L Neuhausen, Heli Nevanlinna, Patrick Neven, William G Newman, Sune F Nielsen, Børge G Nordestgaard, Aaron Norman, Kenneth Offit, Janet E Olson, Håkan Olsson, Nick Orr, V Shane Pankratz, Tjoung-Won Park-Simon, Jose I A Perez, Clara Pérez-Barrios, Paolo Peterlongo, Julian Peto, Mila Pinchev, Dijana Plaseska-Karanfilska, Eric C Polley, Ross Prentice, Nadege Presneau, Darya Prokofyeva, Kristen Purrington, Katri Pylkäs, Brigitte Rack, Paolo Radice, Rohini Rau-Murthy, Gad Rennert, Hedy S Rennert, Valerie Rhenius, Mark Robson, Atocha Romero, Kathryn J Ruddy, Matthias Ruebner, Emmanouil Saloustros, Dale P Sandler, Elinor J Sawyer, Daniel F Schmidt, Rita K Schmutzler, Andreas Schneeweiss, Minouk J Schoemaker, Fredrick Schumacher, Peter Schürmann, Lukas Schwentner, Christopher Scott, Rodney J Scott, Caroline Seynaeve, Mitul Shah, Mark E Sherman, Martha J Shrubsole, Xiao-Ou Shu, Susan Slager, Ann Smeets, Christof Sohn, Penny Soucy, Melissa C Southey, John J Spinelli, Christa Stegmaier, Jennifer Stone, Anthony J Swerdlow, Rulla M Tamimi, William J Tapper, Jack A Taylor, Mary Beth Terry, Kathrin Thöne, Rob A E M Tollenaar, Ian Tomlinson, Thérèse Truong, Maria Tzardi, Hans-Ulrich Ulmer, Michael Untch, Celine M Vachon, Elke M van Veen, Joseph Vijai, Clarice R Weinberg, Camilla Wendt, Alice S Whittemore, Hans Wildiers, Walter Willett, Robert Winqvist, Alicja Wolk, Xiaohong R Yang, Drakoulis Yannoukakos, Yan Zhang, Wei Zheng, Argyrios Ziogas, ABCTB Investigators, kConFab/AOCS Investigators, NBCS Collaborators, Alison M Dunning, Deborah J Thompson, Georgia Chenevix-Trench, Jenny Chang-Claude, Marjanka K Schmidt, Per Hall, Roger L Milne, Paul D P Pharoah, Antonis C Antoniou, Nilanjan Chatterjee, Peter Kraft, Montserrat García-Closas, Jacques Simard, and Douglas F Easton. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.*, 104(1):21–34, January 2019.

[199] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, Yang Luo, Carlo Sidore, Alan Kwong, Nicholas Timpson, Seppo Koskinen, Scott Vrieze, Laura J Scott, He Zhang, Anubha Mahajan, Jan Veldink, Ulrike Peters, Carlos Pato, Cornelia M van Duijn, Christopher E Gillies, Ilaria Gandin, Massimo Mezzavilla, Arthur Gilly, Massimiliano Cocca, Michela Traglia, Andrea Angius, Jeffrey C Barrett, Dorrett Boomsma, Kari Branham, Gerome Breen, Chad M Brummett, Fabio Busonero, Harry Campbell, Andrew Chan, Sai Chen, Emily Chew, Francis S Collins, Laura J Corbin, George Davey Smith, George Dedoussis, Marcus Dorr, Aliki-Eleni Farmaki, Luigi Ferrucci, Lukas Forer, Ross M Fraser, Stacey Gabriel, Shawn Levy, Leif Groop, Tabitha Harrison, Andrew Hattersley, Oddgeir L Holmen, Kristian Hveem, Matthias Kretzler, James C Lee, Matt McGue, Thomas Meitinger, David Melzer, Josine L Min, Karen L Mohlke, John B Vincent, Matthias Nauck, Deborah Nickerson, Aarno Palotie, Michele Pato, Nicola Pirastu, Melvin McInnis, J Brent Richards, Cinzia Sala, Veikko Salomaa, David Schlessinger, Sebastian Schoenherr, P Eline Slagboom, Kerrin Small, Timothy Spector, Dwight Stambolian, Marcus Tuke, Jaakko Tuomilehto, Leonard H Van den Berg, Wouter Van Rheenen, Uwe Volker, Cisca Wijmenga, Daniela Toniolo, Eleftheria Zeggini, Paolo Gasparini, Matthew G Samp-

son, James F Wilson, Timothy Frayling, Paul I W de Bakker, Morris A Swertz, Steven McCarroll, Charles Kooperberg, Annelot Dekker, David Altshuler, Cristen Willer, William Iacono, Samuli Ripatti, Nicole Soranzo, Klaudia Walter, Anand Swaroop, Francesco Cucca, Carl A Anderson, Richard M Myers, Michael Boehnke, Mark I McCarthy, Richard Durbin, Gonçalo Abecasis, Jonathan Marchini, and the Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10):1279–1283, October 2016. Number: 10 Publisher: Nature Publishing Group.

[200] Meghan E McGarry, Clement L Ren, Runyu Wu, Philip M Farrell, and Susanna A McColley. Detection of disease-causing CFTR variants in state newborn screening programs. *Pediatr. Pulmonol.*, 58(2):465–474, February 2023.

[201] Victor A McKusick. Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.*, 80(4):588–604, April 2007.

[202] Gil McVean. A genealogical interpretation of principal components analysis. *PLoS Genet.*, 5(10):e1000686, October 2009.

[203] Graham McVicker, David Gordon, Colleen Davis, and Phil Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics*, 5(5), 2009.

[204] Carolina Medina-Gomez, Janine Frédérique Felix, Karol Estrada, Marjoline Josephine Peters, Lizbeth Herrera, Claudia Jeanette Kruithof, Liesbeth Duijts, Albert Hofman, Cornelia Marja van Duijn, Andreas Gerardus Uitterlinden, Vincent Wilfred Vishal Jaddoe, and Fernando Rivadeneira. Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the generation R study. *Eur. J. Epidemiol.*, 30(4):317–330, March 2015.

[205] Melinda C Mills and Charles Rahal. The GWAS diversity monitor tracks diversity by disease in real time. *Nat. Genet.*, 52(3):242–243, March 2020.

[206] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

[207] Vicki Modell, Bonnie Gee, David B Lewis, Jordan S Orange, Chaim M Roifman, John M Routes, Ricardo U Sorensen, Luigi D Notarangelo, and Fred Modell. Global study of primary immunodeficiency diseases (PI)–diagnosis, treatment, and economic impact: an updated report from the jeffrey modell foundation. *Immunol. Res.*, 51(1):61–70, October 2011.

[208] Vicki Modell, Jessica Quinn, Grant Ginsberg, Ron Gladue, Jordan Orange, and Fred Modell. Modeling strategy to identify patients with primary immunodeficiency utilizing risk management and outcome measurement. *Immunol. Res.*, 65(3):713–720, June 2017.

