Target Journal

American Journal of Human Genetics

Other ideas: *PLoS Genetics*, *Genetic Epidemiology*

Title

Adjusting for principal components can induce spurious associations in genome-wide association studies in admixed populations

Authors and Affiliations

Kelsey E. Grinde,[1]* Brian L. Browning,[2] Alexander P. Reiner,[3,4] Timothy A. Thornton,[5,6] Sharon R. Browning[6] (middle authors are listed alphabetically by last name)

1. Department of Mathematics, Statistics, and Computer Science, Macalester College, Saint Paul, MN, 55105, USA
2. Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, 98195, USA
3. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, USA
4. Department of Epidemiology, University of Washington, Seattle, WA, 98195, USA
5. Regeneron Genetics Center, Tarrytown, New York, 10591, USA
6. Department of Biostatistics, University of Washington, Seattle, WA, 98195, USA

* kgrinde@macalester.edu

Please confirm that your affiliations are correct!

Abstract

Current word count: 244

Maximum allowed by AJHG: 250

Principal component analysis (PCA) is widely used to control for population structure in genome-wide association studies (GWAS). The top principal components (PCs) typically reflect population structure, but deciding exactly how many PCs to include in regression models can be challenging. Often researchers err on the side of including more PCs than may be necessary to ensure that population structure is fully captured. In this paper, we show that adjusting for extraneous PCs can induce spurious associations. Spurious associations can arise when PCs are correlated with multiple local genomic features, such as regions of the genome with atypical linkage disequilibrium (LD), rather than genome-wide ancestry. We investigate the performance of PCA in samples of African American individuals from the Women's Health Initiative SNP Health Association Resource and two Trans-Omics for Precision Medicine Whole Genome Sequencing Project studies (the Jackson Heart Study and the Genetic Epidemiology of Chronic Obstructive Pulmonary Disease Study). In all three samples, problems arise unless careful pre-processing is performed prior to running PCA. We show that LD pruning, using stricter thresholds than is often suggested in the literature, can resolve these issues, whereas excluding high LD regions identified in previous studies does not. We also demonstrate that the rate of spurious associations is controlled when we simply adjust for the first PC or estimated genome-wide ancestry proportions. Our work demonstrates that great care must be taken when using principal components to control for population structure in genome-wide association studies in admixed populations.

1 Introduction

There is considerable variability in global ancestry—the genome-wide proportion of genetic material inherited from each ancestral population—across individuals in admixed populations[1,2,3,4,5]. Heterogeneous global ancestry, as with other types of population structure, can lead to spurious associations in genome-wide association studies[6,7,8,9]. In fact, some authors have cited the ancestral heterogeneity of admixed populations, and the statistical challenges it poses, as one reason why these populations have been historically underrepresented in genome-wide association studies (GWAS)[10,11,12,13,14]. Spurious associations can arise in GWAS in ancestrally heterogeneous populations when global ancestry confounds the association between genotypes and the phenotype of interest. This confounding occurs when the genetic variant being tested differs in frequency across ancestral populations (i.e., global ancestry is associated with genotype) and global ancestry also has an effect on the phenotype via, for example, environmental factors or causal loci elsewhere in the genome that differ in frequency across ancestral groups.

A number of methods for detecting and controlling for ancestral heterogeneity in genetic association studies have been proposed. Early approaches included restricting analyses to subsets of ancestrally homogeneous individuals[15], performing a genome-wide correction for test statistic inflation due to ancestral heterogeneity via genomic control[6], and using family-based designs[16]. More recently, approaches based on mixed models have been proposed[17,18,19], using random effects to control for both close (e.g., due to family-based sampling) and distant (e.g., due to shared ancestry) relatedness across individuals. When studies do not include closely related individuals, a simpler approach is to include inferred global ancestry as a fixed effect in marginal regression models[7,20]. This fixed effects adjustment for global ancestry is used extensively in published studies (e.g.,[21,22,23,24,25,26,27,28]), with global ancestry inferred using either model-based ancestry inference methods or principal component analysis.

Model-based approaches for global ancestry inference (e.g., `frappe`[29], `STRUCTURE`[30], `ADMIXTURE`[31],

RFMix[32]) model the probability of observed genotypes given unobserved ancestry and allele frequencies in each ancestral population. These approaches are used to estimate global ancestry proportions, also known as admixture proportions, which can then be included as covariates in GWAS models to adjust for ancestral heterogeneity. One of the challenges of using these model-based approaches to infer global ancestry is that the total number of ancestral populations usually needs to be pre-specified. In addition, many of these approaches are supervised, requiring reference panel data from each ancestral population of interest to estimate allele frequencies. Furthermore, ancestry inference is typically conducted at a continental level (e.g., African versus European) so finer levels of population structure could be missed, although some recent efforts have considered global ancestry inference on a sub-continental scale[33,34].

Principal component analysis (PCA) is an unsupervised approach for inferring global ancestry that is implemented by a variety of software packages (e.g., EIGENSTRAT[7], SNPRelate[35], PC-AiR[36]). PCA offers the advantages of not requiring reference panel data or pre-specification of the number of ancestral populations of interest, and it is capable of capturing sub-continental structure[23]. The top principal components (PCs) tend to reflect global ancestry[37,38]. To adjust for ancestral heterogeneity in genome-wide association studies, researchers must choose some number of PCs to include as covariates in GWAS regression models. Determining this number of PCs needed to capture global ancestry is non-trivial. Numerous techniques have been proposed, including formal significance tests based on Tracy-Widom theory[7,37], examining inflation factors[5,39] or the proportion of variance explained by each PC[5,39,40], comparing PCs to self-reported race/ethnicity[5], and keeping PCs that are significantly associated with the trait[24,41]. Typically, the number of PCs selected is on the order of one to ten[42], but in practice it is common to see applications in which many more PCs are used—more than may be necessary to capture global ancestry. Prior work has suggested that including higher-order PCs can provide the safeguard of removing "virtually all stratification"[43] at the cost of perhaps only "subtle" decreases in power[44].

Another challenge involves ensuring that PCs reflect global ancestry rather than other features or artifacts. Principal components can capture relatedness across samples[9,36,37,45], array artifacts or other data quality issues[7,9,37,46], and small regions of the genome with unusual patterns of linkage disequilibrium (LD)[7,9,21,37,45,46,47,48,49,50,51]. To address this last issue, many authors have suggested running PCA on a reduced subset of variants after first performing *LD pruning*, using a program such as `PLINK`[52] to remove variants that are in "high" LD (e.g., pairwise-correlation $r^2 > 0.2$) with nearby variants[5,21,22,23,36,39,41,45,46,50,53,54,55,56,57], and/or excluding regions of the genome that are known to have extensive, long-ranging, or otherwise unusual patterns of LD[5,21,22,23,40,46,48,55]. A list of these previously-identified high LD regions and references that recommend their exclusion are provided in Table 1.

LD pruning and filtering are not universally practiced, and the downstream implications of adjusting for PCs that capture features other than global ancestry are not fully understood. Furthermore, much of this prior work was conducted in populations of European ancestry, so recommendations on how best to implement principal component-based adjustment for ancestral heterogeneity in admixed populations are lacking. In this paper, we investigate the impact of LD filtering and pruning choices, as well as choices of the number of principal components to include in analyses, on genome-wide association studies in admixed populations. We show that principal components may capture local genomic features, unless careful pre-processing is performed prior to analysis. We also conduct simulation studies and provide analytic results to show that including too many PCs can induce spurious associations in GWAS, particularly when those extraneous PCs capture local genomic features rather than genome-wide ancestry. To conclude, we provide suggestions regarding best practices for controlling for ancestral heterogeneity in genome-wide association studies in admixed populations.

Chr	Start (bp)	End (bp)	References
1	48000000	52060567	46,48,55
2	85941853	100500000	46,48,55
2	129600000	140000000	5,23,40,46,48,58
2	182882739	190000000	46,48,55
3	47500000	50000000	46,48,55
3	83500000	87000000	46,48,55
3	89000000	97500000	46,48
3	163100000	164900000	58
5	44000000	51500000	22,46,48,55
5	98000000	100500000	46,48
5	129000000	132000000	46,48,55
5	135500000	138500000	46,48
6	23800000	39000000	5,22,23,40,46,48,55,58
6	57000000	64000000	46,48,55
6	140000000	142500000	46,48,55
7	55000000	66193285	46,48,55
8	6300000	13500000	5,22,23,40,46,47,48,55,58
8	43000000	50000000	46,48,55
8	112000000	115000000	46,48,55
10	37000000	43000000	46,48,55
11	45000000	57000000	22,46,48
11	87500000	90500000	46,48,55
12	33000000	40000000	46,48,55
12	109500000	112021663	46,48
14	46600000	47500000	58
17	37800000	42000000	5,23
20	32000000	34500000	46,48,55

Table 1: Regions of the genome with high, long-range, or otherwise unusual linkage disequilibrium that are often recommended for exclusion prior to running PCA. Start and end physical (base pair) positions are provided with respect to genome build 36. This list is also available for download in builds 36, 37, and 38 at https://github.com/kegrinde/PCA/.

2 Material and Methods

2.1 Data and Quality Control

Our analyses focus on genotype and sequence data from three samples of unrelated African American individuals. In particular, we consider genotype data from the Women's Health Initiative SNP Health Association Resource (WHI SHARe), as well as whole genome sequencing data from two contributing studies to the Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Project: the Jackson Heart Study (JHS) and the Genetic Epidemiology of Chronic Obstructive Pulmonary Disease Study (COPDGene). We performed quality control and identified subsets of unrelated African American individuals prior to running any further analyses.

2.1.1 WHI SHARe Genotype Data

The Women's Health Initiative (WHI) is a long-term study of the health of women in the United States. In total, 161,808 postmenopausal women aged 50–79 years old were recruited to participate in this study. Additional details of the study design and cohort characteristics can be found elsewhere[59]. The WHI study includes 12,151 self-identified African American women who consented to genetic research, a subsample of which were selected for genotyping using the Affymetrix Genome-Wide Human SNP Array 6.0. This array contains 906,000 single nucleotide polymorphisms (SNPs) and more than 946,000 probes for detection of copy number variants. In our analyses, we focus only on the SNP data.

