Supplemental Information

July 22, 2022

Contents

List of Supplemental Figures

S1 Comparison of PCs and Model-Based Admixture Proportions

In many African American populations, only one principal component may be needed to capture ancestral heterogeneity, at least with respect to differences in the relative proportion of African and European continental ancestry. We investigated whether this statement holds true in three samples of African American individuals from the Women's Health Initiative SNP Health Association Resource (WHI SHARe) and two Trans-Omics for Precision Medicine (TOPMed) contributing studies: the Jackson Heart Study (JHS) and the Chronic Obstructive Pulmonary Disease Genetic Epidemiology Study (COPDGene). Comparing model-based admixture proportions (estimated using `RFMix`[1] in WHI SHARe and an unsupervised `ADMIXTURE`[2] analysis in JHS and COPDGene) to principal components shows that the first PC is in fact highly correlated with the inferred proportion of African ancestry in these samples, while later PCs show very little correlation with genome-wide continental ancestry. This pattern holds regardless of whether PCs are generated with (Figure S2) or without (Figure S1) prior filtering or pruning based on linkage disequilibrium (LD).

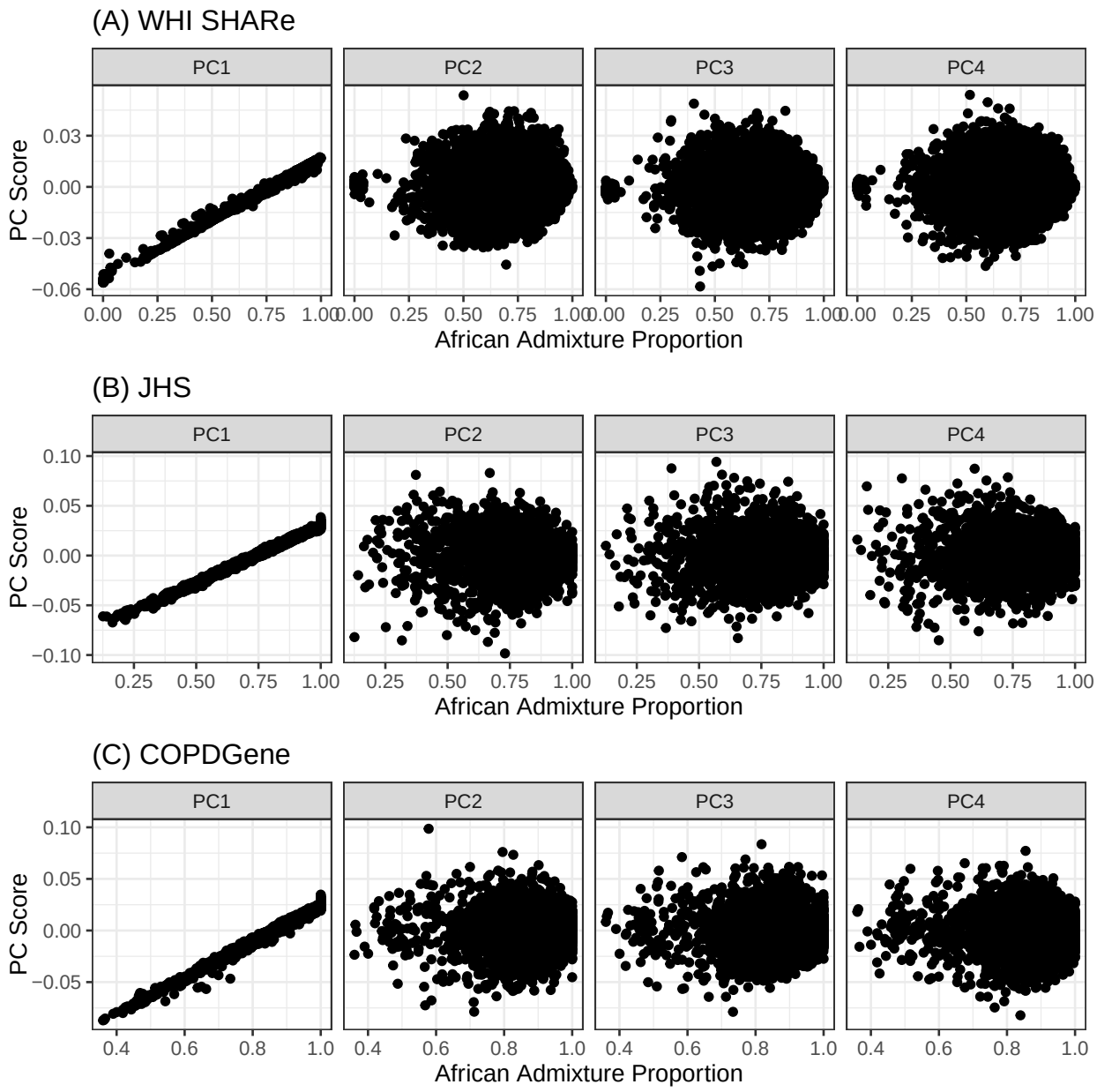

Figure S1: Scatterplots of estimated African admixture proportions versus the first four PCs in (A) WHI SHARe, (B) TOPMed JHS, and (C) TOPMed COPDGene African Americans. Here we consider PCs that were generated on the entire set of SNPs (i.e., without any prior LD-based filtering or pruning).