[209] Theodore J Morley, Lide Han, Victor M Castro, Jonathan Morra, Roy H Perlis, Nancy J Cox, Lisa Bastarache, and Douglas M Ruderfer. Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. *Nat. Med.*, 27(6):1097–1104, June 2021.

[210] Theodore J Morley, Lide Han, Victor M Castro, Jonathan Morra, Roy H Perlis, Nancy J Cox, Lisa Bastarache, and Douglas M Ruderfer. Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. *Nat. Med.*, 27(6):1097–1104, June 2021.

[211] Gerhard Moser, Sang Hong Lee, Ben J Hayes, Michael E Goddard, Naomi R Wray, and Peter M Visscher. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS genetics*, 11(4):e1004969, 2015.

[212] Vivek H Murthy, Harlan M Krumholz, and Cary P Gross. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *JAMA*, 291(22):2720–2726, June 2004.

[213] Arash Naeim, Sarah Dry, David Elashoff, Zhuoer Xie, Antonia Petruse, Clara Magyar, Liliana Johansen, Gabriela Werre, Clara Lajonchere, and Neil Wenger. Electronic Video Consent to Power Precision Health Research: A Pilot Cohort Study. *JMIR Form Res*, 5(9):e29123, September 2021.

[214] Akiko Nagai, Makoto Hirata, Yoichiro Kamatani, Kaori Muto, Koichi Matsuda, Yutaka Kiyohara, Toshiharu Ninomiya, Akiko Tamakoshi, Zentaro Yamagata, Taisei Mushiroda, Yoshinori Murakami, Koichiro Yuji, Yoichi Furukawa, Hitoshi Zembutsu, Toshihiro Tanaka, Yozo Ohnishi, Yusuke Nakamura, BioBank Japan Cooperative Hospital Group, and Michiaki Kubo. Overview of the BioBank japan project: Study design and profile. *J. Epidemiol.*, 27(3S):S2–S8, March 2017.

[215] Aditi Nayak, Albert J Hicks, and Alanna A Morris. Understanding the Complexity of Heart Failure Risk and Treatment in Black Patients. *Circ. Heart Fail.*, 13(8):e007264, August 2020.

[216] Sheryl Hui-Xian Ng, Nabilah Rahman, Ian Yi Han Ang, Srinath Sridharan, Sravan Ramachandran, Debby D Wang, Chuen Seng Tan, Sue-Anne Toh, and Xin Quan Tan. Characterization of high healthcare utilizer groups using administrative data from an electronic medical record database. *BMC health services research*, 19(1):1–14, 2019.

[217] Emily T Norris, Lu Wang, Andrew B Conley, Lavanya Rishishwar, Leonardo Mariño-Ramírez, Augusto Valderrama-Aguirre, and I King Jordan. Genetic ancestry, admixture and health determinants in Latin America. *BMC Genomics*, 19(Suppl 8):861, December 2018.

[218] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, Matthew Stephens, and Carlos D Bustamante. Genes mirror geography within europe. *Nature*, 456(7218):98–101, November 2008.

[219] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, Matthew Stephens, and Carlos D Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, November 2008.

[220] Evangelia E Ntzani, George Liberopoulos, Teri A Manolio, and John P A Ioannidis. Consistency of genome-wide associations across major ancestral groups. *Hum. Genet.*, 131(7):1057–1071, December 2011.

[221] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[222] Luke J O'Connor, Armin P Schoech, Farhad Hormozdiari, Steven Gazal, Nick Patterson, and Alkes L Price. Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105(3):456–476, 2019.

[223] Kimberly J O'Malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: ICD code accuracy. *Health Serv. Res.*, 40(5 Pt 2):1620–1639, October 2005.

[224] Jordan S Orange, Joseph T Glessner, Elena Resnick, Kathleen E Sullivan, Mary Lucas, Berne Ferry, Cecilia E Kim, Cuiping Hou, Fengxiang Wang, Rosetta Chiavacci, Subra Kugathasan, John W Sleasman, Robert Baldassano, Elena E Perez, Helen Chapel, Charlotte Cunningham-Rundles, and Hakon Hakonarson. Genome-wide association identifies diverse causes of common variable immunodeficiency. *J. Allergy Clin. Immunol.*, 127(6):1360–7.e6, June 2011.

[225] Michael D O'Sullivan and Andrew J Cant. The 10 warning signs, 2012.

[226] Aaron Panofsky, Kushan Dasgupta, and Nicole Iturriaga. How white nationalists mobilize genetics: From genetic ancestry and human biodiversity to counterscience and metapolitics. *Am. J. Phys. Anthropol.*, 175(2):387–398, June 2021.

[227] Zachary Parolin and Emma K Lee. The role of poverty and racial discrimination in exacerbating the health consequences of COVID-19. *Lancet Reg Health Am*, 7:100178, March 2022.

[228] Bogdan Pasaniuc and Alkes L Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 18(2):117, 2017.

[229] Roseann E Peterson, Karoline Kuchenbaecker, Raymond K Walters, Chia-Yen Chen, Alice B Popejoy, Sathish Periyasamy, Max Lam, Conrad Iyegbe, Rona J Strawbridge, Leslie Brick, Caitlin E Carey, Alicia R Martin, Jacquelyn L Meyers, Jinni Su, Junfang Chen, Alexis C Edwards, Allan Kalungi, Nastassja Koen, Lerato Majara, Emanuel Schwarz, Jordan W Smoller, Eli A Stahl, Patrick F Sullivan, Evangelos Vassos, Bryan Mowry, Miguel L Prieto, Alfredo Cuellar-Barboza, Tim B Bigdeli, Howard J Edenberg,

Hailiang Huang, and Laramie E Duncan. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*, 179(3):589–603, October 2019.

[230] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573, 2014.

[231] Maria Pino-Yanes, Neeta Thakur, Christopher R Gignoux, Joshua M Galanter, Lindsey A Roth, Celeste Eng, Katherine K Nishimura, Sam S Oh, Hita Vora, Scott Huntsman, Elizabeth A Nguyen, Donglei Hu, Katherine A Drake, David V Conti, Andres Moreno-Estrada, Karla Sandoval, Cheryl A Winkler, Luisa N Borrell, Fred Lurmann, Talat S Islam, Adam Davis, Harold J Farber, Kelley Meade, Pedro C Avila, Denise Serebrisky, Kirsten Bibbins-Domingo, Michael A Lenoir, Jean G Ford, Emerita Brigino-Buenaventura, William Rodriguez-Cintron, Shannon M Thyne, Saunak Sen, Jose R Rodriguez-Santana, Carlos D Bustamante, L Keoki Williams, Frank D Gilliland, W James Gauderman, Rajesh Kumar, Dara G Torgerson, and Esteban G Burchard. Genetic ancestry influences asthma susceptibility and lung function among latinos. *J. Allergy Clin. Immunol.*, 135(1):228–235, January 2015.

[232] Matthias Pinter, Michael Trauner, Markus Peck-Radosavljevic, and Wolfgang Sieghart. Cancer and liver cirrhosis: implications on prognosis and management. *ESMO Open*, 1(2):e000042, March 2016.