The genotype data were processed for quality control. After filtering on call rate, concordance rates for blinded and unblinded duplicates, and sex discrepancy, there were 871,309 SNPs with a missing genotyping rate of 0.2% and 8,421 African American women[24]. We used the iterative procedure suggested by Conomos et al.[60] to identify a subset of 8,064 mutually unrelated individuals, using a kinship threshold of 0.044 (i.e., excluding first-, second-, and third-degree relatives).

2.1.2 TOPMed Whole Genome Sequence Data

The TOPMed Whole Genome Sequencing Project is an ongoing project sponsored by the National Heart, Lung, and Blood Institute. The goal of this project is to collect and analyze whole-genome sequences, other omics data, and extensive phenotypic information for over 100,000 individuals from diverse backgrounds. Data are periodically released on dbGaP for analysis by the broader scientific community. Our analysis uses data from freeze 4, released in 2017, and freeze 5b, released in 2018. These two freezes include samples from a large number of contributing studies. We focus on two of these studies: the Jackson Heart Study (JHS) (accession number: phs000964) and the Genetic Epidemiology of Chronic Obstructive Pulmonary Disease Study (COPDGene) (accession number: phs000951). For TOPMed freezes 4 and 5b, high coverage ($\approx$ 30X) whole genome sequencing was performed by several sequencing centers. Details on TOPMed sequencing and QC methods are available in Taliun et al.[62] and on the TOPMed website: https://topmed.nhlbi.nih.gov/data-sets. In total, the downloaded freeze 4 JHS dataset includes 2,777 African American individuals and the freeze 5b COPDGene dataset includes 8,476 African American and European American individuals consented for biomedical research.

Prior to genetic ancestry inference, we performed two additional stages of variant- and sample-level filtering. We used `bcftools`[63] to restrict our analyses to biallelic single nucleotide variants (SNVs). To identify a subset of mutually unrelated individuals (kinship threshold = 0.044), we used the University of Washington Genetic Analysis Center (UW GAC) TOPMed analysis pipeline. In addition to inferring relatedness using the procedure proposed by Conomos et al.[60], this pipeline also includes code to perform PCA, association testing, and other tasks in whole genome sequence data: more details can be found at https://github.com/UW-GAC/analysis_pipeline. After filtering, 1,928 and 8,406 unrelated samples and 77,136,850 and 135,522,041 variants remained in JHS and COPDGene, respectively.

2.2 Genetic Ancestry Inference

We consider two approaches to inferring genetic ancestry in these admixed samples: model-based approaches and principal component analysis.

2.2.1 Model-Based Approaches

In WHI SHARe African Americans, we inferred both local and global genetic ancestry using model-based ancestry inference techniques. Local ancestry inference was performed using `RFMix`[32] and a reference panel including individuals of European and African descent from the International HapMap Project (HapMap)[64]: see Grinde et al.[65] for more details. We then calculated global ancestry proportions via the genome-wide average local ancestry $\hat{\pi}_{ik} = \frac{1}{2m}\sum_{j=1}^{m} a_{ijk}$, where a_{ijk} is the inferred number of alleles (0, 1, or 2) inherited by individual i at variant j from ancestral population k. We also compared these `RFMix`-based global ancestry estimates to results from supervised and unsupervised `ADMIXTURE`[31] analyses with two ancestral populations ($K = 2$). The supervised analysis used the same HapMap reference panel as was used to infer local ancestry using `RFMix`. All three sets of admixture proportions were highly correlated (pairwise Pearson correlation > 0.998). We proceed with using only the `RFMix`-based admixture proportion estimates for the remainder of our analyses.

In TOPMed JHS and COPDGene samples, we inferred global ancestry via unsupervised `ADMIXTURE` analyses with both two and three ancestral populations (i.e., $K = 2$ and $K = 3$). We also used these inferred global ancestry proportions to identify subsets of admixed individuals. Although JHS is known to focus on African Americans, we identified 40 individuals in the sample with an estimated European ancestry proportion of 100%. These individuals were excluded from further analyses, leaving a total of 1,888 unrelated admixed samples. The COPDGene study, by design, includes both African American and European American individuals. Self-identified race/ethnicity information was not available from dbGaP, so we used inferred admixture proportions to identify and restrict our attention to individuals with at least 29.5% African ancestry. The choice of threshold follows from the results re-

ported by Parker et al.[66], showing that the self-identified African American individuals in the COPDGene study have inferred proportions of African ancestry ranging from 29.5% and above. (We are not suggesting that this same threshold be universally applied to identify African American individuals in other samples.) After filtering, 2,676 individuals remain in COPDGene.

2.2.2 Principal Component Analysis

To infer global ancestry using PCA, we perform a singular value decomposition of the matrix of standardized genotypes (i.e., $\mathbf{X} = \mathbf{UDV}^T$) or, equivalently, an eigenvalue decomposition of the genetic relationship matrix (i.e., $\mathbf{XX}^T = \mathbf{UD}^2\mathbf{U}^T$), where $\mathbf{X}$ is the $n \times m$ matrix of standardized genotypes for n individuals at m single nucleotide variants. One or more of the top eigenvectors, or principal components, $\mathbf{u}_1, \mathbf{u}_2, \ldots$, typically reflect global ancestry.

We ran PCA on the WHI SHARe genotype data using `SNPRelate`[35]. First, we applied PCA to the set of all 551,025 SNPs with available genotypes. We then applied PCA to subsets of SNPs based on the following pre-processing criteria: excluding SNPs falling into regions of the genome that have been cited in the literature as potentially problematic for PCA (Table 1), LD pruning, or both literature-based exclusions and LD pruning. To perform LD pruning, two parameters must be specified: r^2 threshold and window size. Here we use an r^2 threshold of 0.1 and window size of 0.5 mega basepairs (Mb), which is stricter than is often suggested in the literature: for a full discussion of these choices, see Supplemental Information Section S2. Both LD pruning and filtering of regions in Table 1 were implemented using the `SNPRelate` package. Table 2 summarizes the number of SNPs remaining after each set of pre-processing steps. After running PCA on these different sets of variants, we also used `SNPRelate` to assess the contribution of each SNP to each principal component by calculating the SNP loadings and the correlation between PCs and genotypes.

In TOPMed JHS and COPDGene samples, we used the UW GAC TOPMed analysis pipeline to implement pre-processing, run principal component analysis, and calculate and

	WHI SHARe	TOPMed JHS	TOPMed COPDGene
Neither Exclude nor Prune	551,025	14,117,957	13,959,378
Exclude High LD Regions (Table 1)	536,668	13,723,944	13,575,038
LD Prune ($r^2 < 0.1$, 0.5 Mb windows)	49,723	245,003	251,040
Both Exclude and Prune	48,794	239,425	245,378

Table 2: Number of autosomal SNPs remaining after different combinations of pre-processing steps were applied prior to running PCA. Note that TOPMed analyses include only those variants with minor allele frequency greater than 1%.

visualize the contribution of individual variants to each PC. Similar to WHI SHARe, we applied PCA to various subsets of variants based on different pre-processing criteria, including a naive analysis with no prior LD-based pruning or filtering, an analysis after excluding the regions listed in Table 1, an analysis after LD pruning with an r^2 threshold of 0.1, and an analysis after both LD-based pruning and filtering. Following the recommendations of Kirk[67] and the UW GAC pipeline documentation, we also removed variants with a minor allele frequency lower than 0.01 in all cases. Note that the UW GAC pipeline provides the option to exclude some of the regions listed in Table 1 (the *LCT* gene on chromosome 2, the HLA region on chromosome 6, and the locations of large inversions on chromosomes 8 and 17), but we customized the pipeline code to add the other regions identified in our literature review. See Table 2 for the number of variants included in each analysis.

2.3 Simulation Study

To explore the impact of adjusting for principal components that capture local genomic features, we conducted a simulation study using genotype data and simulated traits in the WHI SHARe African American sample.

2.3.1 Trait Simulation

Traits were simulated for each individual $i = 1, \ldots, 8064$ such that they depended only on the genotype g_{ij} at a single causal variant with effect size β_j:

$$y_i = \beta_j g_{ij} + \epsilon_i, \ \epsilon_i \overset{iid}{\sim} N(0, 1).$$

We considered seven choices of effect sizes ($\beta_j = 0, 0.25, 0.5, 1, 2, 4, 8$) and nearly 500 choices for the position j of the causal variant, varying the position of this causal variant across all 22 autosomes.

To choose the location of these causal variants, we first estimated the difference in ancestral allele frequencies for each variant using the observed allele frequencies in our HapMap reference panel which included samples from the CEU (Utah residents with Northern and Western European ancestry) and YRI (Yoruba in Ibadan, Nigeria) populations: see Grinde et al.[65] for more details. We also considered the SNP loadings for the set of PCs that were generated without any prior LD-based filtering or pruning. For each of the first four PCs, we identified the 220 variants (10 per chromosome) with the highest absolute SNP loadings. Many of these variants with large loadings for one PC also had large loadings for another PC, so in total there were 373 unique variants selected according to this procedure. For comparison, we also selected 100 variants across the autosomes with low SNP loadings ($|\text{loading}| < 0.0008$) for all of the first four PCs. Among these 100 variants, 85 were selected such that they had different allele frequencies in the African and European ancestral populations ($|\hat{p}_{CEU} - \hat{p}_{YRI}| > 0.6$), and 15 were selected that had similar allele frequencies in the two ancestral populations ($|\hat{p}_{CEU} - \hat{p}_{YRI}| < 0.005$). Altogether, we selected 473 variants with positions spread across the genome, high or low SNP loadings, and large or small ancestral allele frequency differences to investigate the impact of different characteristics of the causal variant on GWAS results.

2.3.2 GWAS Models

For each simulated trait, we ran genome-wide association studies using models of the general form

$$E[y_i \mid g_{ij}, \mathbf{w}_i] = \alpha + \beta_j g_{ij} + \boldsymbol{\gamma}\mathbf{w}_i,$$

where y_i is the simulated quantitative trait, g_{ij} is the genotype at position j, and $\mathbf{w}_i$ is a vector of additional covariates. Note that we quantify genotype g_{ij} by the number of copies—0, 1, or, 2—of some pre-specified allele (e.g., the minor allele) carried by individual i at position j. We fit these models at every position $j = 1, \ldots, m$ across the genome and test for association between the trait and genotype by testing the null hypothesis $H_0 : \beta_j = 0$ using a traditional Wald test.

