

# 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,<sup>1\*</sup> Timothy A. Thornton,<sup>2,3</sup> Brian L. Browning,<sup>4</sup> Sharon R. Browning<sup>2</sup>

1. Department of Mathematics, Statistics, and Computer Science, Macalester College, Saint Paul, MN, 55105, USA
2. Department of Biostatistics, University of Washington, Seattle, WA, 98195, USA
3. [... Regeneron ...]
4. Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, 98195, USA

\* kgrinde@macalester.edu

# Abstract

## NEEDS UPDATING

Principal component analysis (PCA) is widely used to control for population structure in genome-wide association studies (GWAS). It has been shown that the top principal components (PCs) typically reflect population structure, but deciding exactly how many PCs to include in GWAS regression models can be challenging. Often researchers will err on the side of including more PCs than may be actually necessary in order to ensure that population structure is fully captured. However, through both analytic results and application to TOPMed whole genome sequence data for 1,888 and 2,676 unrelated African American individuals from the Jackson Heart Study (JHS) and Chronic Obstructive Pulmonary Disease Genetic Epidemiology Study (COPDGene), respectively, we show that adjusting for extraneous PCs can actually induce spurious associations. In particular, spurious associations arise when PCs capture local genomic features, such as regions of the genome with atypical linkage disequilibrium (LD) patterns, rather than genome-wide ancestry. In JHS and COPDGene, we show that careful LD pruning prior to running PCA, using stricter thresholds and wider windows than is often suggested in the literature, can resolve these issues, whereas excluding lists of high LD regions identified in previous studies does not. We also show that the rate of spurious associations can be appropriately controlled in these data when we simply adjust for either the first PC or a model-based estimate of admixture 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

Considerable variability in global ancestry—the genome-wide proportion of genetic material inherited from each ancestral population—has been observed in many studies of admixed populations such as African Americans and Hispanics/Latinos<sup>1,2,3,4,5</sup>. It has been widely documented that heterogeneous global ancestry, as with other types of population structure, can lead to spurious associations in genome-wide association studies<sup>6,7,8,9</sup>. In fact, some authors have even cited the ancestral heterogeneity of admixed populations, and the statistical challenges it poses, as one of many reasons why these populations have been historically underrepresented in genome-wide association studies (GWAS)<sup>10,11,12,13,14</sup>. Spurious associations can arise in GWAS in ancestrally heterogeneous populations when global ancestry confounds the association between genotypes and the phenotype of interest (Figure 1). 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.

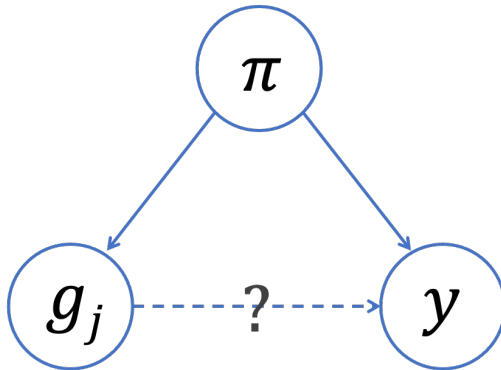


Figure 1: Global ancestry ( $\pi$ ) confounds the association between the genotype at position  $j$  ( $\mathbf{g}_j$ ) and the phenotype of interest ( $\mathbf{y}$ ) if ancestry is associated with both the genotype (e.g., the allele frequencies differ across the ancestral populations) and the phenotype (e.g., there are environmental or other factors that affect the phenotype and differ across the ancestral populations).

A number of methods for detecting and controlling for ancestral heterogeneity in ge-

netic association studies have been proposed. Early approaches included restricting analyses to subsets of ancestrally homogeneous individuals<sup>15</sup>, performing a genome-wide correction for test statistic inflation due to ancestral heterogeneity via *genomic control*<sup>6</sup>, and using family-based designs<sup>16</sup>. More recently, approaches based on mixed models have been proposed<sup>17,18,19</sup>, 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<sup>7,20</sup>. This fixed effects adjustment for global ancestry is currently used extensively throughout the literature, with global ancestry inferred using either model-based ancestry inference methods (e.g., **frappe**<sup>21</sup>, **STRUCTURE**<sup>22</sup>, **ADMIXTURE**<sup>23</sup>) or principal component analysis (e.g., **EIGENSTRAT**<sup>7</sup>, **SNPRelate**<sup>24</sup>, **PC-AiR**<sup>25</sup>).