[233] Matti Pirinen, Peter Donnelly, and Chris C A Spencer. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat. Genet.*, 44(8):848–851, July 2012.

[234] Giorgio Pistis, Eleonora Porcu, Scott I. Vrieze, Carlo Sidore, Maristella Steri, Fabrice Danjou, Fabio Busonero, Antonella Mulas, Magdalena Zoledziewska, Andrea Maschio, Christine Brennan, Sandra Lai, Michael B. Miller, Marco Marcelli, Maria Francesca Urru, Maristella Pitzalis, Robert H. Lyons, Hyun M. Kang, Chris M. Jones, Andrea Angius, William G. Iacono, David Schlessinger, Matt McGue, Francesco Cucca, Gonçalo R. Abecasis, and Serena Sanna. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *European journal of human genetics: EJHG*, 23(7):975–983, July 2015.

[235] Sharon E Plon and Heidi L Rehm. The ancestral pace of variant reclassification. *JNCI Journal of the National Cancer Institute*, 110(10):1133, October 2018.

[236] Mica Pollock, editor. *Everyday Antiracism: Getting Real About Race in School.* The New Press, New York, 10757th edition edition, June 2008.

[237] Ana Cecilia Pontoriero, Julieta Trinks, María Laura Hulaniuk, Mariela Caputo, Lisandro Fortuny, Leandro Burgos Pratx, Analía Frías, Oscar Torres, Félix Nuñez, Adrián Gadano, Pablo Argibay, Daniel Corach, and Diego Flichman. Influence of ethnicity on

the distribution of genetic polymorphisms associated with risk of chronic liver disease in South American populations. *BMC Genet.*, 16:93, July 2015.

[238] Alice B Popejoy. Too many scientists still say Caucasian, August 2021. Publication Title: Nature Publishing Group UK.

[239] Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature*, 538(7624):161–164, October 2016.

[240] Florian Privé, Hugues Aschard, Shai Carmi, Lasse Folkersen, Clive Hoggart, Paul F O'Reilly, and Bjarni J Vilhjálmsson. Portability of 245 polygenic scores when derived from the UK biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.*, 109(2):373, February 2022.

[241] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H Shah, Atul J Butte, Michael D Howell, Claire Cui, Greg S Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*, 1:18, May 2018.

[242] Paula S Ramos, Lindsey A Criswell, Kathy L Moser, Mary E Comeau, Adrienne H Williams, Nicholas M Pajewski, Sharon A Chung, Robert R Graham, Raphael Zidovetzki, Jennifer A Kelly, et al. A comprehensive analysis of shared loci between systemic lupus erythematosus (sle) and sixteen autoimmune diseases reveals limited genetic overlap. *PLoS genetics*, 7(12):e1002406, 2011.

[243] Shereen M Reda, Dalia H El-Ghoneimy, and Hanaa M Afifi. Clinical predictors of primary immunodeficiency diseases in children. *Allergy Asthma Immunol. Res.*, 5(2):88–95, March 2013.

[244] David Reich, Michael A Nalls, W H Linda Kao, Ermeg L Akylbekova, Arti Tandon, Nick Patterson, James Mullikin, Wen-Chi Hsueh, Ching-Yu Cheng, Josef Coresh, Eric Boerwinkle, Man Li, Alicja Waliszewska, Julie Neubauer, Rongling Li, Tennille S Leak, Lynette Ekunwe, Joe C Files, Cheryl L Hardy, Joseph M Zmuda, Herman A Taylor, Elad Ziv, Tamara B Harris, and James G Wilson. Reduced neutrophil count in people of african descent is due to a regulatory variant in the duffy antigen receptor for chemokines gene. *PLoS Genet.*, 5(1):e1000360, January 2009.

[245] Emily M Rencsok, Latifa A Bazzi, Rana R McKay, Franklin W Huang, Adam Friedant, Jake Vinson, Samuel Peisch, Jelani C Zarif, Stacey Simmons, Kelly Hawthorne, Paul Villanti, Philip W Kantoff, Elisabeth Heath, Daniel J George, and Lorelei A Mucci. Diversity of Enrollment in Prostate Cancer Clinical Trials: Current Status and Future Directions. *Cancer Epidemiol. Biomarkers Prev.*, 29(7):1374–1380, July 2020.

[246] Corinne Richard-Miceli and Lindsey A Criswell. Emerging patterns of genetic overlap across autoimmune disorders. *Genome medicine*, 4(1):6, 2012.

[247] Nicholas L Rider, Di Miao, Margaret Dodds, Vicki Modell, Fred Modell, Jessica Quinn, Heidi Schwarzwald, and Jordan S Orange. Calculation of a primary immunodeficiency "risk vital sign" via Population-Wide analysis of claims data to aid in clinical decision support. *Front Pediatr*, 7:70, March 2019.

[248] Cornelius A Rietveld, Sarah E Medland, Jaime Derringer, Jian Yang, Tõnu Esko, Nicolas W Martin, Harm-Jan Westra, Konstantin Shakhbazov, Abdel Abdellaoui, Arpana Agrawal, Eva Albrecht, Behrooz Z Alizadeh, Najaf Amin, John Barnard, Sebastian E Baumeister, Kelly S Benke, Lawrence F Bielak, Jeffrey A Boatman, Patricia A Boyle, Gail Davies, Christiaan de Leeuw, Niina Eklund, Daniel S Evans, Rudolf Ferhmann, Krista Fischer, Christian Gieger, Håkon K Gjessing, Sara Hägg, Jennifer R Harris, Caroline Hayward, Christina Holzapfel, Carla A Ibrahim-Verbaas, Erik Ingelsson, Bo Jacobsson, Peter K Joshi, Astanand Jugessur, Marika Kaakinen, Stavroula Kanoni, Juha Karjalainen, Ivana Kolcic, Kati Kristiansson, Zoltán Kutalik, Jari Lahti, Sang H Lee, Peng Lin, Penelope A Lind, Yongmei Liu, Kurt Lohman, Marisa Loitfelder, George McMahon, Pedro Marques Vidal, Osorio Meirelles, Lili Milani, Ronny Myhre, Marja-Liisa Nuotio, Christopher J Oldmeadow, Katja E Petrovic, Wouter J Peyrot, Ozren Polasek, Lydia Quaye, Eva Reinmaa, John P Rice, Thais S Rizzi, Helena Schmidt, Reinhold Schmidt, Albert V Smith, Jennifer A Smith, Toshiko Tanaka, Antonio Terracciano, Matthijs J H M van der Loos, Veronique Vitart, Henry Völzke, Jürgen Wellmann, Lei Yu, Wei Zhao, Jüri Allik, John R Attia, Stefania Bandinelli, François Bastardot, Jonathan Beauchamp, David A Bennett, Klaus Berger, Laura J Bierut, Dorret I Boomsma, Ute Bültmann, Harry Campbell, Christopher F Chabris, Lynn Cherkas, Mina K Chung, Francesco Cucca, Mariza de Andrade, Philip L De Jager, Jan-Emmanuel De Neve, Ian J Deary, George V Dedoussis, Panos Deloukas, Maria Dimitriou, Guny Eiríksdóttir, Martin F Elderson, Johan G Eriksson, David M Evans, Jessica D Faul, Luigi Ferrucci, Melissa E Garcia, Henrik Grönberg, Vilmundur Gunason, Per Hall, Juliette M Harris, Tamara B Harris, Nicholas D Hastie, Andrew C Heath, Dena G Hernandez, Wolfgang Hoffmann, Adriaan Hofman, Rolf Holle, Elizabeth G Holliday, Jouke-Jan Hottenga, William G Iacono, Thomas Illig, Marjo-Riitta Järvelin, Mika Kähönen, Jaakko Kaprio, Robert M Kirkpatrick, Matthew Kowgier, Antti Latvala, Lenore J Launer, Debbie A Lawlor, Terho Lehtimäki, Jingmei Li, Paul Lichtenstein, Peter Lichtner, David C Liewald, Pamela A Madden, Patrik K E Magnusson, Tomi E Mäkinen, Marco Masala, Matt McGue, Andres Metspalu, Andreas Mielck, Michael B Miller, Grant W Montgomery, Sutapa Mukherjee, Dale R Nyholt, Ben A Oostra, Lyle J Palmer, Aarno Palotie, Brenda W J H Penninx, Markus Perola, Patricia A Peyser, Martin Preisig, Katri Räikkönen, Olli T Raitakari, Anu Realo, Susan M Ring, Samuli Ripatti, Fernando Rivadeneira, Igor Rudan, Aldo Rustichini, Veikko Salomaa, Antti-Pekka Sarin, David Schlessinger, Rodney J Scott, Harold Snieder, Beate St Pourcain, John M Starr, Jae Hoon Sul, Ida Surakka, Rauli Svento, Alexander Teumer, LifeLines Cohort Study, Henning Tiemeier, Frank J A van Rooij, David R Van Wagoner, Erkki Vartiainen, Jorma Viikari, Peter Vollenweider, Judith M