In particular, we consider four models: a model making no adjustment for ancestral heterogeneity (i.e., $\mathbf{w}_i = \emptyset$), a model adjusting for estimated admixture proportions ($\mathbf{w}_i = \hat{\pi}_i$), a model adjusting for the first principal component ($\mathbf{w}_i = u_{1i}$), and a model adjusting for the first four principal components ($\mathbf{w}_i = \begin{bmatrix} u_{1i} & u_{2i} & u_{3i} & u_{4i} \end{bmatrix}$). For the models adjusting for principal components, we consider four sets of PCs based on different pre-processing criteria: *none* (no prior LD-based exclusions or pruning), *exclusions only* (excluding regions from Table 1 but no LD pruning), *pruning only* (LD pruning with $r^2 < 0.1$ and a window size of 0.5 Mb, but not excluding regions from Table 1), and *both* (both Table 1 exclusions and LD pruning).

2.3.3 Spurious Associations

To evaluate these ancestral heterogeneity adjustment approaches, we compared the number of spurious associations that appeared when we used each model. We quantified spurious associations by counting the number of chromosomes, not including the chromosome on which the causal variant was located, with at least one variant reaching genome-wide significance. For all models, the genome-wide significance threshold was set to the $p = 5.0 \times 10^{-8}$ threshold

that is used extensively throughout the GWAS literature[68,69].

3 Results

3.1 Ancestral heterogeneity in admixed populations

Inferred admixture proportions for three samples of African American individuals are presented in Figure 1. In WHI SHARe, we compared admixture proportion estimates from a variety of model-based techniques. Figure 1A presents admixture proportions estimated as genome-wide average local ancestry, using local ancestry calls from `RFMix`. These local ancestry based admixture proportion estimates were highly correlated (Pearson correlation > 0.998) with admixture proportions from supervised and unsupervised `ADMIXTURE` analyses with two ancestral populations ($K = 2$). In TOPMed samples, we performed unsupervised `ADMIXTURE` analyses with varying numbers of ancestral populations. Figures 1B and 1C present results with $K = 2$. Although these analyses were unsupervised, based on prior studies of admixture in African Americans and the distribution of admixture proportions seen in WHI SHARe, we believe that the ancestral population colored dark gray in Figure 1 corresponds to European ancestry and the population colored light gray corresponds to African ancestry. In all three samples, we observe considerable variability in the relative proportions of African and European ancestry across individuals, showing the need to adjust for global ancestry in genome-wide association studies in these admixed samples.

3.2 First PC captures global ancestry

In an African American population, we might expect that only one principal component is needed to capture ancestral heterogeneity, at least with respect to differences in the relative proportion of African and European continental ancestry. Comparing estimated admixture proportions to principal components in WHI SHARe, JHS, and COPDGene confirms this hypothesis. In all three samples of African Americans, the first principal component is highly

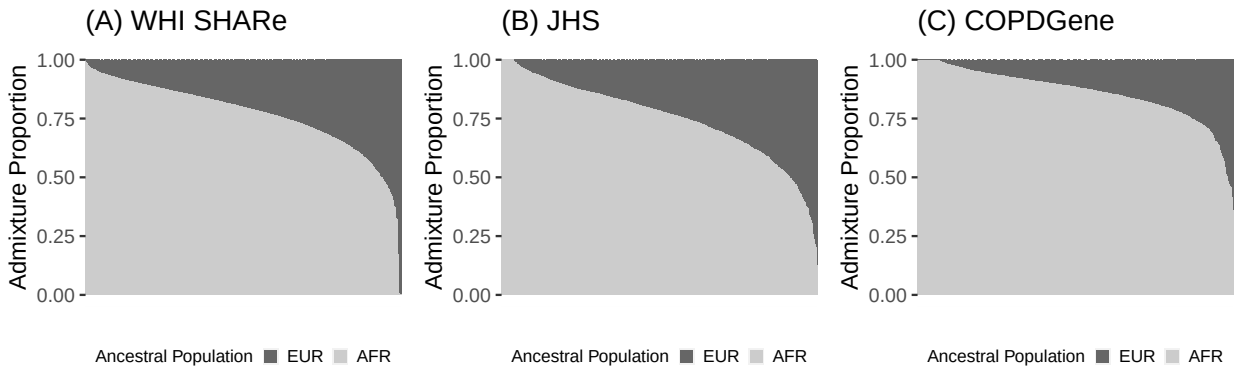

Figure 1: Estimated admixture proportions in three samples of unrelated African American individuals. Each individual is represented by a narrow vertical bar in the plot, and the individuals are sorted by their estimated proportion of African ancestry. The admixture proportions shown here were estimated using `RFMix` in WHI SHARe (A) and an unsupervised `ADMIXTURE` analysis in TOPMed JHS (B) and COPDGene (C).

correlated with the inferred proportion of African ancestry (Pearson correlation: 0.993–0.998), while later PCs show very little correlation with genome-wide continental ancestry (Pearson correlation: -0.015–0.034). We observe similar patterns of correlation between PCs and inferred admixture proportions regardless of the type of LD filtering (or lack thereof) performed prior to running PCA. See Supplemental Figures S1 and S2 for more detail.

3.3 Later PCs may capture local genomic features

As we see above, the first principal component is highly correlated with global ancestry in these African American samples, whereas later PCs are not. While it is possible that these higher-order principal components may be capturing sub-continental structure that is not captured by the estimated admixture proportions, we see in many cases that these later PCs are capturing local genomic features rather than genome-wide ancestry. This is evident from inspection of *SNP loadings*, which represent the contribution of each variant to each principal component, or from investigation of the correlation between principal component scores and genotypes.

Figure 2 presents the correlation between principal components and genotypes in JHS

and COPDGene African Americans when PCs are generated without any prior LD-based pruning or filtering. We see that variants across the genome are contributing relatively equally to the first principal component, whereas the second, third, and fourth PCs are driven by variants on a select number of chromosomes. In JHS, for example, the second PC is particularly highly correlated with variants on chromosomes 6 and 8, and also has a higher correlation with variants on chromosomes 2, 3, and 11. We see similar patterns, although with peaks on different combinations of chromosomes, in COPDGene (Figure 2B) and WHI SHARe African Americans (Figure 3A). The peaks in these genotype-PC correlation plots indicate that those principal components are primarily capturing variation at these positions along the genome rather than genome-wide global ancestry.

3.4 Impact of LD pruning

Previous authors have suggested that this phenomenon of principal components capturing local genomic features arises due to high or otherwise unusual patterns of linkage disequilibrium among variants; as a result, they recommend that variants in high LD with one another be removed prior to running PCA. Following these recommendations, we compare the set of principal components based on all variants to PCs generated after first removing regions of the genome known to have high LD (Table 1), performing LD pruning, or both.

Figure 3 illustrates the impact of these pre-processing steps on the correlation between genotypes and PCs in WHI SHARe African Americans. Panel A presents results for principal components that were generated without any prior LD-based filtering or pruning, and we see that PCs 2–4 are capturing local genomic features rather than genome-wide ancestry. When we exclude the previously-identified high LD regions reported in Table 1 before running PCA (Figure 3B), the pattern of *which* SNPs are driving PCs 2–4 changes, but the issue of PCs capturing local genomic features has not been resolved. However, after LD pruning with an r^2 threshold of 0.1 and a window size of 0.5 Mb (Figure 3C), we now see similar patterns with PCs 2–4 as we do with the first principal component: all variants are now contributing