Principal component analysis (PCA) is a widely-implemented unsupervised approach for inferring global ancestry that does not require reference panel data or pre-specification of the number of ancestral populations of interest and is capable of capturing sub-continental structure<sup>26</sup>. To infer global ancestry using PCA, we perform a singular value decomposition of the matrix of standardized genotypes (i.e.,  $\mathbf{X} = \mathbf{UDV}^\top$ ) or, equivalently, an eigenvalue decomposition of the genetic relationship matrix (i.e.,  $\mathbf{XX}^\top = \mathbf{UD}^2\mathbf{U}^\top$ ). It has been shown that top eigenvectors, or *principal components* (PCs),  $\mathbf{u}_1, \mathbf{u}_2, \dots$  tend to reflect global ancestry<sup>27,28</sup>. To adjust for ancestral heterogeneity, we choose some number of PCs to include as covariates in our GWAS regression models.

Determining the number of PCs needed to capture global ancestry can be difficult. Numerous techniques have been proposed for selecting this number, including formal significance tests based on Tracy-Widom theory<sup>27,7</sup>, examining inflation factors<sup>29,5</sup> and/or the proportion of variance explained by each PC<sup>30,29,5</sup>, comparing PCs to self-reported race/ethnicity<sup>5</sup>, and keeping PCs that are significantly associated with the trait<sup>31,32</sup>. Typically, the number of PCs selected is on the order of one to ten<sup>33</sup>, but in practice it is not uncommon to see applications in which more many more PCs are used—more even than may actually be necessary

to capture global ancestry. This could be due in part to work that has suggested that including higher-order PCs can provide the safeguard of removing “virtually all stratification”<sup>34</sup> at the cost of only “subtle” decreases in power<sup>35</sup>.

Another challenge that can arise in using PCA to adjust for ancestral heterogeneity involves ensuring that PCs actually reflect global ancestry and not some other features or artifacts of the data. Prior work has shown that PCs can capture relatedness across samples<sup>27,9,36,25</sup>, array artifacts or other data quality issues<sup>27,7,9,37</sup>, and/or small regions of the genome with unusual patterns of linkage disequilibrium (LD)<sup>27,7,38,39,40,9,37,41,42,36,43</sup>. To address this last issue, some authors have suggested running PCA on a reduced subset of variants after first performing *LD pruning*, using a program such as PLINK<sup>44</sup> to remove variants that are in “high” LD (e.g., pairwise-correlation  $r^2 > 0.2$ ) with nearby variants<sup>38,45,26,46,47,48,37,42,36,49,25,29,50,5,32</sup>, and/or excluding regions of the genome that are known to have extensive, long-ranging, or otherwise unusual patterns of LD<sup>38,45,26,40,48,37,30,5</sup>. A list of these previously-identified high LD regions and references that recommend their exclusion is provided in Table 1.

The above-cited suggestions regarding LD pruning and filtering are not universally implemented and the downstream implications of adjusting for PCs that capture features other than global ancestry are not fully understood. Furthermore, much of this 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 conduct simulation studies using whole genome sequence data for African American individuals in the Trans-Omics for Precision Medicine (TOPMed) project and provide analytic results to show that including too many PCs can actually induce spurious associations in GWAS, particularly when those extraneous PCs capture local genomic features rather than genome-wide ancestry. — ADD WHI To