Vonk, Gérard Waeber, David R Weir, H-Erich Wichmann, Elisabeth Widen, Gonneke Willemsen, James F Wilson, Alan F Wright, Dalton Conley, George Davey-Smith, Lude Franke, Patrick J F Groenen, Albert Hofman, Magnus Johannesson, Sharon L R Kardia, Robert F Krueger, David Laibson, Nicholas G Martin, Michelle N Meyer, Danielle Posthuma, A Roy Thurik, Nicholas J Timpson, André G Uitterlinden, Cornelia M van Duijn, Peter M Visscher, Daniel J Benjamin, David Cesarini, and Philipp D Koellinger. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, 340(6139):1467–1471, June 2013.

[249] Neil Risch, Esteban Burchard, Elad Ziv, and Hua Tang. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol.*, 3(7):comment2007, July 2002.

[250] Neil Risch, Esteban Burchard, Elad Ziv, and Hua Tang. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol.*, 3(7):comment2007, July 2002.

[251] Kundaje Roadmap Epigenomics Consortium, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Pouya Kheradpour, Zhizhuo Zhang, Alireza Heravi-Moussavi, Yaping Liu, Viren Amin, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015.

[252] D. M. Roden, J. M. Pulley, M. A. Basford, G. R. Bernard, E. W. Clayton, J. R. Balser, and D. R. Masys. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical Pharmacology and Therapeutics*, 84(3):362–369, September 2008.

[253] Howard W Rogers, Martin A Weinstock, Steven R Feldman, and Brett M Coldiron. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the US Population, 2012. *JAMA Dermatol.*, 151(10):1081–1086, October 2015. Publisher: American Medical Association.

[254] Caryn Roth, Randi E Foraker, Philip R O Payne, and Peter J Embi. Community-level determinants of obesity: harnessing the power of electronic health records for retrospective data analysis. *BMC Med. Inform. Decis. Mak.*, 14(1):1–8, May 2014.

[255] Wendy D Roth and Biorn Ivemark. Genetic options: The impact of genetic ancestry testing on consumers' racial and ethnic identities1. *Am. J. Sociol.*, July 2018.

[256] Naga Swetha Samji, Peter D Snell, Ashwani K Singal, and Sanjaya K Satapathy. Racial disparities in diagnosis and prognosis of nonalcoholic fatty liver disease. *Clin. Liver Dis.*, 16(2):66–72, August 2020. Publisher: Wiley.

[257] Saskia C Sanderson, Kyle B Brothers, Nathaniel D Mercaldo, Ellen Wright Clayton, Armand H Matheny Antommaria, Sharon A Aufox, Murray H Brilliant, Diego Campos, David S Carrell, John Connolly, Pat Conway, Stephanie M Fullerton, Nanibaa' A Garrison, Carol R Horowitz, Gail P Jarvik, David Kaufman, Terrie E Kitchner, Rongling Li,

Evette J Ludman, Catherine A McCarty, Jennifer B McCormick, Valerie D McManus, Melanie F Myers, Aaron Scrol, Janet L Williams, Martha J Shrubsole, Jonathan S Schildcrout, Maureen E Smith, and Ingrid A Holm. Public Attitudes toward Consent and Data Sharing in Biobank Research: A Large Multi-site Experimental Survey in the US. *Am. J. Hum. Genet.*, 100(3):414–427, March 2017.

[258] Sharon A Savage, Meg R Gerstenblith, Alisa M Goldstein, Lisa Mirabello, Maria Concetta Fargnoli, Ketty Peris, and Maria Teresa Landi. Nucleotide diversity and population differentiation of the melanocortin 1 receptor gene, MC1R. *BMC Genet.*, 9(1):1–8, April 2008.

[259] Enrique F Schisterman, Stephen R Cole, and Robert W Platt. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20(4):488–495, July 2009.

[260] K A Schulman, J A Berlin, W Harless, J F Kerner, S Sistrunk, B J Gersh, R Dubé, C K Taleghani, J E Burke, S Williams, J M Eisenberg, and J J Escarce. The effect of race and sex on physicians' recommendations for cardiac catheterization. *N. Engl. J. Med.*, 340(8):618–626, February 1999.

[261] Robert C Schwartz and David M Blankenship. Racial disparities in psychotic disorder diagnosis: A review of empirical literature. *World Journal of Psychiatry*, 4(4):133, December 2014.

[262] Marco Scutari, Ian Mackay, and David Balding. Using genetic distance to infer the accuracy of genomic prediction. *PLoS Genet.*, 12(9):e1006288, September 2016.

[263] Jodi B. Segal, John Eng, Leonardo J. Tamariz, and Eric B. Bass. Review of the evidence on diagnosis of deep venous thrombosis and pulmonary embolism. *Annals of Family Medicine*, 5(1):63–73, 2007.

[264] V L Shavers, C F Lynch, and L F Burmeister. Knowledge of the Tuskegee study and its impact on the willingness to participate in medical research studies. *J. Natl. Med. Assoc.*, 92(12):563–572, December 2000.