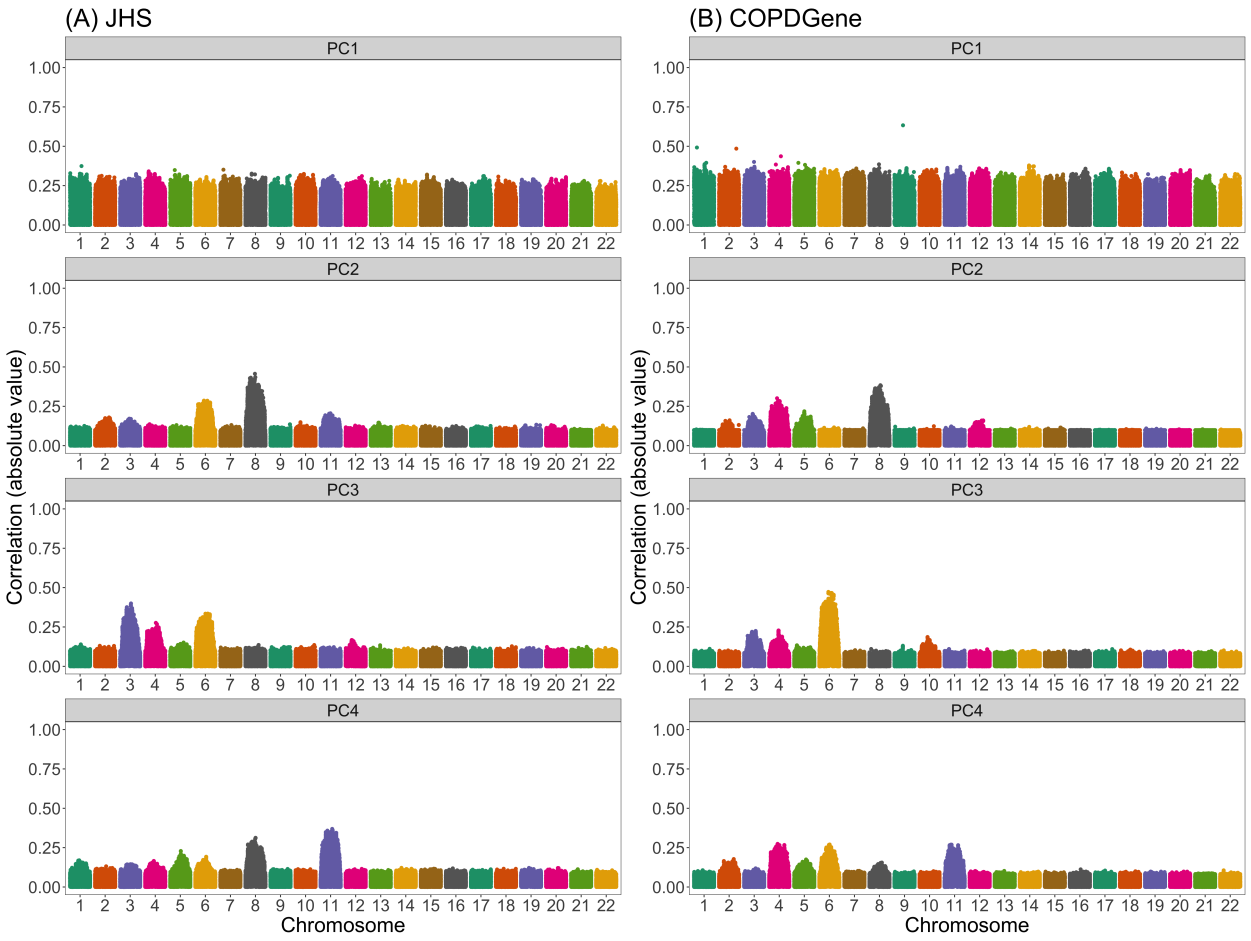

Figure 2: Correlation between naively generated PCs (i.e., PCs that were constructed without any prior LD-based filtering or exclusions) and genotypes in JHS and COPDGene African Americans. Each panel plots the absolute value of the correlation between principal components and genotypes (on the y-axis) versus the position along the genome (x-axis). Panels are organized vertically according to which PC is being investigated (1, 2, 3, 4) and horizontally according to the sample (A: JHS, B: COPDGene). Peaks in this plot indicate that a variant has a larger *loading*, i.e., a larger contribution to that principal component.

relatively equally to each PC. If we then also remove previously-identified high LD regions in addition to performing LD pruning (Figure 3D), the patterns of correlation between PCs and genotypes are indistinguishable from those with LD pruning alone. Similar patterns are observed in JHS and COPDGene (data not shown). — may want to change this if I add point about LD pruning not always working to Discussion.

Note that the thresholds for LD pruning that we use here ($r^2 < 0.1$) are stricter than the default for many software programs and the threshold used in many studies of European populations ($r^2 < 0.2$). If we use the larger r^2 threshold, we see improvement for the second and third principal components, but the fourth continues to capture local genomic features (Supplemental Figure S3).

3.5 Adjusting for PCs that capture local genomic features can induce spurious associations

We have demonstrated that, especially without strict LD pruning, principal components can capture local genomic features rather than global ancestry in admixed populations. However, it remains to be fully understood what the downstream implications would be of adjusting for these PCs in genome-wide association studies. We conducted a simulation study to investigate these implications further.

Figure 4 presents Manhattan plots from one replicate of our simulation study. In this setting, there is a single causal variant on chromosome 4, and we compare results from GWAS models using different ancestral heterogeneity adjustment approaches. As expected, we see extreme inflation, i.e., statistically significant associations on *every* chromosome, when we do not make any adjustment for ancestral heterogeneity (Figure 4A). Otherwise, when we infer and adjust for ancestral heterogeneity using either PCA or estimated admixture proportions, we see a single peak in our Manhattan plot on chromosome 4—as hoped, given that is where the causal variant is located—with one notable exception. When we adjust for the first four principal components (as has been done in previous GWAS in WHI SHARe[24,25]), where those

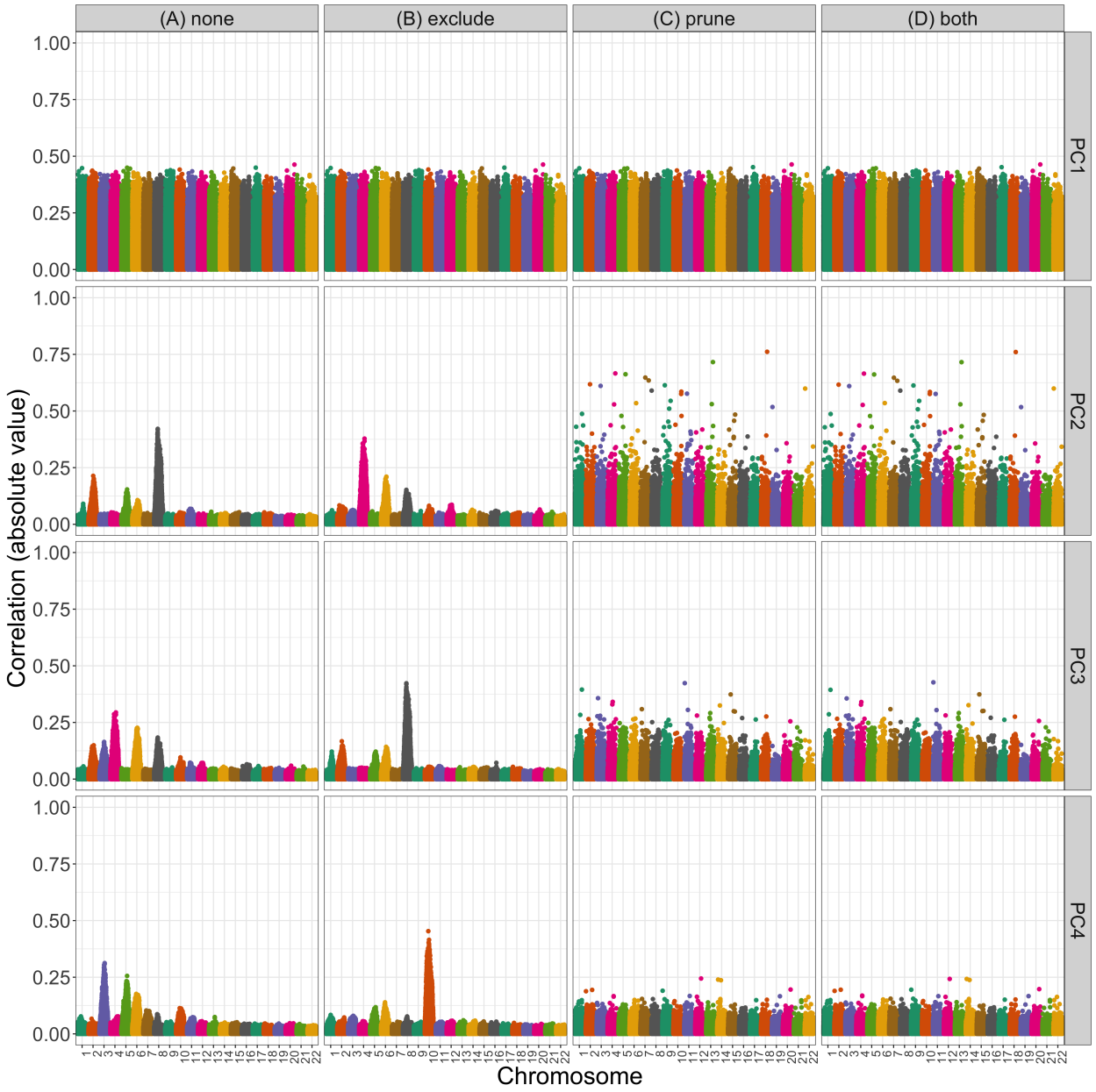

Figure 3: Correlation between PCs and genotypes in WHI SHARe African Americans with different choices of pre-processing. Each panel plots the absolute value of the correlation between principal components and genotypes (on the y-axis) versus the position along the genome (x-axis). Panels are organized vertically according to which PC is being investigated (1, 2, 3, 4) and horizontally according to the level of filtering that was applied prior to running PCA (*none*: all SNPs, *exclude*: after excluding regions in Table 1, *prune*: after LD pruning with an r^2 threshold of 0.1 and window size of 0.5 Mb, and *both*: after both exclusions and LD pruning).

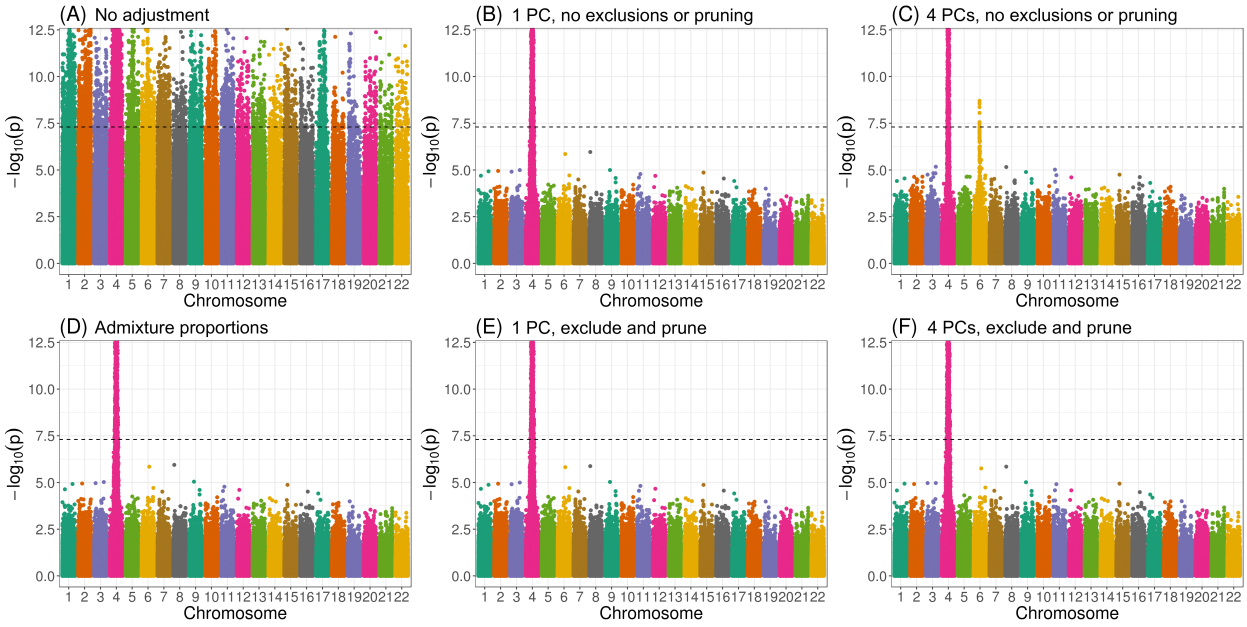

Figure 4: Manhattan plots from genome-wide association studies in WHI SHARe African Americans using different approaches to adjust for ancestral heterogeneity. In this example, the simulated trait depends only on the genotype at a single variant on chromosome 4. Panels present results using different adjustment approaches: (A) no adjustment; (B) one PC, with PCs calculated using all variants; (C) four PCs, with PCs calculated using all variants; (D) estimated admixture proportions; (E) one PC, with PCs calculated after LD pruning ($r^2 < 0.1$, window size = 0.5 Mb) and Table 1 exclusions; and (F) four PCs, with PCs calculated after LD pruning and exclusions. The horizontal dashed line in all panels represents the genome-wide significance threshold of 5×10^{-8}.