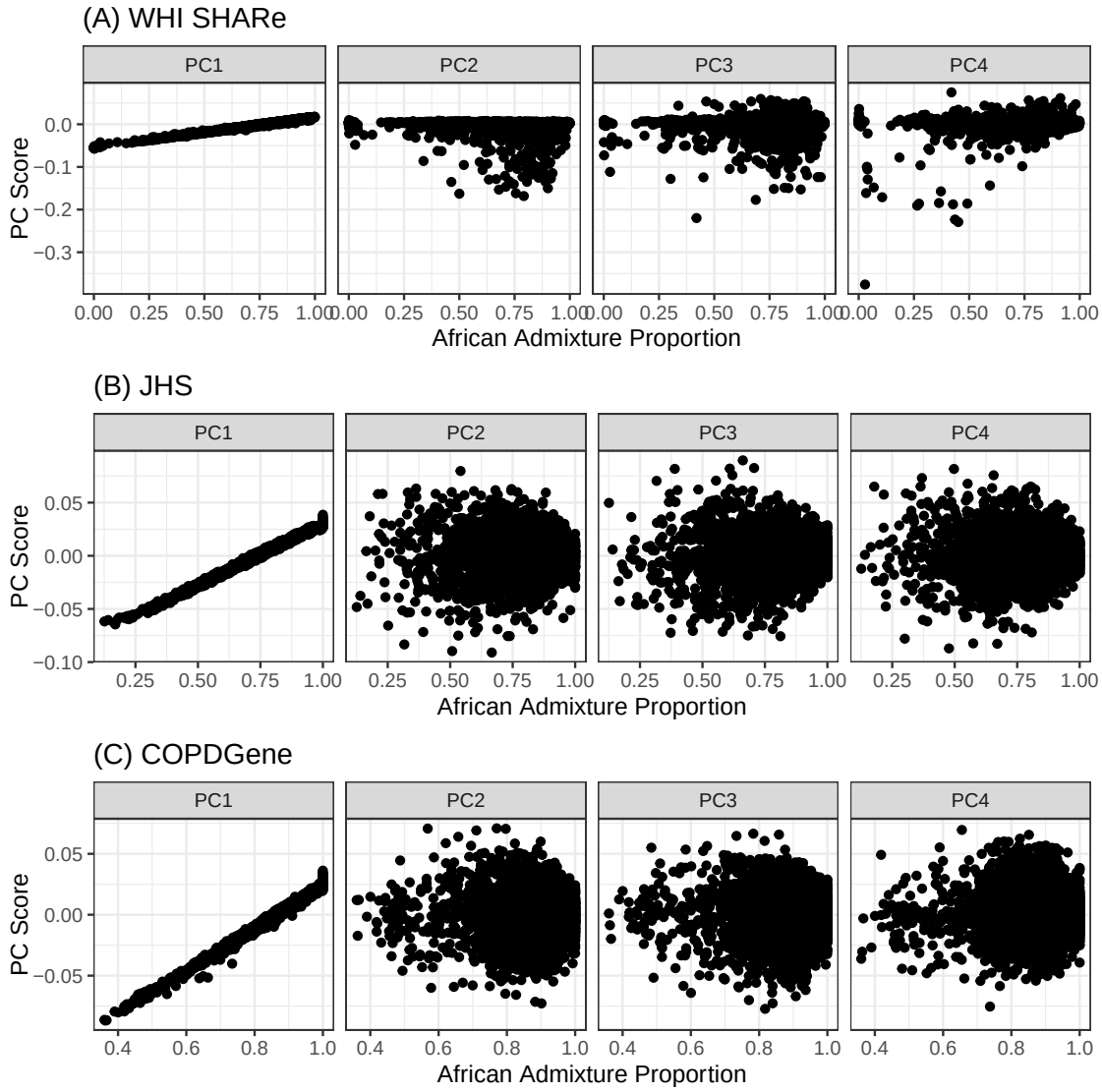

Figure S2: Scatterplots of estimated African admixture proportions versus the first four PCs in (A) WHI SHARe, (B) TOPMed JHS, and (C) TOPMed COPDGene African Americans. Here we consider PCs that were generated after both LD pruning ($r^2 = 0.1$, window size = 0.5 Mb) and filtering previously identified high-LD regions (Table 1).

S2 Comparison of PCA Pre-Processing Choices

We have shown that adjusting for principal components that capture small regions of the genome rather than genome-wide ancestry can induce spurious associations in genome-wide association studies. This problematic behavior occurred in our analysis of genotype data from WHI SHARe African Americans when PCs were generated using all 551,025 available SNPs or if we excluded regions identified in the literature as being potentially problematic for PCA (Table 1). However, problems were ameliorated when we used PCs that were generated after strict LD pruning, using an r^2 threshold of 0.1 and window size of 0.5 Mb. In this section, we further investigate the behavior of PCs generated after different filtering techniques.

Many authors have suggested using an r^2 threshold of 0.2 for LD pruning prior to running PCA[3,4,5,6,7,8,9,10]. Furthermore, this threshold is the default for LD pruning software such as `SNPRelate`[11]. However, in our analysis of WHI SHARe data, we found that using an r^2 threshold of 0.2 prior to running PCA still led to one of the top PCs (the fourth) being highly correlated with small regions of the genome, while if we used a stricter threshold of 0.1 the peaks have disappeared (at least for the top four PCs). See Figure S3 for a comparison of the correlation between PCs and genotypes in WHI SHARe African Americans across different choices of r^2 threshold.

When performing LD pruning, another choice that practitioners have to make is the window size. In the literature, various window sizes have been suggested, including 10 Mb[12], 2 Mb[4], or 0.5 Mb (the `SNPRelate` default), and others have suggested that window size may not have a big impact[6]. Similar to the latter, in our analysis of WHI SHARe data we see little difference in the correlation between PCs and genotypes across different choices of window sizes: see Figure S4. Smaller window sizes are less computationally intensive, so we used the window size of 0.5 Mb for the remainder of our analyses.

Finally, we also considered filtering out regions that were highly correlated with PCs in our own data, as has been done previously[9,3]. To implement this data-based filtering, we

investigated the SNP loadings for each of the top four PCs. Starting with the second PC, we found the SNP on each chromosome with the largest loading: if this loading was larger than 0.005, we excluded the SNP and all SNPS within M Mb; if the loading was small, we kept all SNPs on the chromosome. (We considered $M = 1, 5, 10$, and 20 Mb.) We repeated this process for PCs 3 and 4, then re-ran PCA using the remaining SNPs. Using these new PCs, we re-calculated SNP loadings and looked to see if there were still regions of the genome that were driving the PCs. If so, we repeated the entire process. This data-based filtering process is very tedious, and even after four rounds of exclusions with $M = 5$ Mb we found that the problematic behavior did not totally go away (Figure S5).

In WHI SHARe data, at least, strict LD pruning is the most effective of the pre-processing steps that we considered in eliminating the correlation between PCs and genotypes in small regions of the genome.