[265] Huwenbo Shi, Kathryn S Burch, Ruth Johnson, Malika K Freund, Gleb Kichaev, Nicholas Mancuso, Astrid M Manuel, Natalie Dong, and Bogdan Pasaniuc. Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. *Am. J. Hum. Genet.*, 106(6):805, June 2020.

[266] Huwenbo Shi, Steven Gazal, Masahiro Kanai, Evan M Koch, Armin P Schoech, Katherine M Siewert, Samuel S Kim, Yang Luo, Tiffany Amariuta, Hailiang Huang, Yukinori Okada, Soumya Raychaudhuri, Shamil R Sunyaev, and Alkes L Price. Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.*, 12(1):1–15, February 2021.

[267] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99(1):139–153, 2016.

[268] Huwenbo Shi, Nicholas Mancuso, Sarah Spendlove, and Bogdan Pasaniuc. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *The American Journal of Human Genetics*, 101(5):737–751, 2017.

[269] Daniel Shriner. Overview of admixture mapping. *Curr. Protoc. Hum. Genet.*, Chapter 1:Unit 1.23, 2013.

[270] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. *CA Cancer J. Clin.*, 71(1):7–33, January 2021. Publisher: Wiley.

[271] Laura Sikstrom, Marta M Maslej, Katrina Hui, Zoe Findlay, Daniel Z Buchman, and Sean L Hill. Conceptualising fairness: three pillars for medical algorithms and health equity. *BMJ Health Care Inform*, 29(1), January 2022.

[272] Susana L Silva, Mariana Fonseca, Marcelo L M Pereira, Sara P Silva, Rita R Barbosa, Ana Serra-Caetano, Elena Blanco, Pedro Rosmaninho, Martin Pérez-Andrés, Ana Berta Sousa, Alexandre A S F Raposo, Margarida Gama-Carvalho, Rui M M Victorino, Lennart Hammarstrom, and Ana E Sousa. Monozygotic Twins Concordant for Common Variable Immunodeficiency: Strikingly Similar Clinical and Immune Profile Associated With a Polygenic Burden. *Front. Immunol.*, 10:2503, November 2019.

[273] Giorgio Sirugo, Sarah A Tishkoff, and Scott M Williams. The quagmire of race, genetic ancestry, and health disparities. *J. Clin. Invest.*, 131(11), June 2021.

[274] Giorgio Sirugo, Scott M Williams, and Sarah A Tishkoff. The Missing Diversity in Human Genetic Studies. *Cell*, 177(1):26–31, March 2019.

[275] Charlotte A Slade, Julian J Bosco, Tran Binh Giang, Elizabeth Kruse, Robert G Stirling, Paul U Cameron, Fiona Hore-Lacy, Michael F Sutherland, Sara L Barnes, Stephen Holdsworth, Samar Ojaimi, Gary A Unglik, Joseph De Luca, Mittal Patel, Jeremy McComish, Kymble Spriggs, Yang Tran, Priscilla Auyeung, Katherine Nicholls, Robyn E O'Hehir, Philip D Hodgkin, Jo A Douglass, Vanessa L Bryant, and Menno C van Zelm. Delayed diagnosis and complications of predominantly antibody deficiencies in a cohort of australian adults, 2018.

[276] José Manuel Soria, Pierre-Emmanuel Morange, Joan Vila, Juan Carlos Souto, Manel Moyano, David-Alexandre Trégouët, José Mateo, Noémi Saut, Eduardo Salas, and Roberto Elosua. Multilocus genetic risk scores for venous thromboembolism risk assessment. *Journal of the American Heart Association*, 3(5):e001060, October 2014.

[277] Elizabeth K Speliotes, Cristen J Willer, Sonja I Berndt, Keri L Monda, Gudmar Thorleifsson, Anne U Jackson, Hana Lango Allen, Cecilia M Lindgren, Jian'an Luan,

Reedik Mägi, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, 42(11):937, 2010.

[278] Asbjørg Stray-Pedersen, Hanne Sørmo Sorte, Pubudu Samarakoon, Tomasz Gambin, Ivan K Chinn, Zeynep H Coban Akdemir, Hans Christian Erichsen, Lisa R Forbes, Shen Gu, Bo Yuan, Shalini N Jhangiani, Donna M Muzny, Olaug Kristin Rødningen, Ying Sheng, Sarah K Nicholas, Lenora M Noroski, Filiz O Seeborg, Carla M Davis, Debra L Canter, Emily M Mace, Timothy J Vece, Carl E Allen, Harshal A Abhyankar, Philip M Boone, Christine R Beck, Wojciech Wiszniewski, Børre Fevang, Pål Aukrust, Geir E Tjønnfjord, Tobias Gedde-Dahl, Henrik Hjorth-Hansen, Ingunn Dybedal, Ingvild Nordøy, Silje F Jørgensen, Tore G Abrahamsen, Torstein Øverland, Anne Grete Bechensteen, Vegard Skogen, Liv T N Osnes, Mari Ann Kulseth, Trine E Prescott, Cecilie F Rustad, Ketil R Heimdal, John W Belmont, Nicholas L Rider, Javier Chinen, Tram N Cao, Eric A Smith, Maria Soledad Caldirola, Liliana Bezrodnik, Saul Oswaldo Lugo Reyes, Francisco J Espinosa Rosales, Nina Denisse Guerrero-Cursaru, Luis Alberto Pedroza, Cecilia M Poli, Jose L Franco, Claudia M Trujillo Vargas, Juan Carlos Aldave Becerra, Nicola Wright, Thomas B Issekutz, Andrew C Issekutz, Jordan Abbott, Jason W Caldwell, Diana K Bayer, Alice Y Chan, Alessandro Aiuti, Caterina Cancrini, Eva Holmberg, Christina West, Magnus Burstedt, Ender Karaca, Gözde Yesil, Hasibe Artac, Yavuz Bayram, Mehmed Musa Atik, Mohammad K Eldomery, Mohammad S Ehlayel, Stephen Jolles, Berit Flatø, Alison A Bertuch, I Celine Hanson, Victor W Zhang, Lee-Jun Wong, Jianhong Hu, Magdalena Walkiewicz, Yaping Yang, Christine M Eng, Eric Boerwinkle, Richard A Gibbs, William T Shearer, Robert Lyle, Jordan S Orange, and James R Lupski. Primary immunodeficiency diseases: Genomic approaches delineate heterogeneous mendelian disorders. *J. Allergy Clin. Immunol.*, 139(1):232–245, January 2017.

[279] Anbezhil Subbarayan, Gloria Colarusso, Stephen M Hughes, Andrew R Gennery, Mary Slatter, Andrew J Cant, and Peter D Arkwright. Clinical features that identify children with primary immunodeficiency diseases, 2011.

[280] Jae Hoon Sul, Lana S Martin, and Eleazar Eskin. Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genet.*, 14(12):e1007309, December 2018.

[281] Maxine Sun, Pierre I Karakiewicz, Jesse D Sammon, Shyam Sukumar, Mai-Kim Gervais, Paul L Nguyen, Toni K Choueiri, Mani Menon, and Quoc-Dien Trinh. Disparities in selective referral for cancer surgeries: implications for the current healthcare delivery system. *BMJ Open*, 4(3):e003921, March 2014.