PCs were generated without any prior LD-based pruning or filtering, then we see a spurious association on chromosome 6 (Figure 4C). However, this spurious association disappears if we only adjust for the first of these PCs (Figure 4B). Likewise, no spurious association arises if we adjust for estimated admixture proportions (Figure 4D) or if we use PCs that were generated after LD pruning and Table 1 exclusions (Figure 4, panels E and F). Note that the causal variant, on chromosome 4, and the spurious signal, on chromosome 6, are both located in regions of the genome that are highly correlated with the PCs that were generated without any prior LD pruning (Figure 3).

These results are not unique to this particular causal variant. In Figure 5, we see that adjusting for PCs that capture local genomic features leads to higher numbers of spurious associations, on average, across all simulation settings. Comparing models that make some

sort of adjustment for ancestral heterogeneity, we observe the most spurious associations when GWAS models adjust for four principal components without any prior LD-based pruning or exclusions (represented by the orange solid line with circles in Figure 5). Excluding the high LD regions from Table 1 prior to running PCA (the orange solid line with triangles) reduces the number of observed spurious associations slightly, but not to the levels of the other approaches. Given what we saw in Figure 3, this is not surprising: even with these exclusions, PCs 2–4 still capture local genomic features—unless those exclusions are also combined with strict LD pruning. On the other hand, when models only include PCs that do not capture local genomic features, the rate of spurious associations drops. This includes models that adjust for just the first PC (the green lines in Figure 5) and models that include four PCs, but only after strict LD pruning (the orange dashed lines). Models that adjust for estimated admixture proportions (the purple solid line) perform nearly identically to models that adjust for the first PC. This, again, is not surprising, given the high correlation between admixture proportions and the first PC observed in this sample (Supplemental Figures S1 and S2).

3.6 Factors that influence the rate of spurious associations

Our simulation results highlight various factors that influence when, and how many, spurious associations arise when adjusting for PCs that capture local genomic features. First, we note that there are very few spurious associations, regardless of the adjustment approach (or even lack thereof), when there are small differences in ancestral allele frequencies at the causal variant (Figure 5B). This is to be expected: in this scenario, the causal variant is not associated with global ancestry, so global ancestry is not a confounding variable and there is no need for adjustment. Considering other simulation settings in which the causal variant has a larger difference in ancestral allele frequencies (panels C, D, and E of Figure 5), so adjusting for ancestral heterogeneity is needed, the number of observed spurious associations remains low for models that adjust for admixture proportions, a single principal component

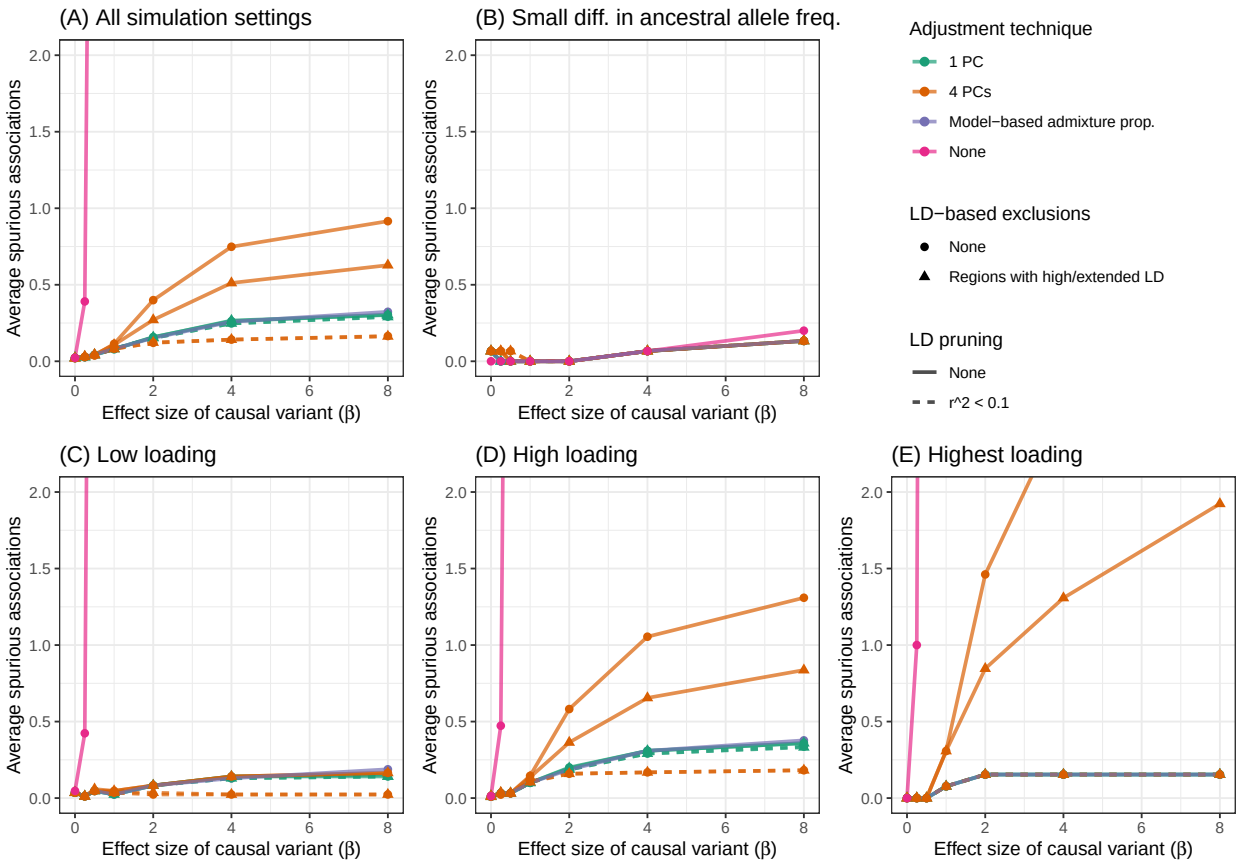

Figure 5: Comparison of the number of spurious associations in genome-wide association studies in WHI SHARe African Americans using different approaches to adjust for ancestral heterogeneity. Panel (A) displays the average number of spurious associations that were observed across all simulation settings. Remaining panels focus on the subset of simulation settings in which the causal variant has (B) a small difference in ancestral allele frequencies, (C) low SNP loadings for each of the first four PCs, (D) a high SNP loading for at least one of the first four PCs, or (D) the highest SNP loading on its chromosome for one of the first four PCs. Within each panel, we compare the number of spurious associations when GWAS models adjust for estimated admixture proportions, 1 PC (with or without LD pruning and/or Table 1 exclusions), or 4 PCs (with or without LD pruning and/or Table 1 exclusions). Results shown here are for simulated traits with a single causal variant, with effect size (β) ranging from 0 to 8.

(regardless of pre-processing), or four PCs—if those PCs were generated after strict LD pruning. For the two models that adjust for PCs capturing local genomic features (i.e., the models that adjust for 4 PCs that were generated with or without Table 1 exclusions, but no LD pruning), however, we see a higher rate of spurious associations, particularly when the causal variant is highly correlated with one of those PCs. Notably, as the size of the causal variant's SNP loading increases from low (Figure 5C), to high (Figure 5D), to the highest on its chromosome (Figure 5E), we see an increasing number of spurious associations for these two approaches. This confirms the pattern we saw in Figure 4, where a spurious association arose when we adjusted for PCs that were highly correlated with variants in several regions across the genome, and both the causal variant and spurious signal were located in one of those regions. Finally, we note that these problems worsen as the effect size of the causal variant increases.

To better understand the patterns observed in our simulation study, we compare the expected effect size estimates from GWAS models in admixed populations with two ancestral populations using different techniques for adjusting for ancestral heterogeneity. As in our simulations, we assume that the trait depends on a single causal variant:

$$y_i \overset{iid}{\sim} N(\beta_1 g_{i1} + \beta_\pi \pi_i, 1),$$

where g_{i1} represents the number of minor alleles carried by individual i at the causal variant, which we will refer to as *Variant 1*, and π_i is the individual's admixture proportion. (Note that $\beta_\pi = 0$ in our simulation study, but we consider the more general setting here.) We can then derive the expected effect size estimate at that causal variant, as well as a second variant that is not associated with the trait and sits on a different chromosome than the causal variant. When we consider a GWAS model that adjusts for the true admixture proportions, the expected effect size estimates at the causal variant (Variant 1) and the unlinked neutral

variant (Variant 2) are

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1 \\ E[\hat{\beta}_2] &= 0, \end{aligned} \tag{1}$$

where β_1 is the true effect size of the causal variant and $\beta_2 = 0$ is the true effect size of the neutral variant. In other words, models that perfectly adjust for ancestral heterogeneity will yield unbiased estimates of the effect size at the causal and unlinked neutral variants. A proof of this result can be found in Supplemental Information Section S4.