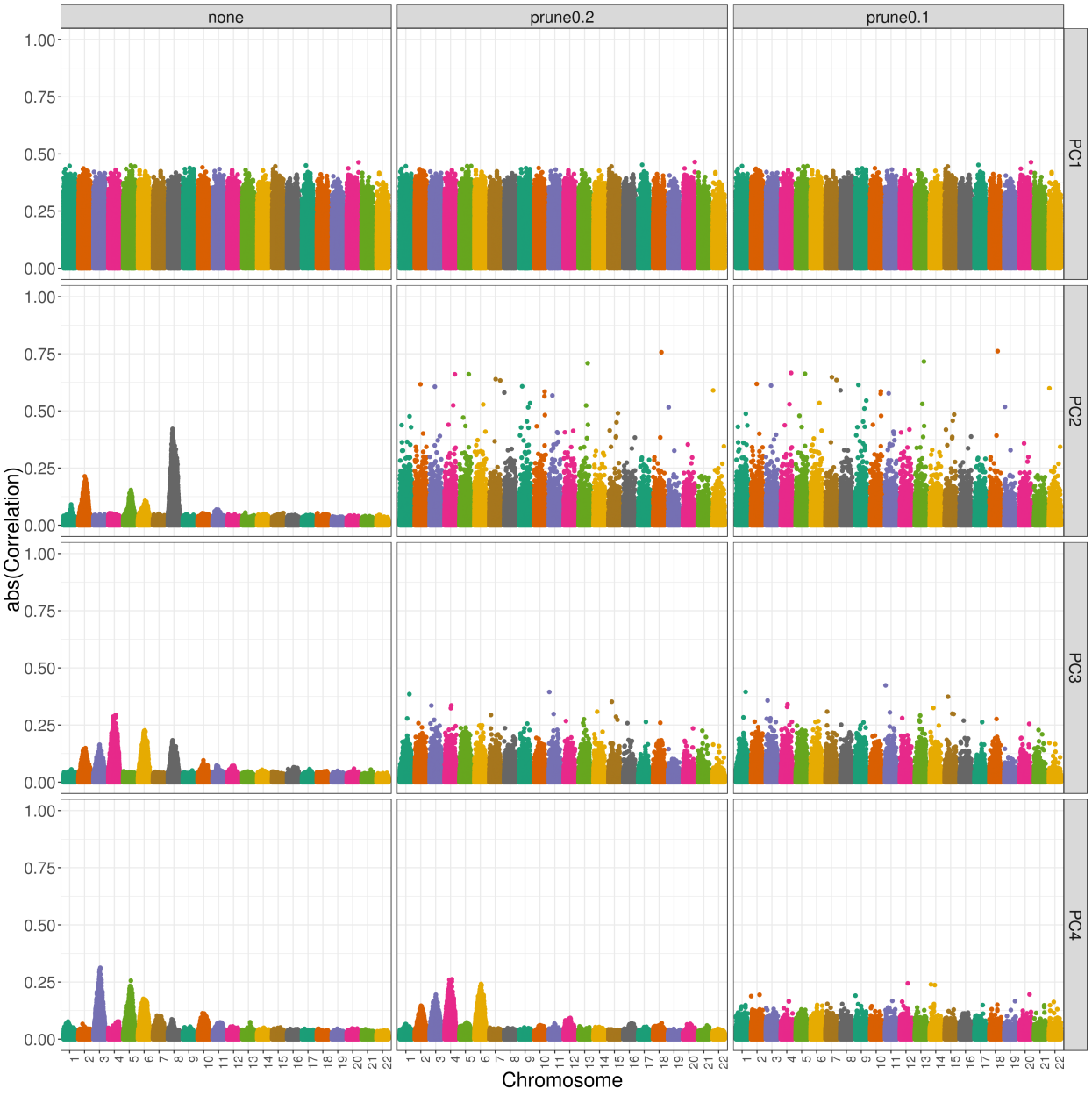

Figure S3: Correlation between PCs and genotypes in WHI SHARe African Americans using different LD pruning thresholds. Each panel plots the absolute value (abs) of the correlation between principal components and genotypes on the y-axis versus the position along the genome on the x-axis. Panels are organized vertically according to which PC is being investigated (1, 2, 3, 4) and horizontally according to what r^2 threshold was used when running LD pruning prior to PCA (*none*: no LD pruning, *prune0.2*: LD pruning with an r^2 threshold of 0.2 and window size of 0.5 Mb, and *prune0.1*: LD pruning with an r^2 threshold of 0.1 and window size of 0.5 Mb).