[282] Binit Sureka, Kalpana Bansal, Yashwant Patidar, S Rajesh, Amar Mukund, and Ankur Arora. Neurologic Manifestations of Chronic Liver Disease and Liver Cirrhosis. *Curr. Probl. Diagn. Radiol.*, 44(5):449–461, September 2015.

[283] C K Svensson. Representation of American blacks in clinical trials of new drugs. *JAMA*, 261(2):263–265, January 1989.

[284] Samina T Syed, Ben S Gerber, and Lisa K Sharp. Traveling towards disease: transportation barriers to health care access. *J. Community Health*, 38(5):976–993, October 2013.

[285] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, Achilleas N Pitsillides, Jonathon LeFaive, Seung-Been Lee, Xiaowen Tian, Brian L Browning, Sayantan Das, Anne-Katrin Emde, Wayne E Clarke, Douglas P Loesch, Amol C Shetty, Thomas W Blackwell, Albert V Smith, Quenna Wong, Xiaoming Liu, Matthew P Conomos, Dean M Bobo, François Aguet, Christine Albert, Alvaro Alonso, Kristin G Ardlie, Dan E Arking, Stella Aslibekyan, Paul L Auer, John Barnard, R Graham Barr, Lucas Barwick, Lewis C Becker, Rebecca L Beer, Emelia J Benjamin, Lawrence F Bielak, John Blangero, Michael Boehnke, Donald W Bowden, Jennifer A Brody, Esteban G Burchard, Brian E Cade, James F Casella, Brandon Chalazan, Daniel I Chasman, Yii-Der Ida Chen, Michael H Cho, Seung Hoan Choi, Mina K Chung, Clary B Clish, Adolfo Correa, Joanne E Curran, Brian Custer, Dawood Darbar, Michelle Daya, Mariza de Andrade, Dawn L DeMeo, Susan K Dutcher, Patrick T Ellinor, Leslie S Emery, Celeste Eng, Diane Fatkin, Tasha Fingerlin, Lukas Forer, Myriam Fornage, Nora Franceschini, Christian Fuchsberger, Stephanie M Fullerton, Soren Germer, Mark T Gladwin, Daniel J Gottlieb, Xiuqing Guo, Michael E Hall, Jiang He, Nancy L Heard-Costa, Susan R Heckbert, Marguerite R Irvin, Jill M Johnsen, Andrew D Johnson, Robert Kaplan, Sharon L R Kardia, Tanika Kelly, Shannon Kelly, Eimear E Kenny, Douglas P Kiel, Robert Klemmer, Barbara A Konkle, Charles Kooperberg, Anna Köttgen, Leslie A Lange, Jessica Lasky-Su, Daniel Levy, Xihong Lin, Keng-Han Lin, Chunyu Liu, Ruth J F Loos, Lori Garman, Robert Gerszten, Steven A Lubitz, Kathryn L Lunetta, Angel C Y Mak, Ani Manichaikul, Alisa K Manning, Rasika A Mathias, David D McManus, Stephen T McGarvey, James B Meigs, Deborah A Meyers, Julie L Mikulla, Mollie A Minear, Braxton D Mitchell, Sanghamitra Mohanty, May E Montasser, Courtney Montgomery, Alanna C Morrison, Joanne M Murabito, Andrea Natale, Pradeep Natarajan, Sarah C Nelson, Kari E North, Jeffrey R O'Connell, Nicholette D Palmer, Nathan Pankratz, Gina M Peloso, Patricia A Peyser, Jacob Pleiness, Wendy S Post, Bruce M Psaty, D C Rao, Susan Redline, Alexander P Reiner, Dan Roden, Jerome I Rotter, Ingo Ruczinski, Chloé Sarnowski, Sebastian Schoenherr, David A Schwartz, Jeong-Sun Seo, Sudha Seshadri, Vivien A Sheehan, Wayne H Sheu, M Benjamin Shoemaker, Nicholas L Smith, Jennifer A Smith, Nona Sotoodehnia, Adrienne M Stilp, Weihong Tang, Kent D Taylor, Marilyn Telen, Timothy A Thornton, Russell P Tracy, David J Van Den Berg, Ramachandran S Vasan, Karine A Viaud-Martinez, Scott Vrieze, Daniel E Weeks, Bruce S Weir, Scott T Weiss, Lu-Chen Weng, Cristen J Willer, Yingze Zhang, Xutong Zhao, Donna K Arnett, Allison E Ashley-Koch, Kathleen C Barnes, Eric Boerwinkle, Stacey Gabriel, Richard Gibbs, Kenneth M Rice, Stephen S Rich, Edwin K Silverman, Pankaj Qasba, Weiniu Gan, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, George J Papanicolaou, Deborah A Nickerson, Sharon R Browning, Michael C Zody, Sebastian Zöllner, James G Wilson, L Adrienne Cupples, Cathy C Laurie, Cashell E Jaquish,

Ryan D Hernandez, Timothy D O'Connor, and Gonçalo R Abecasis. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845):290–299, February 2021.

[286] Stuart G Tangye, Waleed Al-Herz, Aziz Bousfiha, Talal Chatila, Charlotte Cunningham-Rundles, Amos Etzioni, Jose Luis Franco, Steven M Holland, Christoph Klein, Tomohiro Morio, Hans D Ochs, Eric Oksenhendler, Capucine Picard, Jennifer Puck, Troy R Torgerson, Jean-Laurent Casanova, and Kathleen E Sullivan. Human inborn errors of immunity: 2019 update on the classification from the international union of immunological societies expert committee. *J. Clin. Immunol.*, 40(1):24–64, January 2020.

[287] Hale M Thompson, Brihat Sharma, Sameer Bhalla, Randy Boley, Connor McCluskey, Dmitriy Dligach, Matthew M Churpek, Niranjan S Karnik, and Majid Afshar. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. *J. Am. Med. Inform. Assoc.*, 28(11):2393–2403, October 2021.

[288] Ruth Thorlby, Selena Jorgensen, Bruce Siegel, and John Z Ayanian. How health care organizations are using data on patients' race and ethnicity to improve quality of care. *Milbank Q.*, 89(2):226, June 2011.

[289] Peixin Tian, Tsai Hor Chan, Yong-Fei Wang, Wanling Yang, Guosheng Yin, and Yan Dora Zhang. Multiethnic polygenic risk prediction in diverse populations through transfer learning. *Front. Genet.*, 13:906965, August 2022.

[290] Nicholas J Timpson, Celia M T Greenwood, Nicole Soranzo, Daniel J Lawson, and J Brent Richards. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.*, 19(2):110–124, February 2018.

[291] Sarah A Tishkoff and Kenneth K Kidd. Implications of biogeography of human populations for 'race' and medicine. *Nat. Genet.*, 36(11):S21–S27, October 2004. Publisher: Nature Publishing Group.

[292] Sarah A Tishkoff and Brian C Verrelli. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.*, 4:293–340, 2003.