Chr	Start (bp)	End (bp)	References
1	48000000	52060567	48,40,37
2	85941853	100500000	48,40,37
2	129600000	140000000	40,26,37,30,5,51
2	182882739	190000000	48,40,37
3	47500000	50000000	48,40,37
3	83500000	87000000	48,40,37
3	89000000	97500000	40,37
3	163100000	164900000	51
5	44000000	51500000	45,48,40,37
5	98000000	100500000	40,37
5	129000000	132000000	48,40,37
5	135500000	138500000	40,37
6	23800000	39000000	45,48,40,26,37,30,5,51
6	57000000	64000000	48,40,37
6	140000000	142500000	48,40,37
7	55000000	66193285	48,40,37
8	6300000	13500000	45,48,40,26,39,37,30,5,51
8	43000000	50000000	48,40,37
8	112000000	115000000	48,40,37
10	37000000	43000000	48,40,37
11	45000000	57000000	45,40,37
11	87500000	90500000	48,40,37
12	33000000	40000000	48,40,37
12	109500000	112021663	40,37
14	46600000	47500000	51
17	37800000	42000000	26,5
20	32000000	34500000	48,40,37

Table 1: Regions of the genome with high, long-range, or otherwise unusual patterns of linkage disequilibrium (LD) that are often recommended for exclusion prior to running PCA. This list of regions was generated on the basis of an extensive literature review. Start and end physical (base pair) positions are provided with respect to genome build 36. Also available for download (in builds 36, 37, or 38) at <https://github.com/kegrinde/PCA/>. **UPDATE TO REFLECT WHI ANALYSES**

conclude, we provide suggestions regarding best practice for appropriately controlling for ancestral heterogeneity in genome-wide association studies in admixed populations.

## 2 Material and Methods

ADD WHI — see dissertation

### 2.1 TOPMed Whole Genome Sequence Data

The Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Project is an ongoing project sponsored by the National Heart, Lung, and Blood Institute (NHLBI) that is working to collect and analyze whole-genome sequences, other 'omics data, and rich 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 such studies: the Jackson Heart Study (JHS) (accession number: phs000964) and the Genetic Epidemiology of Chronic Obstructive Pulmonary Disease Study (COPDGene) (accession number: phs000951). In total, the freeze 4 JHS dataset includes 3,406 African American individuals and the freeze 5b COPDGene dataset includes 8,742 African American and European American individuals.

For TOPMed freezes 4 and 5b, high coverage ( $\approx 30X$ ) whole genome sequencing was performed by several sequencing centers. Variant discovery and genotype calling was performed by the TOPMed Informatics Resources Center (IRC) using the `GotCloud` pipeline<sup>52</sup>. Quality control (QC) was performed by the sequencing centers, IRC, and TOPMed Data Coordinating Center, and only those samples and variants that passed these stages of QC are included in the VCF downloaded from dbGaP. Details on TOPMed sequencing and QC methods are available in Taliun et al.<sup>53</sup> and on the TOPMed website: <https://topmed.nhlbi.nih.gov/data->

sets.

## 2.2 Additional Quality Control and Filtering

Prior to principal component analysis, we perform additional variant- and sample-level filtering. We use `bcftools`<sup>54</sup> to remove indels and otherwise restrict our analyses to biallelic single nucleotide variants (SNVs). We also remove variants with low minor allele frequency ( $< 1\%$ ) and/or **high rates of missing calls ( $> 1\%$ ) — STILL NEEDS TO BE IMPLEMENTED**. After this filtering, a total of **???** SNVs remain in JHS and **???** SNVs remain in COPDGene.

At the sample level, we use the iterative procedure proposed by Conomos et al.<sup>55</sup> and implemented in the TOPMed Analysis Pipeline to identify a subset of mutually unrelated individuals using a kinship threshold of 0.044 (third degree relatives). We also perform an unsupervised **ADMIXTURE**<sup>23</sup> analysis with both  $K = 2$  and  $K = 3$  to identify admixed (African American) and non-admixed (European American) individuals; we restrict remaining analyses to admixed individuals only. Prior to both of these analyses we implement LD pruning/filtering as recommended in their respective user manuals. After exclusions, a total of **???** and **???** unrelated African Americans remain in JHS and COPDGene, respectively.