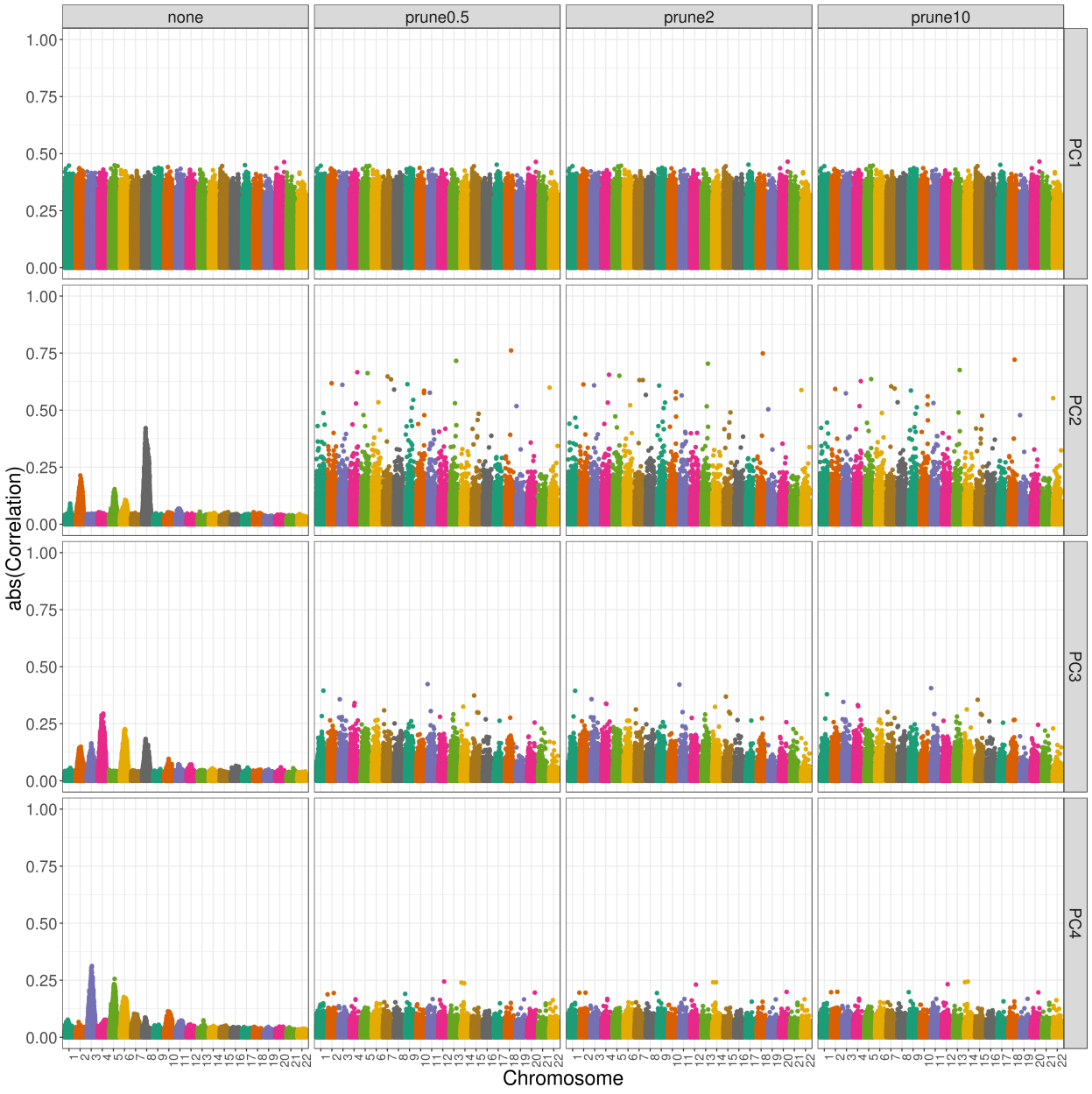

Figure S4: Correlation between PCs and genotypes in WHI SHARe African Americans using different LD pruning window sizes. Each panel plots the absolute value (abs) of the correlation between principal components and genotypes on the y-axis versus the position along the genome on the x-axis. Panels are organized vertically according to which PC is being investigated (1, 2, 3, 4) and horizontally according to what window size was used when running LD pruning prior to PCA (*none*: no LD pruning, *prune0.5*: LD pruning with an r^2 threshold of 0.1 and window size of 0.5 Mb, *prune2*: LD pruning with an r^2 threshold of 0.1 and window size of 2 Mb, and *prune10*: LD pruning with an r^2 threshold of 0.1 and window size of 10 Mb).

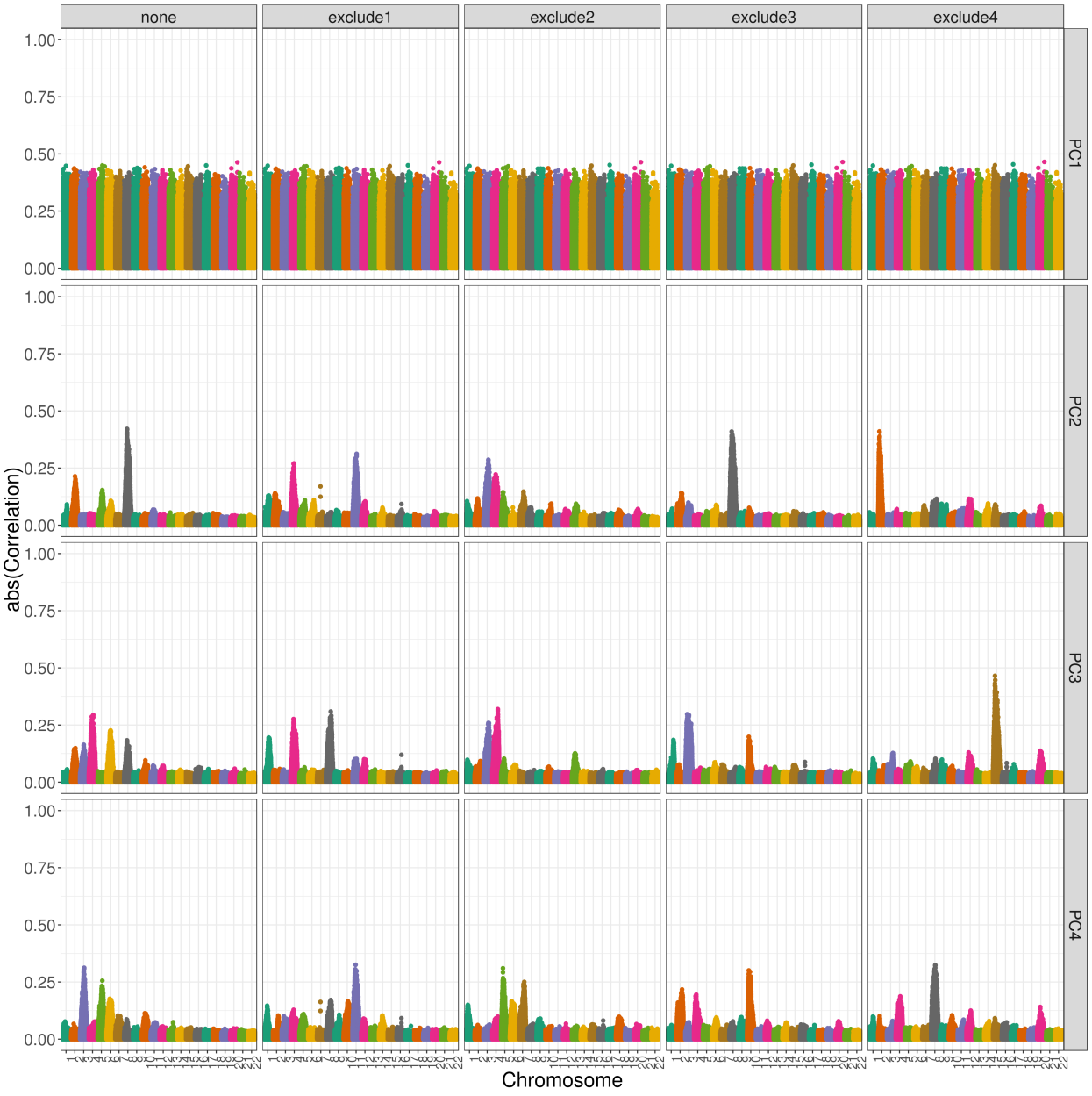

Figure S5: Correlation between PCs and genotypes in WHI SHARe African Americans after multiple rounds of data-based exclusions. Each panel plots the absolute value (abs) of the correlation between principal components and genotypes on the y-axis versus the position along the genome on the x-axis. Panels are organized vertically according to which PC is being investigated (1, 2, 3, 4) and horizontally according to the number of iterations of our procedure for excluding regions highly correlated with PCs in this sample (*none*: no exclusions, *exclude1*: one round of exclusions, *exclude2*: two rounds of exclusions, etc.).

S3 Investigation of PCs in a European American Population

We have shown that principal components can capture multiple local genomic features, rather than genome-wide ancestry, unless careful pre-processing is performed prior to running PCA. This observation is not in itself novel, but note that the patterns we observe in WHI SHARe, JHS, and COPDGene African Americans do differ slightly from what has previously been observed in European populations. In particular, in European populations a principal component might capture variation on a single chromosome[6,13] whereas in these admixed populations we see PCs driven by contributions from variants across several chromosomes. Although the focus of our work has been on admixed individuals, we were also able to run PCA on a sample of individuals with European ancestry using the COPDGene European Americans that we had excluded from our primary analyses. In this sample, we see patterns similar to those observed by previous authors, with the second and third principal components driven primarily by variants on a single chromosome: chromosome 11 (Figure S6). This difference in what is captured by principal components in European populations versus admixed populations (i.e., variants on one chromosome versus multiple) has important implications: only when a PC captures *multiple* local genomic features does the possibility of collider bias arise. Thus, particular care must be taken when performing genome-wide association studies in admixed populations to ensure that models to do not adjust for principal components that are highly correlated with variants on distinct chromosomes.