In comparison, GWAS models that do not make any adjustment for ancestral heterogeneity will yield effect size estimates of

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1 + \frac{(p_{11} - p_{10})V_\pi \beta_\pi}{p_{10}(1-p_{10}) + (p_{11}-p_{10})(1-p_{11}-p_{10})E_\pi + (p_{11}-p_{10})^2(V_\pi + E_\pi - E_\pi^2)} \\ E[\hat{\beta}_2] &= 0 + \frac{(p_{21} - p_{20})V_\pi \{\beta_\pi + 2\beta_1(p_{11}-p_{10})\}}{p_{20}(1-p_{20}) + (p_{21}-p_{20})(1-p_{21}-p_{20})E_\pi + (p_{21}-p_{20})^2(V_\pi + E_\pi - E_\pi^2)}, \end{aligned} \tag{2}$$

where E_π and V_π are the population mean and variance of the admixture proportions, β_π is the direct effect of admixture proportions on the trait, p_{11}, p_{10} are the allele frequencies of the causal variant in the two ancestral populations, and p_{21}, p_{20} are the ancestral allele frequencies of the unlinked neutral variant. (See Supplemental Information Section S4 for the derivation of Equation 2.) From these results, we see that the unadjusted model will yield a biased estimate of the effect size of the causal variant ($E[\hat{\beta}_1] \neq \beta_1]$) unless there is no ancestral heterogeneity (i.e., $V_\pi = 0$), global ancestry does not have a direct effect on the trait (i.e., $\beta_\pi = 0$), or the causal variant does not have different allele frequencies in the ancestral populations (i.e., $p_{11} = p_{10}$). We see, also, that the model can yield a biased effect size at the unlinked neutral variant ($E[\hat{\beta}_2] \neq 0$) even if global ancestry does not have a direct effect on the trait, provided that both the causal variant and the variant being tested have allele frequencies that differ between the two ancestral populations (i.e., $p_{11} \neq p_{10}$ and $p_{21} \neq p_{20}$). These biased effect size estimates at neutral variants will translate into spurious associations as sample sizes increase, just as we saw in our simulations (Figures 4 and 5).

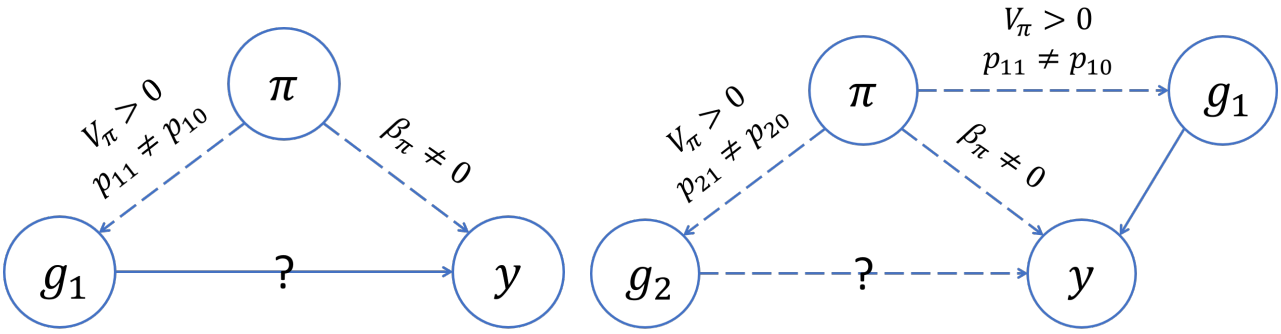

Figure 6: Directed acyclic graphs (DAGs) summarizing the conditions for confounding by global ancestry in GWAS. On the left, we see that global ancestry confounds the association at the causal variant (Variant 1) if there is ancestral heterogeneity in the population ($V_\pi > 0$), the causal variant has different allele frequencies in the ancestral populations ($p_{11} \neq p_{10}$), and global ancestry has a direct effect on the trait ($\beta_\pi \neq 0$). On the right, we see that global ancestry can confound the association at an unlinked neutral variant (Variant 2) even if global ancestry does not have a direct effect on the trait ($\beta_\pi = 0$), provided that there is ancestral heterogeneity ($V_\pi > 0$) and both the causal variant and the variant being tested have different allele frequencies in the ancestral population ($p_{11} \neq p_{10}, p_{21} \neq p_{20}$).

These results are summarized in Figure 6 and underscore the importance of adjusting for ancestral heterogeneity even when global ancestry does not have a direct effect on the trait.

To mimic the idea of adjusting for principal components that adjust for local genomic features, we also consider a scenario in which our GWAS model adjusts for two "principal components". We assume that the first principal component captures global ancestry (i.e., $\mathbf{u}_1 = \boldsymbol{\pi}$) but the second principal component captures some feature other than global ancestry (i.e., $\mathbf{u}_2 = \mathbf{z}$ for some variable z). Then, we can show (see Supplemental Information Section S4) that the expected effect size estimates at the causal variant and an unlinked neutral variant will be

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1 \\ E[\hat{\beta}_2] &= 0 + \beta_1 \frac{-V_\pi E\{\mathrm{Cov}(g_1, z \mid \pi)\} E\{\mathrm{Cov}(g_2, z \mid \pi)\}}{V_z(V_\pi V_{g_2} - C_{g_2,\pi}^2) - V_\pi C_{g_2,z}^2 + C_{\pi,z}(2C_{g_2,\pi}C_{g_2,z} - V_{g_2}C_{\pi,z})}, \end{aligned} \tag{3}$$

where $V_a = \mathrm{Var}(a)$ and $C_{a,b} = \mathrm{Cov}(a, b)$. We see that this model adjusting for an extraneous principal component will yield an unbiased effect size estimate at the causal variant, but the same is not true for the unlinked neutral variant. In particular, the effect size estimate

at this neutral variant will be biased away from zero when there is ancestral heterogeneity (i.e., $V_\pi \neq 0$) and the second principal component is correlated with both the causal variant and the variant being tested (i.e., $\text{Cov}(g_1, z \mid \pi) \neq 0$ and $\text{Cov}(g_2, z \mid \pi) \neq 0$)). In other words, these results indicate that if a model adjusts for a PC that is correlated with the causal variant as well as a second variant that is not associated with the trait, then spurious associations will arise at that second neutral variant in large enough samples. This is exactly what we observe in our simulations (Figure 4C, Figure 5D, Figure 5E). However, if the extra PC is not correlated with the causal variant, then spurious associations will not arise (Figure 4F, Figure 5C).

See Sections S4 and S5 of the Supplemental Information for derivations and simulations validating these analytic results.

4 Discussion

We observe considerable variability in global ancestry proportions across all three admixed populations studied in this paper: the Women's Health Initiative SNP Health Association Resource (WHI SHARe), Trans-Omics for Precision Medicine Jackson Heart Study (TOPMed JHS), and TOPMed Genetic Epidemiology of Chronic Obstructive Pulmonary Disease Study (COPDGene) African Americans. It is widely understood that adjusting for this ancestral heterogeneity in genome-wide association studies is needed in order to control for potential confounding by global ancestry and the spurious associations that can arise as a result. As we've shown above, this confounding can occur even when global ancestry does not have a direct effect on the trait itself, provided that there is a causal variant elsewhere in the genome that has different allele frequencies across the ancestral populations of interest. Although this fact has been recognized previously[70], it is sometimes overlooked. Our theoretical work (Equation 2) identifies the factors that impact the magnitude of the bias incurred by GWAS models that fail to adjust for global ancestry. We hope that our

results will serve as a reminder to researchers of the various ways in which global ancestry can confound genetic studies in admixed populations and the importance of ensuring that GWAS models appropriately adjust for ancestral heterogeneity.

A common approach for adjusting for ancestral heterogeneity in GWAS involves including global ancestry as a covariate in marginal regression models, with global ancestry estimated using either model-based approaches or principal component analysis. In WHI SHARe, TOPMed JHS, and TOPMed COPDGene African Americans, the first principal component is highly correlated with estimates of the genome-wide proportion of African ancestry and models adjusting for either perform similarly. Later PCs, however, do not correlate with global ancestry and models that adjust for these PCs anyway—a common practice in the literature—can yield elevated rates of spurious associations, particularly when those PCs capture local genomic features rather than genome-wide ancestry. Prior work (e.g., [49,51]) has shown that PCs can detect regions with high, extensive, or otherwise unusual patterns of linkage disequilibrium (Table 1), but the patterns observed in those studies—primarily involving individuals of European ancestry—differ from what we see here. These prior studies typically saw PCs that were driven by variants in a single high-LD region, just as we see in TOPMed COPDGene European Americans (Supplemental Figure S6). However, in all three of the admixed samples that we investigated, we see instead that PCs are often correlated with variants in *multiple* regions, across multiple chromosomes. This has important downstream implications.

Our simulations show that GWAS models adjusting for PCs that capture multiple local genomic features have elevated rates of spurious associations. This can be explained by the concept of *collider bias* (Figure 7). If a PC is correlated with the genotype of multiple variants, and at least one of those variants is associated with the trait, then the PC becomes a *collider variable* when testing the association between the other variants and the trait. Adjusting for collider variables can induce a spurious association between variables that are otherwise unlinked[71]. This is precisely what we see above. Our theoretical work (Equation

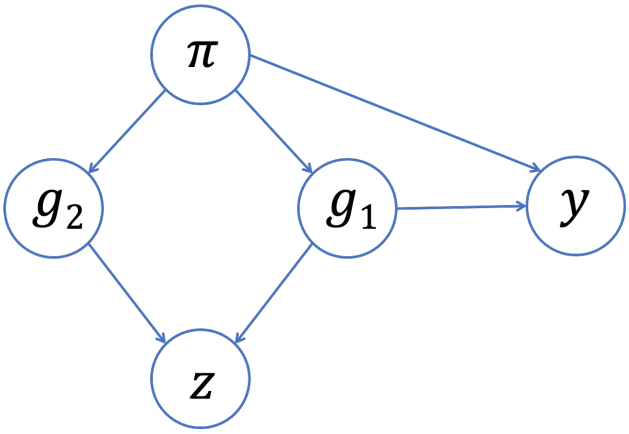

Figure 7: Collider bias in GWAS. Suppose that, instead of genome-wide global ancestry (π), a principal component (z) captures the genotype of two variants (g_1, g_2). If one of those variants (g_1) is associated with the trait (y), then the PC will be a collider variable—rather than a confounding variable—when testing the association between the other variant (g_2) and the trait. Adjusting for the PC can then induce a spurious association between the trait and the neutral variant as a result of collider bias.