### 2.2.1 LD-Based Filtering

In addition to the filtering described above, we also implement different types of LD-based filtering. [... **Describe the different types of LD-based filtering we compared (see below).** ...] These analyses are also compared to a *naive* analysis that did not perform any LD-based filtering. The number of variants that remain after each type of filtering is presented in Table 2.

- Exclude
  - None
  - Lit Review (Table 1)



- Auto-Detect ([... using Prive package — implement or skip for now?? ...])
- Prune
  - None
  - Default 0.2
  - Stricter 0.1
  - [... Stricter 0.05 ??? ...]
  - [... Different window size ??? ...]

	Naive	Exclude	Prune	Stricter Prune	Exclude + Stricter Prune
JHS	[... ? ...]	[... ? ...]	[... ? ...]	[... ? ...]	[... ? ...]
COPDGene	[... ? ...]	[... ? ...]	[... ? ...]	[... ? ...]	[... ? ...]

Table 2: Single nucleotide variants that remained in Jackson Heart Study (JHS) and Genetic Epidemiology of COPD (COPDGene) datasets after varying levels of LD-based filtering.

- what filtering was performed, and how many variants left after filtering
  - JHS, ADMIXTURE: see above
  - JHS, PCA: exclude regions (TRUE/FALSE), r-squared (1, 0.1, 0.2, 0.05), window size (0, 0.5, 10), and MAF (0, 0.01)
    - \* no filtering: FALSE-1-0-0
    - \* MAF filtering: FALSE-1-0-0.01
    - \* exclude but no prune: TRUE-1-0-0.01
    - \* prune but no exclude: FALSE-0.1-0.5-0.01 and FALSE-0.1-10-0.01 and FALSE-0.2-0.5-0.01 and FALSE-0.05-0.5-0.01
    - \* prune and exclude: TRUE-0.1-0.5-0.01 and TRUE-0.1-10-0.01 and TRUE-0.2-0.5-0.01 and TRUE-0.05-0.5-0.01
  - COPD, ADMIXTURE: see above

- COPD, PCA: exclude regions (TRUE/FALSE), r-squared (1, 0.1, 0.2, 0.05), window size (0, 0.5, 10), MAF (0, 0.01)
  - \* no filtering: FALSE-1-0-0
  - \* MAF filtering: FALSE-1-0-0.01
  - \* exclude but no prune: TRUE-1-0-0.01
  - \* prune but no exclude: FALSE-0.1-0.5-0.01, FALSE-0.1-10-0.01, FALSE-0.2-0.5-0.01, FALSE-0.05-0.5-0.01
  - \* prune and exclude: TRUE-0.1-0.5-0.01, TRUE-0.1-10-0.01, TRUE-0.05-0.5-0.01, TRUE-0.2-0.5-0.01
- COPD, also ran SNPRelate on Europeans with different levels of filtering (FALSE-0.1-0.5-0.01, FALSE-0.2-0.5-0.01, FALSE-1-0-0.01, FALSE-1-0-0, TRUE-0.1-0.5-0.01, TRUE-0.2-0.5-0.01, TRUE-1-0-0.01)

## 2.3 Principal Component Analysis

We use the `SNPRelate` package in R to run principal component analysis using each of the subsets of SNVs described in Section 2.2. For each set of principal components, we also use the `SNPRelate` package to assess the contribution of each variant to each PC by calculating and plotting the correlation between genotypes and PCs. [... also look at loadings? shouldn't this give us the same picture as corr? ...]

## 2.4 Simulation Study to Investigate Rates of Spurious Associations

We implement a simulation study to explore the impact of different variant-level filtering choices, particularly with respect to linkage disequilibrium, on rates of spurious associations in genome-wide association studies using models that adjust for ancestral heterogeneity using principal components.

### 2.4.1 Simulated Traits

[...

- find loading peaks from "naive" approach
- simulate trait that is  $\beta \cdot x + \text{rnorm}(0, 1)$ , where  $\beta = 1$  or  $2$  and  $x = \text{genotype}$  at one of the peaks

...]