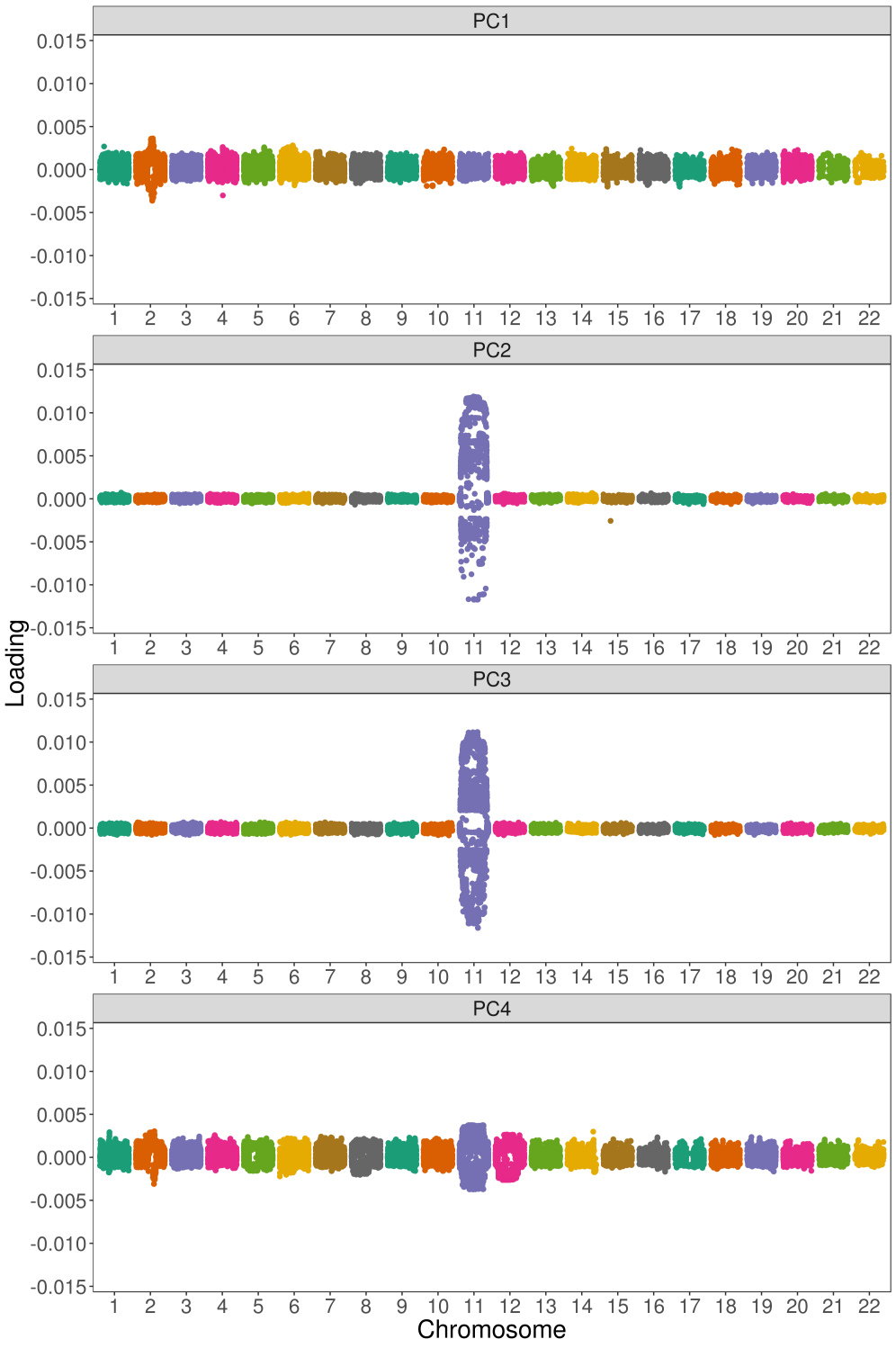

Figure S6: SNP loadings for naively generated PCs in COPDGene European Americans. Each panel plots the principal component loading (y-axis) versus the position along the genome (x-axis) for each variant. Panels are organized vertically according to which PC is being investigated (1, 2, 3, 4). Unlike in admixed populations, we see a single peak on chromosome 11.

S4 Derivation of Theoretical Results

In the main paper, we present the expected effect size estimates for GWAS models using different techniques for adjusting for ancestral heterogeneity: see Equations (1), (2), and (3). Here, we provide details and simulation studies validating these theoretical results.

S4.1 Assumed data-generating mechanism

We consider an admixed population with two ancestral populations, n individuals, and admixture proportions $\boldsymbol{\pi}_i = \begin{pmatrix} \pi_i & 1-\pi_i \end{pmatrix}^\top$ that are allowed to vary across the population. We refer to the two ancestral populations as *Ancestral Population 1* and *Ancestral Population 2*, with π_i representing the genome-wide proportion of genetic material inherited by individual i from Ancestral Population 1 and $1-\pi_i$ representing the proportion of genetic material inherited from Ancestral Population 2. We denote local ancestry by $\mathbf{a}_{ij} = \begin{pmatrix} a_{ij} & 2-a_{ij} \end{pmatrix}^\top$, where a_{ij} and $2-a_{ij}$ are the number of alleles inherited by individual i from Ancestral Populations 1 and 2, respectively, at position j. Genotypes, quantified as the number of copies of some pre-specified allele carried by individual i at position j, are represented by g_{ij}. We consider two *unlinked* variants $j = 1, 2$ (e.g., variants on distinct chromosomes) and assume that data are generated according to the following hierarchical model:

$$
\begin{aligned}
\pi_i &\overset{\text{i.i.d.}}{\sim} F \text{ for some distribution } F \\
a_{ij} \mid \pi_i &\overset{\text{i.i.d}}{\sim} \text{Binomial}(2, \pi_i),\ j = 1, 2 \\
g_{ij} \mid a_{ij}, \mathbf{p}_j &\overset{\text{ind.}}{\sim} \text{Binomial}(a_{ij}, p_{j1}) + \text{Binomial}(2 - a_{ij}, p_{j2}),\ j = 1, 2
\end{aligned}
$$

where p_{j1}, p_{j2} are allele frequencies at position j in Ancestral Populations 1 and 2, respectively. Since the two variants under consideration are unlinked, we assume that local ancestry and genotypes at these positions are conditionally independent.

We assume that our quantitative trait of interest $\mathbf{y}$ depends only on the genotype at

position 1 ($j = 1$), and we allow for the possibility that the admixture proportions $\boldsymbol{\pi}$ have a direct effect on the trait (e.g., through environmental differences across ancestral populations). More specifically, we assume that this trait is generated according to

$$y_i = \beta_0 + \beta_1 g_{i1} + \beta_\pi \pi_i + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d}}{\sim} (0, \sigma_\epsilon^2).$$

We refer to β_1 and β_2 as the true *effect sizes* of variants 1 and 2, respectively. Since the trait only depends on the genotype at position 1, the true effect size of position 2 is $\beta_2 = 0$.