[293] Sergio Tofanelli, Luca Taglioli, Stefania Bertoncini, Paolo Francalacci, Anatole Klyosov, and Luca Pagani. Mitochondrial and Y chromosome haplotype motifs as diagnostic markers of jewish ancestry: a reconsideration. *Front. Genet.*, 5, November 2014.

[294] K Toms, E Gkrania-Klotsas, and D Kumararatne. Analysis of scoring systems for primary immunodeficiency diagnosis in adult immunology clinics. *Clin. Exp. Immunol.*, 203(1):47–54, January 2021.

[295] Johanne Tremblay, Mounsif Haloui, Redha Attaoua, Ramzan Tahir, Camil Hishmih, François Harvey, François-Christophe Marois-Blanchet, Carole Long, Paul Simon, Lara Santucci, Candan Hizel, John Chalmers, Michel Marre, Stephen Harrap, Renata Cífková, Alena Krajčoviechová, David R Matthews, Bryan Williams, Neil Poulter, Sophia Zoungas, Stephen Colagiuri, Giuseppe Mancia, Diederick E Grobbee, Anthony Rodgers, Liusheng Liu, Mawussé Agbessi, Vanessa Bruat, Marie-Julie Favé, Michelle P Harwood, Philip Awadalla, Mark Woodward, Julie G Hussin, and Pavel Hamet. Polygenic risk scores predict diabetes complications and their response to intensive blood pressure and glucose control. *Diabetologia*, 64(9):2012–2025, September 2021.

[296] Eric Trépo, Stefano Romeo, Jessica Zucman-Rossi, and Pierre Nahon. PNPLA3 gene in liver diseases. *J. Hepatol.*, 65(2):399–412, August 2016.

[297] Eric Turkheimer, Kathryn Paige Harden, and Richard E Nisbett. There's still no good reason to believe black-white IQ differences are due to genes. https://www.vox.com/the-big-idea/2017/6/15/15797120/race-black-white-iq-response-critics, June 2017. Accessed: 2023-3-6.

[298] United States Census Bureau. QuickFacts: Santa monica city, california.

[299] United States Census Bureau. QuickFacts: Los Angeles city, California, 2020.

[300] Simon Urschel, Lale Kayikci, Uwe Wintergerst, Gundula Notheis, Annette Jansson, and Bernd H Belohradsky. Common variable immunodeficiency disorders in children: delayed diagnosis despite typical clinical presentation. *J. Pediatr.*, 154(6):888–894, June 2009.

[301] US Census Bureau. About the topic of race. https://www.census.gov/topics/population/race/about.html. Accessed: 2022-3-9.

[302] Sara L Van Driest, Noura S Abul-Husn, Joseph T Glessner, Lisa Bastarache, Sharon Nirenberg, Jonathan S Schildcrout, Meghana S Eswarappa, Gillian M Belbin, Christian M Shaffer, Frank Mentch, John Connolly, Mingjian Shi, C Michael Stein, Dan M Roden, Hakon Hakonarson, Nancy J Cox, Scott C Borinstein, and Jonathan D Mosley. Association Between a Common, Benign Genotype and Unnecessary Bone Marrow Biopsies Among African American Patients. *JAMA Intern. Med.*, 181(8):1100–1105, August 2021.

[303] Candelaria Vergara, Margaret M Parker, Liliana Franco, Michael H Cho, Ana V Valencia-Duarte, Terri H Beaty, and Priya Duggal. Genotype imputation performance of three reference panels using african ancestry individuals. *Hum. Genet.*, 137(4):281–292, April 2018.

[304] Anurag Verma, Noah L Tsao, Lauren O Thomann, Yuk-Lam Ho, Sudha K Iyengar, Shiuh-Wen Luoh, Rotonya Carr, Dana C Crawford, Jimmy T Efird, Jennifer E Huffman, Adriana Hung, Kerry L Ivey, Michael G Levin, Julie Lynch, Pradeep Natarajan, Saiju Pyarajan, Alexander G Bick, Lauren Costa, Giulio Genovese, Richard Hauger,

Ravi Madduri, Gita A Pathak, Renato Polimanti, Benjamin Voight, Marijana Vujkovic, Seyedeh Maryam Zekavat, Hongyu Zhao, Marylyn D Ritchie, VA Million Veteran Program COVID-19 Science Initiative, Kyong-Mi Chang, Kelly Cho, Juan P Casas, Philip S Tsao, J Michael Gaziano, Christopher O'Donnell, Scott M Damrauer, and Katherine P Liao. A Phenome-Wide association study of genes associated with COVID-19 severity reveals shared genetics with complex diseases in the million veteran program. *PLoS Genet.*, 18(4):e1010113, April 2022.

[305] Joshua R Vest and Ofir Ben-Assuli. Prediction of emergency department revisits using area-level social determinants of health measures and health information exchange information. *Int. J. Med. Inform.*, 129:205–210, September 2019.

[306] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms, 2020.

[307] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *N. Engl. J. Med.*, 383(9):874–882, August 2020. Publisher: Massachusetts Medical Society.

[308] Lynne E Wagenknecht, Nicholette D Palmer, Donald W Bowden, Jerome I Rotter, Jill M Norris, Julie Ziegler, Yii-Der I Chen, Steven Haffner, Ann Scherzinger, and Carl D Langefeld. Association of PNPLA3 with non-alcoholic fatty liver disease in a minority cohort: the Insulin Resistance Atherosclerosis Family Study. *Liver Int.*, 31(3):412–416, March 2011.

[309] Jennifer K. Wagner, Joon-Ho Yu, Jayne O. Ifekwunigwe, Tanya M. Harrell, Michael J. Bamshad, and Charmaine D. Royal. Anthropologists' views on race, ancestry, and genetics. *American Journal of Physical Anthropology*, 162(2):318–327, February 2017.

[310] Niels Weifenbach, Annalena A C Schneckenburger, and Stefan Lötters. Global Distribution of Common Variable Immunodeficiency (CVID) in the Light of the UNDP Human Development Index (HDI): A Preliminary Perspective of a Rare Disease. *J Immunol Res*, 2020:8416124, September 2020.

[311] Karin Weissenborn, Martin Bokemeyer, Jochen Krause, Jochen Ennen, and Björn Ahl. Neurological and neuropsychiatric syndromes associated with liver disease. *AIDS*, 19 Suppl 3:S93–8, October 2005.

[312] Ken Wiley, Jr and Robb Rowley. Polygenic RIsk MEthods in diverse populations (PRIMED) consortium. https://www.genome.gov/Funded-Programs-Projects/PRIMED-Consortium. Accessed: 2023-2-26.

[313] Melissa Wills. Are clusters races? a discussion of the rhetorical appropriation of rosenberg et al.'s 'genetic structure of human populations'. *Philosophy & Theory in Biology*, 9, 2017.

[314] Thomas W Winkler, Felix R Day, Damien C Croteau-Chonka, Andrew R Wood, Adam E Locke, Reedik Mägi, Teresa Ferreira, Tove Fall, Mariaelisa Graff, Anne E Justice, Jian'an Luan, Stefan Gustafsson, Joshua C Randall, Sailaja Vedantam, Tsegaselassie Workalemahu, Tuomas O Kilpeläinen, André Scherag, Tonu Esko, Zoltán Kutalik, Iris M Heid, Ruth J F Loos, and Genetic Investigation of Anthropometric Traits (GIANT) Consortium. Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.*, 9(5):1192–1212, May 2014.