### 2.4.2 GWAS Models

To perform genome-wide association studies in samples of unrelated admixed individuals, we use marginal regression models, regressing the trait of interest on the genotype at each position across the genome. At a given position  $j$ , we quantify genotype  $g_{ij}$  as the number of copies (0, 1, or 2) of some pre-specified allele (e.g., the minor allele) carried by individual  $i$  at that position. Considering a quantitative trait  $y_i$ , we fit one linear regression model at each position ( $j = 1, \dots, m$ ):

$$E[y_i \mid g_{ij}, \mathbf{z}_i] = \beta_0 + \beta_j g_{ij} + \boldsymbol{\beta}_z \mathbf{z}_i,$$

where  $\mathbf{z}_i$  is a vector of additional covariates (e.g., potential confounding variables) that we want to include in the model. We test for an association between the trait and genotype by testing the null hypothesis  $H_0 : \beta_j = 0$  at each position  $j = 1, \dots, m$ .

To adjust for ancestral heterogeneity, we include inferred global ancestry in the vector  $\mathbf{z}_i$  of potential confounders in our regression models. We infer global ancestry using one of two techniques: model-based global ancestry inference or principal component analysis. [... describe which models we compare ...]

[...

- for each of 188\*2 simulated phenotypes

- for each set of PCs
- including 1, 4, or 10 PCs

...]

### 2.4.3 Evaluation

[... defining spurious associations ...]

## 2.5 Code and Data Availability

All code and data used throughout this paper are publicly available online:

- TOPMed Sequence Data: <https://www.ncbi.nlm.nih.gov/gap/>
- Code: <https://github.com/kegrinde/PCA>

## 3 Results

[...

- streamline to just focus on WHI in results section, and refer quickly to TOPMed analyses in Discussion/Supplement

...]

### 3.1 Ancestral heterogeneity in admixed populations

Inferred global ancestry proportions for TOPMed JHS African Americans, TOPMed COPDGene African Americans, WHI SHARe African Americans, and [... WHI SHARe Hispanic Americans ...] are presented in Figure 2. As has been the case in other studies of admixed populations, we observe considerable variability in the relative proportions of African and European (and, in the case of WHI SHARe Hispanic Americans, Native American) ancestry

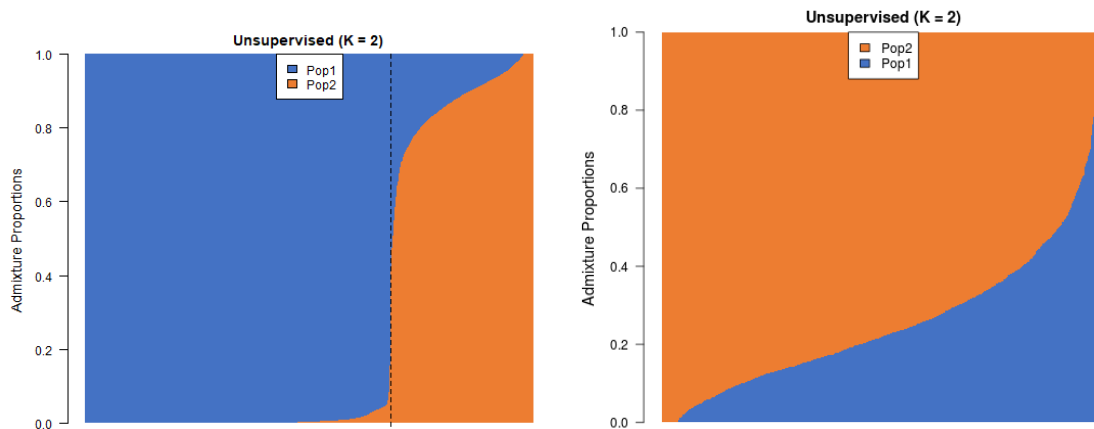


Figure 2: Barplots of estimated admixture proportions from unsupervised ADMIXTURE analyses with  $K = 2$  in TOPMed COPDGene African Americans (left panel) and JHS African Americans (right panel).

across these samples. This ancestral heterogeneity motivates the need to carefully adjust for global ancestry in genome-wide association studies in these, and other, admixed samples.