Assuming that data are generated according to the above-described mechanisms, and defining $E_\pi := E(\pi)$ and $V_\pi := \text{Var}(\pi)$, then the following statements are true. For notational simplicity, we drop the subscript i.

- $E(a_j) = 2E_\pi,\ j = 1, 2$
- $V(a_j) = 2\{V_\pi + E_\pi(1 - E_\pi)\},\ j = 1, 2$
- $\text{Cov}(a_1, a_2) = 4V_\pi$
- $\text{Cov}(a_j, \pi) = 2V_\pi,\ j = 1, 2$
- $E(g_j) = 2\{p_{j2} + (p_{j1} - p_{j2})E_\pi\},\ j = 1, 2$
- $V(g_j) = 2[p_{j2}(1-p_{j2})+(p_{j1}-p_{j2})(1-p_{j1}-p_{j2})E_\pi+(p_{j1}-p_{j2})^2\{V_\pi+E_\pi(1-E_\pi)\}],\ j = 1, 2$
- $\text{Cov}(g_1, g_2) = 4(p_{11} - p_{12})(p_{21} - p_{22})V_\pi$
- $\text{Cov}(g_j, g_j) = 2(p_{j1} - p_{j2})\{V_\pi + E_\pi(1 - E\pi)\},\ j = 1, 2$
- $\text{Cov}(g_j, g_k) = 4(p_{j1} - p_{j2})V_\pi,\ j \neq k$
- $\text{Cov}(g_j, \pi) = 2(p_{j1} - p_{j2})V_\pi,\ j = 1, 2$

Furthermore, suppose we define a random variable $z_g = h(g_1, g_2) + e,\ e \sim (\mu_e, \sigma_e^2)$ for some function h. Then:

- $E(z_g) = \mu_e + E\{h(g_1, g_2)\}$
- $V(z_g) = \sigma_e^2 + V\{h(g_1, g_2)\}$
- $\text{Cov}(\pi, z_g) = \text{Cov}[\pi, E\{h(g_1, g_2) \mid \pi\}]$
- $\text{Cov}(a_j, z_g) = 2\text{Cov}(\pi, z_g) + E[\text{Cov}\{a_j, h(g_1, g_2) \mid \pi\}],\ j = 1, 2$
- $\text{Cov}(g_j, z_g) = 2(p_{j1} - p_{j2})\text{Cov}(\pi, z_x) + E[\text{Cov}\{g_j, h(g_1, g_2) \mid \pi\}],\ j = 1, 2$

These results are straightforward to derive, using our assumed hierarchical data-generating model and the laws of total expectation

$$E[x] = E\{E[x \mid y]\},$$

total variance

$$V[x] = V\{E[x \mid y]\} + E\{V[x \mid y]\},$$

and total covariance

$$\text{Cov}[x, y] = \text{Cov}\{E[x \mid z], E[y \mid z]\} + E\{\text{Cov}[x, y \mid z]\}.$$

S4.2 Expected effect size estimates

Suppose we conduct a genome-wide association study using either an *unadjusted*, *admixture proportion adjusted*, or *principal component adjusted* GWAS model. These models can be written as follows:

$$\begin{aligned}
\text{Unadjusted:}\ & E[y_i \mid g_{ij}] = \alpha + \beta_j g_{ij}, \\
\text{Admixture Proportion Adjusted:}\ & E[y_i \mid g_{ij}, \pi_i] = \alpha + \beta_j g_{ij} + \gamma \pi_i, \\
\text{Principal Component Adjusted:}\ & E[y_i \mid g_{ij}, u_{1i}, \ldots, u_{pi}] = \alpha + \beta_j g_{ij} + \gamma_1 u_{1i} + \cdots + \gamma_p u_{pi},
\end{aligned}$$

for some number of PCs p. The expected effect size estimates from these models can be derived using the theory of linear models (and a lot of algebra).

S4.2.1 Unadjusted model

If we fit an unadjusted GWAS model, as defined above, then the estimate of the effect size of the variant at position j takes the following form in expectation:

$$E[\hat{\beta}_j] = \frac{\beta_1 \widehat{\text{Cov}}(g_1, g_j) + \beta_\pi \widehat{\text{Cov}}(\pi, g_j)}{\widehat{\text{Var}}(g_j)},$$

where $\widehat{\text{Var}}$ and $\widehat{\text{Cov}}$ are the sample variance and covariance, respectively, across all n individuals in the sample (e.g., $\widehat{\text{Var}}(g_j) = \frac{1}{n-1}\sum_{i=1}^{n}(g_{ij} - \bar{g}_j)^2$).

Proof. Let $\boldsymbol{\pi}, \mathbf{a}_1, \mathbf{a}_2, \mathbf{g}_1, \mathbf{g}_2$ be drawn from the hierarchical model specified above. Assume that the trait $\mathbf{y}$ is generated such that $\mathbf{y} = \beta_0\mathbf{1} + \beta_1\mathbf{g}_1 + \beta_\pi\boldsymbol{\pi} + \boldsymbol{\epsilon}$, where ϵ_i are drawn *i.i.d.* from some distribution with mean 0 and variance σ_ϵ^2. Suppose that at position j we fit the unadjusted GWAS model $E[\mathbf{y} \mid \mathbf{g}_j] = \beta_0\mathbf{1} + \beta_j\mathbf{g}_j$. Then, the estimated regression coefficients for this model will take the form

$$\hat{\boldsymbol{\beta}}_j = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_j \end{pmatrix} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \text{ for } \mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{g}_j \end{pmatrix},$$

with expected value

$$E[\hat{\boldsymbol{\beta}}_j] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}^*\boldsymbol{\beta}, \text{ for } \mathbf{X}^* = \begin{pmatrix} \mathbf{1} & \mathbf{g}_1 & \boldsymbol{\pi} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_\pi \end{pmatrix}.$$