3) shows that GWAS models adjusting for an extraneous PC will yield biased estimates of variant effect sizes, with the magnitude of that bias increasing with the effect size of the causal variant, the strength of the correlation between the PC and the variants it captures, and the amount of ancestral heterogeneity within the population. When that bias is large enough, it can lead to a spurious association—as our simulations show. Given the similarities in terms of which regions of the genome tend to be correlated with PCs across different admixed populations (see Figures 2 and 3, for example), it is possible that these spurious associations may even replicate across studies.

Increasing attention has been paid to the issue of collider bias in genetic association studies in recent years[72,73,74,75], but to our knowledge this is the first paper to fully demonstrate the concerns related to including extraneous principal components in GWAS models. In fact, the focus in the literature has typically been on demonstrating issues that can arise when adjusting for *too few* principal components[7,18,76]. One recent paper does show that adjusting for principal components can lead to replicating spurious associations in gene expression studies due to collider bias[77]. As a short aside, the authors mention that collider bias could arise in GWAS, but they imply that the magnitude of this bias would be small—

and thus not of practical concern—given that "causal SNPs have much lower leverage on genetic PCs" (i.e., the correlation between PCs and causal variants is small). Our findings support this conclusion that the magnitude of collider bias depends on the strength of the correlation between the PCs and the variants they capture. However, we see that this correlation between PCs and genetic variants can be non-trivial in admixed populations (without careful pre-processing prior to running PCA), implying that collider bias is of more concern in this setting than previous authors may have realized. We also see that the amount of bias depends on the variance of admixture proportions across a population (denoted V_π in Equation 3), which could explain why work that has focused on more ancestrally homogenous populations may not have identified this issue previously. That said, it is important to raise the question of the magnitude of bias that could be expected in a "typical" genome-wide association study. The smaller the effect size of the causal variant, the smaller the number of spurious associations that we observed in our simulation study (Figure 5), and when studying complex traits/diseases we would expect the causal variant effect sizes to be fairly small. However, in these settings we would also expect that there are *multiple* causal variants—not just a single causal variant, as we assumed for the sake of simplicity in our simulations and theoretical work—and then these effects may be aggregated[78]. In any case, it is worth considering what steps can be taken to reduce or eliminate concerns about potential collider bias altogether.

In our analysis of genotype and sequence data from unrelated WHI SHARe, TOPMed JHS, and TOPMed COPDGene African Americans, we found that all but the first principal component were largely driven by small regions of the genome—and thus have the potential to be collider variables—unless careful pre-processing of genotype data was performed prior to running PCA. As mentioned earlier, previous studies have found that PCs can be driven primarily by small regions of the genome, and as a result have suggested that these regions be excluded (Table 1) and/or that LD pruning be performed prior to running PCA. However, the motivation for this LD-based filtering has typically been framed in terms of the ability of the

principal components to capture global ancestry, as well as the computational complexity of running PCA, rather than the downstream implications on association testing that we have highlighted here. Furthermore, we have found that excluding the regions listed in Table 1, without also performing LD pruning, does not seem to solve these issues in admixed populations: we still see peaks in the PC-genotype correlation plots and models adjusting for these PCs continue to show elevated rates of spurious associations. A more tedious iterative approach of identifying and removing potentially problematic regions based on our own data also does not prevent PCs from capturing local genomic features within a reasonable number of iterations (Supplemental Figure S5). LD pruning, on the other hand, proves to be more successful, at least when looking at the first four principal components. However, our results highlight that a stricter threshold (e.g., $r^2 = 0.1$) is needed for LD pruning in admixed populations than the $r^2 = 0.2$ threshold that is often suggested in the literature: see Supplemental Information Section S2 for a comparison. Given that linkage disequilibrium patterns differ between admixed populations and the European populations upon which much of this prior work was based, it is not surprising that different LD-based filtering techniques are required here. [... even strict LD pruning will not protect against false positive associations when the causal variant is associated with the PC and the sample size is sufficiently large ...]

Ultimately, our work demonstrates the challenges that can arise in appropriately adjusting for ancestral heterogeneity in genome-wide association studies in admixed populations. For populations where we have a good idea of the number of ancestral populations of interest and relevant reference panel data is readily available, GWAS models adjusting for estimated global ancestry proportions rather than principal components perform well. PCA can offer advantages over model-based ancestry inference methods, but careful consideration must be given to how many PCs should be included and what those PCs are capturing. In the African American samples studied here, for example, a single principal component was sufficient for controlling spurious associations induced by population structure. In some simulation set-

tings, we did see a small drop in the number of spurious associations when we included three additional PCs, but only after strict LD pruning. Careful pre-processing of data prior to running PCA, combined with thorough diagnostics (e.g., by calculating and plotting SNP loadings or the correlation between PCs and genotypes), is critical to ensure that models do not include principal components that could cause the very problem—spurious associations—that these techniques aim to solve.

5 Supplemental Information

Supplemental Information includes 13 figures, as well as proofs and simulations validating the theoretical results presented in the main paper.

6 Declaration of Interests

If you have anything to disclose (see https://www.cell.com/declaration-of-interests), please let me know.

The authors declare no competing interests.

7 Acknowledgments

The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A listing of WHI investigators can be found at https://www.whi.org/doc/WHIInvestigator-Long-List.pdf.

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung, and Blood Institute (NHLBI). Core support including centralized genomic-read mapping and genotype calling along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The Jackson Heart Study is supported

and conducted in collaboration with Jackson State University (HHSN268201300049C and HHSN268201300050C), Tougaloo College (HHSN268201300048C), and the University of Mississippi Medical Center (HHSN268201300046C and HHSN268201300047C) contracts from NHLBI and the National Institute for Minority Health and Health Disparities (NIMHD); genome sequencing was funded by HHSN268201100037C. The COPDGene study was supported by NIH grants U01 HL089856 and U01 HL089897. The COPDGene project is also supported by the COPD Foundation through contributions made by an Industry Advisory Board comprised of Pfizer, AstraZeneca, Boehringer Ingelheim, Novartis, and Sunovion.

K.E.G. was supported in part by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1256082. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. S.R.B. and B.L.B. were supported by the National Human Genome Research Institute of the National Institutes of Health under award number HG010869. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

If there is anything else you would like me to add, please let me know.

8 Data and Code Availability

WHI SHARe genotype data is available on dbGaP (accession number: phs000386) or directly through WHI according to the policy outlined at https://www.whi.org/doc/WHI-genetic-data-transfer-policy.pdf. TOPMed whole genome sequence data is also available from dbGaP (Jackson Heart Study: phs000964, Genetic Epidemiology of Chronic Pulmonary Disease Study: phs000951). All software packages used throughout this paper are freely available online:

- bcftools[63] (quality control): https://samtools.github.io/bcftools/

- RFMix[32] (local ancestry inference): https://sites.google.com/site/rfmixlocalancestryinference/
- ADMIXTURE[31] (global ancestry inference): https://dalexander.github.io/admixture/
- PCRelate[60] and PC-AiR[36] (identifying unrelated individuals): https://rdrr.io/bioc/GENESIS/
- SNPRelate[35] (LD pruning, PCA, and PCA-related diagnostics): https://www.bioconductor.org/packages/release/bioc/html/SNPRelate.html
- PLINK[52] (GWAS): https://zzz.bwh.harvard.edu/plink/
- TOPMed Analysis Pipeline (identifying unrelated individuals, LD pruning, PCA, PCA-related diagnostics, and association studies in whole genome sequence data): https://github.com/UW-GAC/analysis_pipeline
- R (analyzing and visualizing results): https://cran.r-project.org/

Other resources pertaining to this paper, including download-able lists of the high LD regions in Table 1 in various builds, can be found on the lead author's GitHub page:

- https://github.com/kegrinde/PCA.

The GitHub repository is currently private, but I'll make it public once the paper is submitted. If anyone would like to see the repository before I submit, let me know and I'll add you as a collaborator.

9 References

[1] Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., Forrester, T., Allison, D. B., Deka, R., Ferrell, R. E. *et al.* (1998). Estimating african american admixture proportions by use of population-specific alleles. The American Journal of Human Genetics *63*, 1839–1851.

[2] Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O. *et al.* (2009). The genetic structure and history of africans and african americans. Science *324*, 1035–1044.

[3] Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S. A. *et al.* (2010). Genome-wide patterns of population structure and admixture in west africans and african americans. Proceedings of the National Academy of Sciences *107*, 786–791.

[4] Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C. D., and Ostrer, H. (2010). Genome-wide patterns of population structure and admixture among hispanic/latino populations. Proceedings of the National Academy of Sciences *107*, 8954–8961.

[5] Conomos, M. P., Laurie, C. A., Stilp, A. M., Gogarten, S. M., McHugh, C. P., Nelson, S. C., Sofer, T., Fernández-Rhodes, L., Justice, A. E., Graff, M. *et al.* (2016). Genetic diversity and association studies in us hispanic/latino populations: applications in the hispanic community health study/study of latinos. The American Journal of Human Genetics *98*, 165–184.

[6] Devlin, B. and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004.

[7] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics *38*, 904–909.

[8] Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. Nature Genetics *36*, 512–517.

[9] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics *11*, 459–463.

[10] Need, A. C. and Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. Trends in Genetics *25*, 489–494.

[11] Bustamante, C. D., Francisco, M., and Burchard, E. G. (2011). Genomics for the world. Nature *475*, 163–165.

[12] Popejoy, A. B. and Fullerton, S. M. (2016). Genomics is failing on diversity. Nature News *538*, 161.

[13] Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E., Hutter, C. M., Manolio, T. A., and Green, E. D. (2018). Prioritizing diversity in human genomics research. Nature Reviews Genetics *19*, 175.

[14] Manolio, T. A. (2019). Using the data we have: improving diversity in genomic research. The American Journal of Human Genetics *105*, 233–236.

[15] Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. Science *265*, 2037–2048.

[16] Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for

linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). American journal of human genetics *52*, 506.