[... Barplots:

- JHS African Americans only (remove EA)
- COPDGene African Americans only (remove EA)
- WHI SHARe African Americans
- WHI SHARe Hispanic Americans

...]

### 3.2 Comparison of principal components and model-based admixture proportions

Comparing model-based admixture proportions generated by ADMIXTURE to principal components confirms that initial PCs are capturing genome-wide continental ancestry. In all three African American samples (Figure 3), the first PC is highly correlated with the inferred proportion of African/European ancestry, while later PCs show very little correlation

with global ancestry. [... In WHI SHARe Hispanic Americans (Figure 4, we see ... ...)] 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 Figure ??? ...]).

[... Scatterplot grids:

- AA: scatterplots of AFR admixture proportions vs the first four PCs (across columns) for each sample JHS/COPD/WHI (across rows)
- HA: scatterplots of admixture proportions AFR/EUR/NAM (across rows) vs the first four PCs (across columns)
- supplement: show that similar patterns arise even after strict LD pruning

...]

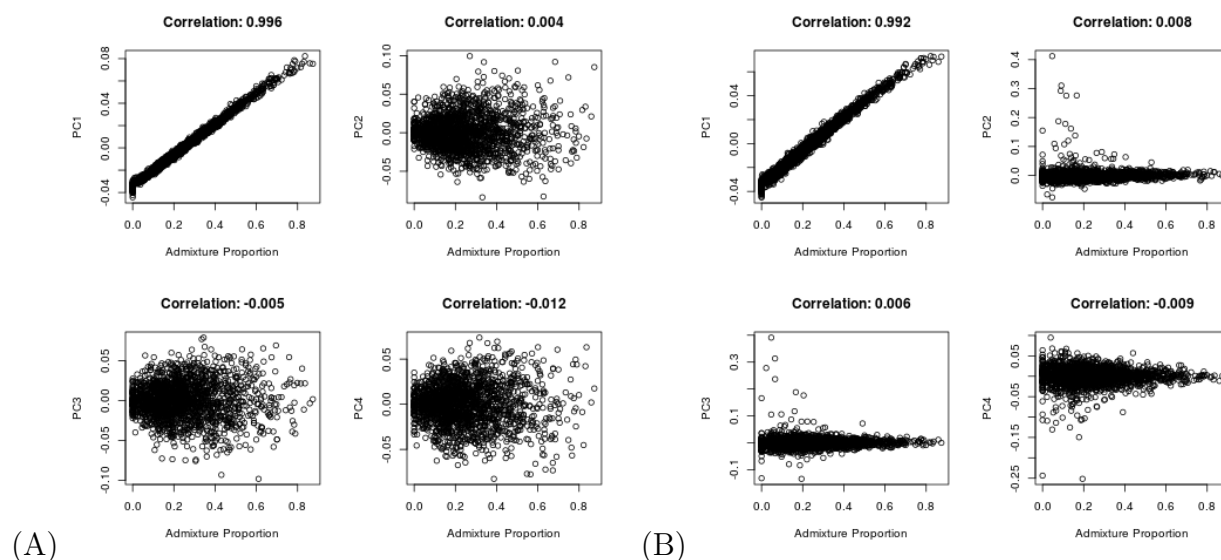


Figure 3: Comparison of model-based admixture proportions (from an unsupervised ADMIXTURE with  $K = 2$ ) and the first four principal components from various sets of PCs generated after different levels of LD filtering:

(A) PCs generated in JHS African Americans without any LD filtering.

(B) PCs generated in JHS African Americans after excluding known high LD regions and pruning with an  $r^2$  threshold of 0.1 and window size of 10 Mb. — NEEDS UPDATING PER DIRECTIONS ABOVE

Figure 4: Comparison of model-based admixture proportions and the first four principal components in WHI SHARe Hispanic Americans. PCs were generated without any LD filtering.