But

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{pmatrix} \mathbf{1}^\top\mathbf{1} & \mathbf{1}^\top\mathbf{g}_j \\ \mathbf{g}_j^\top\mathbf{1} & \mathbf{g}_j^\top\mathbf{g}_j \end{pmatrix}^{-1} = \frac{1}{n\widehat{\text{Var}}(g_j)} \begin{pmatrix} \widehat{\text{Var}}(g_j) + \hat{E}(g_j)^2 & -\hat{E}(g_j) \\ -\hat{E}(g_j) & 1 \end{pmatrix}$$

and

$$\mathbf{X}^\top \mathbf{X}^* = \begin{pmatrix} \mathbf{1}^\top\mathbf{1} & \mathbf{1}^\top\mathbf{g}_1 & \mathbf{1}^\top\boldsymbol{\pi} \\ \mathbf{g}_j^\top\mathbf{1} & \mathbf{g}_j^\top\mathbf{g}_1 & \mathbf{g}_j^\top\boldsymbol{\pi} \end{pmatrix} = n \begin{pmatrix} 1 & \hat{E}(g_1) & \hat{E}(\pi) \\ \hat{E}(g_j) & \widehat{\mathrm{Cov}}(g_1, g_j) + \hat{E}(g_1)\hat{E}(g_j) & \widehat{\mathrm{Cov}}(\pi, g_j) + \hat{E}(\pi)\hat{E}(g_j) \end{pmatrix}.$$

It follows that

$$E[\hat{\boldsymbol{\beta}}_j] = \frac{1}{\widehat{\mathrm{Var}}(g_j)} \begin{pmatrix} \widehat{\mathrm{Var}}(g_j) & \widehat{\mathrm{Var}}(g_j)\hat{E}(g_1) - \widehat{\mathrm{Cov}}(g_j, g_1)\hat{E}(g_j) & \widehat{\mathrm{Var}}(g_j)\hat{E}(\pi) - \hat{E}(g_j)\widehat{\mathrm{Cov}}(g_j, \pi) \\ 0 & \widehat{\mathrm{Cov}}(g_j, g_1) & \widehat{\mathrm{Cov}}(g_j, \pi) \end{pmatrix} \boldsymbol{\beta},$$

and thus

$$E[\hat{\beta}_j] = \frac{\beta_1 \widehat{\mathrm{Cov}}(g_j, g_1) + \beta_\pi \widehat{\mathrm{Cov}}(\pi, g_j)}{\widehat{\mathrm{Var}}(g_j)},$$

as desired. □

Using the results from Section S4.1, we can simplify this result further. We will consider studies with large sample sizes, such that we can replace the sample variance $\widehat{\mathrm{Var}}$ and covariance $\widehat{\mathrm{Cov}}$ with their population equivalent. At the causal variant (position $j = 1$), the expected effect size estimate becomes

$$\begin{aligned} E[\hat{\beta}_1] &= \frac{\beta_1 \mathrm{Cov}(g_1, g_1) + \beta_\pi \mathrm{Cov}(\pi, g_1)}{\mathrm{Var}(g_1)} \\ &= \frac{\beta_1 \mathrm{Var}(g_1) + \beta_\pi \mathrm{Cov}(\pi, g_1)}{\mathrm{Var}(g_1)} \\ &= \beta_1 + \frac{\beta_\pi V_\pi (p_{11} - p_{12})}{p_{12}(1 - p_{12}) + (p_{11} - p_{12})(1 - p_{11} - p_{12})E_\pi + (p_{11} - p_{12})^2(V_\pi + E_\pi - E_\pi^2)}, \end{aligned}$$

and at the unlinked neutral variant (position $j = 2$),

$$\begin{aligned} E[\hat{\beta}_2] &= \frac{\beta_1 \mathrm{Cov}(g_1, g_2) + \beta_\pi \mathrm{Cov}(\pi, g_2)}{\mathrm{Var}(g_2)} \\ &= \frac{(p_{21} - p_{22}) V_\pi [2\beta_1 (p_{11} - p_{12}) + \beta_\pi]}{p_{22}(1 - p_{22}) + (p_{21} - p_{22})(1 - p_{21} - p_{22})E_\pi + (p_{21} - p_{22})^2(V_\pi + E_\pi - E_\pi^2)}. \end{aligned}$$

This confirms the results presented in Equation (2) in the main paper.

S4.2.2 Admixture proportion adjusted model

Now suppose that we fit a GWAS model that adjusts for the true admixture proportions, π_i. The expected effect size estimate at variant j is:

$$E[\hat{\beta}_j] = \beta_1 \frac{\widehat{\text{Var}}(\pi)\widehat{\text{Cov}}(g_1, g_j) - \widehat{\text{Cov}}(g_1, \pi)\widehat{\text{Cov}}(g_j, \pi)}{\widehat{\text{Var}}(\pi)\widehat{\text{Var}}(g_j) - \widehat{\text{Cov}}(g_j, \pi)^2}$$

Proof. The proof of this result follows similar arguments to that for the unadjusted model, replacing the design matrix $\mathbf{X}$ with $\begin{pmatrix} \mathbf{1} & \mathbf{g}_j & \boldsymbol{\pi} \end{pmatrix}$. With a bit of algebra, the rest follows. □

Again, we can simplify this result further using the results from Section S4.1 and the assumption of large sample sizes. At the causal variant (position $j = 1$), the expected effect size estimate simplifies to

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1 \frac{\text{Var}(\pi)\text{Cov}(g_1, g_1) - \text{Cov}(g_1, \pi)\text{Cov}(g_1, \pi)}{\text{Var}(\pi)\text{Var}(g_1) - \text{Cov}(g_1, \pi)^2} \\ &= \beta_1 \frac{\text{Var}(\pi)\text{Var}(g_1) - \text{Cov}(g_1, \pi)^2}{\text{Var}(\pi)\text{Var}(g_1) - \text{Cov}(g_1, \pi)^2} \\ &= \beta_1, \end{aligned}$$

and at the unlinked neutral variant (position $j = 2$), we have

$$\begin{aligned} E[\hat{\beta}_2] &= \beta_1 \frac{\text{Var}(\pi)\text{Cov}(g_1, g_2) - \text{Cov}(g_1, \pi)\text{Cov}(g_2, \pi)}{\text{Var}(\pi)\text{Var}(g_2) - \text{Cov}(g_2, \pi)^2} \\ &= 0. \end{aligned}$$

This confirms the results presented in Equation (1) in the main paper.

S4.2.3 Principal component adjusted model

Last, we consider a model that adjusts for two principal components, $\mathbf{u}_1, \mathbf{u}_2$, supposing that the first PC captures global ancestry (i.e., $u_{i1} = \pi_i \ \forall i$) and the second captures some other

feature quantified by a random variable $\mathbf{z}$ (i.e., $u_{i2} = z_i \; \forall i$). The expected effect size estimate at variant j is:

$$E[\hat{\beta}_j] = \beta_1 \frac{V_z(V_\pi C_{g_1,g_j} - C_{g_1,\pi}C_{g_j,\pi}) - V_\pi C_{g_1,z}C_{g_j,z} + C_{\pi,z}(C_{g_1,\pi}C_{g_j,z} + C_{g_1,z}C_{g_j,\pi} - C_{g_1,g}C_{\pi,z})}{V_z(V_\pi V_{g_j} - C^2_{g_j,\pi}) - V_\pi C^2_{g_j,z} + C_{\pi,z}(2C_{g_j,\pi}C_{g_j,z} - V_{g_j}C_{\pi,z})},$$

where $V_a = \widehat{\text{Var}}(a)$ and $C_{a,b} = \widehat{\text{Cov}}(a,b)$.