[315] Lauren A Wise, Edward A Ruiz-Narvaez, Julie R Palmer, Yvette C Cozier, Arti Tandon, Nick Patterson, Rose G Radin, Lynn Rosenberg, and David Reich. African ancestry and genetic risk for uterine leiomyomata. *Am. J. Epidemiol.*, 176(12):1159–1168, December 2012.

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

[317] Brooke N Wolford, Cristen J Willer, and Ida Surakka. Electronic health records: the next wave of complex disease genetics. *Hum. Mol. Genet.*, 27(R1):R14–R21, May 2018.

[318] Patrick Wu, Aliya Gifford, Xiangrui Meng, Xue Li, Harry Campbell, Tim Varley, Juan Zhao, Robert Carroll, Lisa Bastarache, Joshua C Denny, Evropi Theodoratou, and Wei-Qi Wei. Mapping ICD-10 and ICD-10-CM codes to phecodes: Workflow development and initial evaluation. *JMIR Med Inform*, 7(4):e14325, November 2019.

[319] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[320] Jian Yang, Teri A Manolio, Louis R Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M Cunningham, Mariza De Andrade, Bjarke Feenstra, Eleanor Feingold, M Geoffrey Hayes, et al. Genome partitioning of genetic variation for complex traits using common snps. *Nature genetics*, 43(6):519, 2011.

[321] Ruqaiijah Yearby, Brietta Clark, and José F Figueroa. Structural racism in historical and modern US health care policy. *Health Aff.*, 41(2):187–194, February 2022.

[322] Jian Zeng, Ronald Vlaming, Yang Wu, Matthew R Robinson, Luke R Lloyd-Jones, Loic Yengo, Chloe X Yap, Angli Xue, Julia Sidorenko, Allan F McRae, et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nature genetics*, 50(5):746, 2018.

[323] Yan Zhang, Guanghao Qi, Ju-Hyun Park, and Nilanjan Chatterjee. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature genetics*, 50(9):1318, 2018.

[324] Zhangchen Zhao, Lars G Fritsche, Jennifer A Smith, Bhramar Mukherjee, and Seunggeun Lee. The construction of cross-population polygenic risk scores using transfer learning. *Am. J. Hum. Genet.*, 109(11):1998–2008, November 2022.

[325] Hou-Feng Zheng, Martin Ladouceur, Celia M. T. Greenwood, and J. Brent Richards. Effect of genome-wide genotyping and reference panels on rare variants imputation. *Journal of Genetics and Genomics = Yi Chuan Xue Bao*, 39(10):545–550, October 2012.

[326] Wei Zhou, Masahiro Kanai, Kuan-Han H. Wu, Humaira Rasheed, Kristin Tsuo, Jibril B. Hirbo, Ying Wang, Arjun Bhattacharya, Huiling Zhao, Shinichi Namba, Ida Surakka, Brooke N. Wolford, Valeria Lo Faro, Esteban A. Lopera-Maya, Kristi Läll, Marie-Julie Favé, Juulia J. Partanen, Sinéad B. Chapman, Juha Karjalainen, Mitja Kurki, Mutaamba Maasha, Ben M. Brumpton, Sameer Chavan, Tzu-Ting Chen, Michelle Daya, Yi Ding, Yen-Chen A. Feng, Lindsay A. Guare, Christopher R. Gignoux, Sarah E. Graham, Whitney E. Hornsby, Nathan Ingold, Said I. Ismail, Ruth Johnson, Triin Laisk, Kuang Lin, Jun Lv, Iona Y. Millwood, Sonia Moreno-Grau, Kisung Nam, Priit Palta, Anita Pandit, Michael H. Preuss, Chadi Saad, Shefali Setia-Verma, Unnur Thorsteinsdottir, Jasmina Uzunovic, Anurag Verma, Matthew Zawistowski, Xue Zhong, Nahla Afifi, Kawthar M. Al-Dabhani, Asma Al Thani, Yuki Bradford, Archie Campbell, Kristy Crooks, Geertruida H. de Bock, Scott M. Damrauer, Nicholas J. Douville, Sarah Finer, Lars G. Fritsche, Eleni Fthenou, Gilberto Gonzalez-Arroyo, Christopher J. Griffiths, Yu Guo, Karen A. Hunt, Alexander Ioannidis, Nomdo M. Jansonius, Takahiro Konuma, Ming Ta Michael Lee, Arturo Lopez-Pineda, Yuta Matsuda, Riccardo E. Marioni, Babak Moatamed, Marco A. Nava-Aguilar, Kensuke Numakura, Snehal Patil, Nicholas Rafaels, Anne Richmond, Agustin Rojas-Muñoz, Jonathan A. Shortt, Peter Straub, Ran Tao, Brett Vanderwerff, Manvi Vernekar, Yogasudha Veturi, Kathleen C. Barnes, Marike Boezen, Zhengming Chen, Chia-Yen Chen, Judy Cho, George Davey Smith, Hilary K. Finucane, Lude Franke,

Eric R. Gamazon, Andrea Ganna, Tom R. Gaunt, Tian Ge, Hailiang Huang, Jennifer Huffman, Nicholas Katsanis, Jukka T. Koskela, Clara Lajonchere, Matthew H. Law, Liming Li, Cecilia M. Lindgren, Ruth J.F. Loos, Stuart MacGregor, Koichi Matsuda, Catherine M. Olsen, David J. Porteous, Jordan A. Shavit, Harold Snieder, Tomohiro Takano, Richard C. Trembath, Judith M. Vonk, David C. Whiteman, Stephen J. Wicks, Cisca Wijmenga, John Wright, Jie Zheng, Xiang Zhou, Philip Awadalla, Michael Boehnke, Carlos D. Bustamante, Nancy J. Cox, Segun Fatumo, Daniel H. Geschwind, Caroline Hayward, Kristian Hveem, Eimear E. Kenny, Seunggeun Lee, Yen-Feng Lin, Hamdi Mbarek, Reedik Mägi, Hilary C. Martin, Sarah E. Medland, Yukinori Okada, Aarno V. Palotie, Bogdan Pasaniuc, Daniel J. Rader, Marylyn D. Ritchie, Serena Sanna, Jordan W. Smoller, Kari Stefansson, David A. van Heel, Robin G. Walters, Sebastian Zöllner, Alicia R. Martin, Cristen J. Willer, Mark J. Daly, and Benjamin M. Neale. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics*, 2(10):100192, October 2022.

[327] Wei Zhou, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A. Gagliano, Aliya Gifford, Lisa A. Bastarache, Wei-Qi Wei, Joshua C. Denny, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R. Abecasis, Cristen J. Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341, September 2018.

[328] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264, 2013.

[329] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, 44(7):821–824, June 2012.

[330] Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The annals of applied statistics*, 11(3):1561, 2017.