### 3.3 Investigation of PCs capturing local genomic features and the impact of LD pruning

As we see in Figures 3 and 4, initial principal components seem to be capturing global ancestry, 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 model-based admixture proportions, we see in many cases that these later PCs are actually capturing local genomic features rather than genome-wide ancestry. This is evident upon inspection of SNP *loadings*—which represent the contribution of each SNP to each PC—and the correlation between PCs and the original genotypes. For example, Figure 5 presents the correlation between PCs and genotypes in WHI SHARe African Americans. We see (in the leftmost panel of the figure) that SNPs across the genome are contributing relatively equally to the first PC, whereas the second, third, and fourth PCs are driven more-so by SNPs on just a select number of chromosomes. [... For example, the second PC seems to be capturing ... ...] Similar patterns are observed in JHS African Americans, COPDGene African Americans, and [... WHI SHARe Hispanic Americans ...] (Supplemental Figures [... ?? ...]): after the first PC (or the second in the case of WHI SHARe Hispanic Americans), all later PCs exhibit multiple peaks in their genotype-PC correlation plots, indicating that those PCs are capturing local genomic features on a handful of chromosomes rather than genome-wide global ancestry.

As mentioned above, previous authors have suggested that this phenomenon arises due to LD among SNPs; as a result, they recommend that SNPs 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 SNPs to PCs generated after first removing regions of the genome known to have high LD (Table 1), performing LD pruning, or both. The

remaining panels of Figure 5 show the correlation between genotypes and these other sets of PCs in WHI SHARe African Americans. When we exclude previously-identified high LD regions before running PCA (the second column of Figure 5), 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 (third column), we now see similar patterns with PCs 2–4 as we do with the first principal component — SNPs are now contributing relatively equally to each PC. If we then also remove previously-identified high LD regions in addition to performing LD pruning (last column), the patterns of correlation between PCs and genotypes are indistinguishable from running LD pruning alone. [... What do we see in WHI SHARe Hispanic Americans? ...] [... What about later PCs? ...] [... What do we see in JHS and COPD? refer to supplement ...] [... Correlation plots:

- Correlation plot of WHI AA with different levels of LD filtering (none vs exclude vs recommended prune vs exclude + prune)
- For supplement: correlation plots for JHS, COPD, and WHI HA (none, exclude, best prune, both) — or should we put WHI HA in main body of paper?
- For supplement: correlation plots for WHI AA with PCs 1–10
- For supplement: correlation plots for WHI AA comparing LD pruning choices

...]

Figure 5: Correlation between PCs and genotypes in WHI SHARe African Americans. Each panel plots the absolute value of the correlation (y-axis) between principal components and genotypes at each position along the genome (x-axis). Panels are stratified according to which PC is being investigated (1, 2, 3, or 4) and what level of LD filtering 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 a window size of 0.5 Mb, or *both* (after exclusions and LD pruning).

[... for discussion:



- why is this happening? (LD)
- how does what we see compare to what's been observed in Europeans?
- does LD pruning universally fix the problem (i.e., does it seem to work better in some samples than others)? how many PCs does it help (i.e., what do PCs 5–10 look like)?
- why do we think exclusions didn't work? (high LD regions identified in Europeans, patterns of LD differ—more extensive—in admixed populations)

...]

### 3.4 Implications of adjusting for PCs that capture local genomic features

We have demonstrated that, especially without strict LD pruning, principal components can capture local genomic features rather than global ancestry in admixed populations — but what are the downstream implications of adjusting for these PCs in genome-wide association models? Figure 6 presents a comparison of the rate of spurious associations in genome-wide association studies in WHI SHARe African Americans. [... Adjusting for PCs that capture local genomic features leads to higher rates of spurious associations ...]

Figure 6: Comparison of the average number of spurious associations in genome-wide association studies in WHI SHARe African Americans.

Figure 7: Manhattan plots from GWAS in WHI SHARe African Americans using different approaches to adjust for ancestral heterogeneity.

[... when/where do spurious assoc arise?