[17] Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B. *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics *38*, 203–208.

[18] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E. *et al.* (2010). Variance component model to account for sample structure in genome-wide association studies. Nature Genetics *42*, 348–354.

[19] Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nature Genetics *46*, 100–106.

[20] Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. The American Journal of Human Genetics *67*, 170–181.

[21] Consortium, W. T. C. C. *et al.* (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661.

[22] Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A. *et al.* (2007). A whole-genome association study of major determinants for host control of hiv-1. Science *317*, 944–947.

[23] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R. *et al.* (2008). Genes mirror geography within europe. Nature *456*, 98–101.

[24] Reiner, A. P., Beleza, S., Franceschini, N., Auer, P. L., Robinson, J. G., Kooperberg, C., Peters, U., and Tang, H. (2012). Genome-wide association and population genetic analysis of c-reactive protein in african american and hispanic american women. The American Journal of Human Genetics *91*, 502–512.

[25] Carty, C. L., Johnson, N. A., Hutter, C. M., Reiner, A. P., Peters, U., Tang, H., and Kooperberg, C. (2012). Genome-wide association study of body height in African Americans: The Women's Health Initiative SNP Health Association Resource (SHARe). Human Molecular Genetics *21*, 711–720.

[26] Pino-Yanes, M., Gignoux, C. R., Galanter, J. M., Levin, A. M., Campbell, C. D., Eng, C., Huntsman, S., Nishimura, K. K., Gourraud, P.-A., Mohajeri, K. *et al.* (2015). Genome-wide association study and admixture mapping reveal new loci associated with total ige levels in latinos. Journal of Allergy and Clinical Immunology *135*, 1502–1510.

[27] Akenroye, A. T., Brunetti, T., Romero, K., Daya, M., Kanchan, K., Shankar, G., Chavan, S., Boorgula, M. P., Ampleford, E. A., Fonseca, H. F. *et al.* (2021). Genome-wide association study of asthma, total ige, and lung function in a cohort of peruvian children. Journal of Allergy and Clinical Immunology *148*, 1493–1504.

[28] Conti, D. V., Darst, B. F., Moss, L. C., Saunders, E. J., Sheng, X., Chou, A., Schumacher, F. R., Olama, A. A. A., Benlloch, S., Dadaev, T. *et al.* (2021). Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. Nature genetics *53*, 65–75.

[29] Tang, H., Peng, J., Wang, P., and Risch, N. J. (2005). Estimation of individual admixture: analytical and study design considerations. Genetic Epidemiology *28*, 289–301.

[30] Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics *164*, 1567–1587.

[31] Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research *19*, 1655–1664.

[32] Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics *93*, 278–288.

[33] Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet *8*, e1002453.

[34] Durand, E. Y., Do, C. B., Mountain, J. L., and Macpherson, J. M. (2014). Ancestry composition: a novel, efficient pipeline for ancestry deconvolution. biorxiv , 010512.

[35] Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of snp data. Bioinformatics *28*, 3326–3328.

[36] Conomos, M. P., Miller, M. B., and Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genetic epidemiology *39*, 276–293.

[37] Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet *2*, e190.

[38] McVean, G. (2009). A genealogical interpretation of principal components analysis. PLoS Genet *5*, e1000686.

[39] Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., and Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. Statistics in Medicine *34*, 3769–3792.

[40] Raska, P., Iversen, E., Chen, A., Chen, Z., Fridley, B. L., Permuth-Wey, J., Tsai, Y.-Y., Vierkant, R. A., Goode, E. L., Risch, H. *et al.* (2012). European american

stratification in ovarian cancer case control data: the utility of genome-wide data for inferring ancestry. Plos one *7*, e35235.

[41] Daya, M., Rafaels, N., Brunetti, T. M., Chavan, S., Levin, A. M., Shetty, A., Gignoux, C. R., Boorgula, M. P., Wojcik, G., Campbell, M. *et al.* (2019). Association study in african-admixed populations across the americas recapitulates asthma risk loci in non-african populations. Nature Communications *10*, 1–13.

[42] Abegaz, F., Chaichoompu, K., Génin, E., Fardo, D. W., König, I. R., Mahachie John, J. M., and Van Steen, K. (2019). Principals about principal components in statistical genetics. Briefings in Bioinformatics *20*, 2200–2216.

[43] Mathieson, I. and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nature Genetics *44*, 243–246.

[44] Liu, N., Zhao, H., Patki, A., Limdi, N. A., and Allison, D. B. (2011). Controlling population structure in human genetic association studies with samples of unrelated individuals. Statistics and its interface *4*, 317.

[45] Abdellaoui, A., Hottenga, J.-J., De Knijff, P., Nivard, M. G., Xiao, X., Scheet, P., Brooks, A., Ehli, E. A., Hu, Y., Davies, G. E. *et al.* (2013). Population structure, migration, and diversifying selection in the netherlands. European Journal of Human Genetics *21*, 1277–1285.

[46] Weale, M. E. (2010). Quality control for genome-wide association studies. Genetic Variation , 341–372.

[47] Tian, C., Plenge, R. M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., Klareskog, L., Pulver, A. E., Qi, L., Gregersen, P. K. *et al.* (2008). Analysis and application of european genetic substructure using 300 k snp information. PLoS Genet *4*, e4.

[48] Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D. *et al.* (2008). Long-range ld can confound genome scans in admixed populations. The American Journal of Human Genetics *83*, 132–135.

[49] Zou, F., Lee, S., Knowles, M. R., and Wright, F. A. (2010). Quantification of population structure using correlated snps by shrinkage principal components. Human Heredity *70*, 9–22.

[50] Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Cornelis, M. C., Edenberg, H. J. *et al.* (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic Epidemiology *34*, 591–602.

[51] Privé, F., Luu, K., Blum, M. G., McGrath, J. J., and Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. Bioinformatics *36*, 4449–4457.

[52] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J. *et al.* (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics *81*, 559–575.

[53] Yu, K., Wang, Z., Li, Q., Wacholder, S., Hunter, D. J., Hoover, R. N., Chanock, S., and Thomas, G. (2008). Population substructure and control selection in genome-wide association studies. PloS one *3*, e2551.

[54] Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., Briley, L. P., Maruyama, Y., Waterworth, D. M., Waeber, G. *et al.* (2008). The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. The American Journal of Human Genetics *83*, 347–358.

[55] Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. Nature Protocols *5*, 1564–1573.

[56] Zhang, Y., Guan, W., and Pan, W. (2013). Adjustment for population stratification via principal components in association analysis of rare variants. Genetic epidemiology *37*, 99–109.

[57] Galinsky, K. J., Bhatia, G., Loh, P.-R., Georgiev, S., Mukherjee, S., Patterson, N. J., and Price, A. L. (2016). Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. The American Journal of Human Genetics *98*, 456–472.

[58] Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M. G. (2018). Efficient analysis of large-scale genome-wide data with two r packages: bigstatsr and bigsnpr. Bioinformatics *34*, 2781–2787.

[59] Hays, J., Hunt, J. R., Hubbell, F. A., Anderson, G. L., Limacher, M., Allen, C., and Rossouw, J. E. (2003). The Women's Health Initiative recruitment methods and results. Annals of Epidemiology *13*, S18–S77.

[60] Conomos, M. P., Reiner, A. P., Weir, B. S., and Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. The American Journal of Human Genetics *98*, 127–148.

[61] Jun, G., Wing, M. K., Abecasis, G. R., and Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale dna sequence data. Genome Research *25*, 918–925.

[62] Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M. *et al.* (2021). Sequencing of 53,831 diverse genomes from the nhlbi topmed program. Nature *590*, 290–299.

[63] Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. *et al.* (2021). Twelve years of samtools and bcftools. Gigascience *10*, giab008.

[64] Consortium, I. H. . *et al.* (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52.

[65] Grinde, K. E., Brown, L. A., Reiner, A. P., Thornton, T. A., and Browning, S. R. (2019). Genome-wide significance thresholds for admixture mapping studies. The American Journal of Human Genetics *104*, 454–465.

[66] Parker, M. M., Foreman, M. G., Abel, H. J., Mathias, R. A., Hetmanski, J. B., Crapo, J. D., Silverman, E. K., Beaty, T. H., and Investigators, C. (2014). Admixture mapping identifies a quantitative trait locus associated with fev1/fvc in the copdgene study. Genetic Epidemiology *38*, 652–659.

[67] Kirk, J. L. (2016). *Statistical methods for inferring population structure with human genome squence data*. PhD thesis, University of Washington.

[68] Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic Epidemiology *32*, 381–385.

[69] Jannot, A.-S., Ehret, G., and Perneger, T. (2015). P¡ 5× 10- 8 has emerged as a standard of statistical significance for genome-wide association studies. Journal of Clinical Epidemiology *68*, 460–465.

[70] Wacholder, S., Rothman, N., and Caporaso, N. (2002). Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiology Biomarkers & Prevention *11*, 513–520.

[71] Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. Annual Review of Sociology *40*, 31.

[72] Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L., and Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. The American Journal of Human Genetics *96*, 329–339.

[73] Day, F. R., Loh, P.-R., Scott, R. A., Ong, K. K., and Perry, J. R. (2016). A robust example of collider bias in a genetic association study. The American Journal of Human Genetics *98*, 392–393.

[74] Cai, S., Hartley, A., Mahmoud, O., Tilling, K., and Dudbridge, F. (2022). Adjusting for collider bias in genetic association studies using instrumental variable methods. Genetic Epidemiology .

[75] Hemani, G., Tilling, K. M., and Smith, G. D. (2022). Collider bias from selecting disease samples distorts causal inferences. Genetic Epidemiology *46*, 213–215.

[76] Yao, Y. and Ochoa, A. (2022). Limitations of principal components in quantitative genetic association models for human studies. bioRxiv .

[77] Dahl, A., Guillemot, V., Mefford, J., Aschard, H., and Zaitlen, N. (2019). Adjusting for principal components of molecular phenotypes induces replicating false positives. Genetics *211*, 1179–1189.

[78] Di, Y., Mi, G., Sun, L., Dong, R., Zhu, H., and Peng, L. (2011). Power of association tests in the presence of multiple causal variants. In BMC Proceedings volume 5 pages 1–5, Springer.