- when causal SNP has high contribution to one of the PCs you adjusted for (6)
- spurious assoc arises at another SNP that is also correlated with same PC (7)

why do spurious assoc arise?

- discuss theoretical results, connect to idea of collider bias

...]

## 4 Discussion

Need to address ancestral heterogeneity in admixed populations

- we observe considerable heterogeneity in global ancestry proportions in admixed populations studied here, as in other studies
- well-established that global ancestry is a potential confounding variable
- this confounding can exist even if global ancestry does not have a direct effect on the trait (as demonstrated by our simulation studies and theoretical results)
- → important to carefully measure and adjust for ancestral heterogeneity in GWAS in admixed populations

Comparing (naive) PCs and admixture proportions

- both widely used for measuring and adjusting for ancestral heterogeneity
- in AA, first PC correlated with global ancestry but later PCs are not (in HL: TBD)
- instead, later PCs often capture local genomic features (e.g., regions with extensive LD)
- while this has been documented before, note that, in contrast to what has been observed in EUR, we see that PCs seem to capture SNPs on more than one chromosome (multiple peaks in SNP loading plots); whereas in EUR we often just see one peak (cite Zou, Prive, other examples?) → why? LD patterns in admixed pops differ from those in EUR

## Spurious associations

- adjusting for these PCs can lead to spurious associations
- this is due to a phenomenon known as collider bias
- ADD: what do theory and sims tell us about when/how likely a spurious association is to occur?

## Impact of LD pruning and removing high LD regions

- after LD pruning, PCs no longer exhibit patterns of being driven by select few SNPs (at least for PCs 2-4)
- note that we had to use smaller  $r^2$  and wider windows than often recommended in literature for this to be true  $\rightarrow$  why? LD patterns in admixed pops differ from those in EUR
- excluding previously-identified high LD regions doesn't seem to be as effective  $\rightarrow$  why? LD patterns in admixed pops differ from those in EUR
- note, too, that even strict LD pruning doesn't seem to remove all correlation between SNPs and genotypes (e.g., later PCs in WHI, TOPMed COPDGene)

## Recommendations

- If using PCs, carefully inspect SNP loadings and/or correlation between PCs and genotypes
- If using PCs, don't use more than you need
- Consider using global ancestry proportions (although further work is needed to reliably capture sub-continental structure)

## 5 Appendices

### 5.1 Regions Removed Prior to PCA

- a list of all "high-LD" regions removed prior to running PCA

### 5.2 Mathematical Derivations

- theoretical results
- proofs
- simulations validating theory

## Supplemental Data

Supplemental Data include [...] figures and [...] tables.

## Declaration of Interests

The authors declare no competing interests.

## Acknowledgments

K.E.G. was supported 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.

## Web Resources

GitHub Repository: lists of regions to exclude, code for LD pruning, excluding, and plotting loadings

## Data and Code Availability

## 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] 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.
  - [22] 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.
  - [23] Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* *19*, 1655–1664.
  - [24] 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.
- [25] 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.
  - [26] 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.
  - [27] Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet* *2*, e190.
  - [28] McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genet* *5*, e1000686.
  - [29] 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.
  - [30] 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.
  - [31] 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.

- [32] 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.
- [33] 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.
- [34] Mathieson, I. and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics* *44*, 243–246.
- [35] 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.
- [36] 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.
- [37] Weale, M. E. (2010). Quality control for genome-wide association studies. *Genetic Variation* , 341–372.
- [38] 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.
- [39] 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.

- [40] 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.
- [41] 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.
- [42] 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.
- [43] Privé, F., Luu, K., Blum, M. G., McGrath, J. J., and Vilhjálmsdóttir, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* *36*, 4449–4457.
- [44] 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.
- [45] 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.
- [46] 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.

- [47] 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.
- [48] 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.
- [49] 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.
- [50] 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.
- [51] 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.
- [52] 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.
- [53] 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.
- [54] 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.

- [55] 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.
  
- [56] 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.

## Figure Titles and Legends

## Tables