Proof. Again, this proof follows from similar arguments to that for the unadjusted model, now replacing the design matrix $\mathbf{X}$ with $\begin{pmatrix} \mathbf{1} & \mathbf{g}_j & \boldsymbol{\pi} & \mathbf{z} \end{pmatrix}$. After making this substitution, the rest follows. □

Making the same assumptions as above, we first simplify these results considering a general form of $\mathbf{z}$. At the causal variant ($j = 1$),

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1 \frac{V_z(V_\pi C_{g_1,g_1} - C_{g_1,\pi}C_{g_1,\pi}) - V_\pi C_{g_1,z}C_{g_1,z} + C_{\pi,z}(C_{g_1,\pi}C_{g_1,z} + C_{g_1,z}C_{g_1,\pi} - C_{g_1,g_1}C_{\pi,z})}{V_z(V_\pi V_{g_1} - C^2_{g_1,\pi}) - V_\pi C^2_{g_1,z} + C_{\pi,z}(2C_{g_1,\pi}C_{g_1,z} - V_{g_1}C_{\pi,z})} \\ &= \beta_1 \frac{V_z(V_\pi V_{g_1} - C^2_{g_1,\pi}) - V_\pi C^2_{g_1,z} + C_{\pi,z}(2C_{g_1,\pi}C_{g_1,z} - V_{g_1}C_{\pi,z})}{V_z(V_\pi V_{g_1} - C^2_{g_1,\pi}) - V_\pi C^2_{g_1,z} + C_{\pi,z}(2C_{g_1,\pi}C_{g_1,z} - V_{g_1}C_{\pi,z})} \\ &= \beta_1, \end{aligned}$$

and at the unlinked neutral variant (position $j = 2$),

$$\begin{aligned} E[\hat{\beta}_2] &= \beta_1 \frac{V_z(V_\pi C_{g_1,g_2} - C_{g_1,\pi}C_{g_2,\pi}) - V_\pi C_{g_1,z}C_{g_2,z} + C_{\pi,z}(C_{g_1,\pi}C_{g_2,z} + C_{g_1,z}C_{g_2,\pi} - C_{g_1,g_2}C_{\pi,z})}{V_z(V_\pi V_{g_2} - C^2_{g_2,\pi}) - V_\pi C^2_{g_2,z} + C_{\pi,z}(2C_{g_2,\pi}C_{g_2,z} - V_{g_2}C_{\pi,z})} \\ &= \beta_1 \frac{-V_\pi E[\text{Cov}(g_1, z \mid \pi)]E[\text{Cov}(g_2, z \mid \pi)]}{V_z(V_\pi V_{g_2} - C^2_{g_2,\pi}) - V_\pi C^2_{g_2,z} + C_{\pi,z}(2C_{g_2,\pi}C_{g_2,z} - V_{g_2}C_{\pi,z})}. \end{aligned}$$

This confirms the results presented in Equation (3) in the main paper.

Now, suppose we make an additional assumption about the form of the second principal component. In particular, assume that $\mathbf{z} = \mathbf{z}_g = z_1\mathbf{g}_1 + z_2\mathbf{g}_2 + \mathbf{e}, \; \mathbf{e} \sim (\mu_e, \sigma^2_e)$ for some scalars z_1, z_2. In other words, the 2nd PC captures genotypes at two variants, one of which is the causal variant ($j = 1$) and the other is an unlinked neutral variant ($j = 2$). Then, at

the causal variant, the expected effect size estimate remains

$$E[\hat{\beta}_1] = \beta_1,$$

and at the unlinked neutral variant, we now have

$$E[\hat{\beta}_2] = \beta_1 \frac{-4z_1 z_2 V_\pi}{V_z(V_\pi V_{x_2} - C^2_{x_2,\pi}) - V_\pi C^2_{x_2,z} + C_{\pi,z}(2C_{x_2,\pi}C_{x_2,z} - V_{x_2}C_{\pi,z})} \times \prod_{j=1}^{2}[p_{j2}(1-p_{j2}) + (p_{j1}-p_{j2})(1-p_{j1}-p_{j2})E_\pi + (p_{j1}-p_{j2})^2(E_\pi - E^2_\pi - V_\pi)].$$

Here we can see more directly that the magnitude of the bias at the unlinked neutral variant ($j = 2$) depends on the effect size of the causal variant (β_1), the variability of admixture proportions in the population (V_π), and the strength of the contribution of the causal variant and neutral variant to the principal component (z_1 and z_2, respectively).

S4.3 Simulations validating theory

To support these theoretical results, we performed a small simulation study. We generated data according to the data generating mechanism described in Section S4.1, and then we compared the observed effect size estimates from GWAS models fit to these simulated data to the expected effect size estimates derived in Section S4.2.

We considered a variety of simulation settings, but present results from just a single setting here. Admixture proportions for $n = 5000$ individuals were generated from the distribution $F = \text{Beta}(7, 2)$ (see Figure S7), allele frequencies at the causal variant were 0.7 (ancestral population 1) and 0.2 (ancestral population 2), allele frequencies at the neutral variant were 0.7 (ancestral population 1) and 0.2 (ancestral population 2), admixture proportions did not have a direct effect on the trait ($\beta_\pi = 0$), and the 2nd PC was generated according to $z_i = z_1 g_{i1} + z_2 g_{i2} + e_i$ for a range of values z_1, z_2 and added noise $e_i \overset{iid}{\sim} N(0, 0.25^2)$.

In Figures S8, S9, and S10 we plot the effect size estimates from GWAS models that

Figure S7: Barplot of simulated admixture proportions.

we observe in our simulation study, and compare these observed effect size estimates to the expected effect sizes based on our analytic results (Section S4.2), as well as the true effect sizes based on the data-generating mechanism (Section S4.1). For all three types of models, we see a perfect correspondence between the observed effect sizes and the expected effect sizes provided in Section S4.2, validating our theoretical derivations.

Comparing the observed and expected effect sizes to the true effect sizes provides insight into the magnitude of bias that can be expected from each model. Figure S8 presents results using the unadjusted GWAS model. We see a departure between the true effect size (0) and the observed and expected effect sizes at the unlinked neutral variant (SNP 2), confirming that models that fail to adjust for ancestral heterogeneity can yield biased estimates of the effect size even when global ancestry does not have a direct effect on the trait ($\beta_\pi = 0$).

In Figure S9, we consider the case of a GWAS model that adjusts for ancestral heterogeneity by including the true admixture proportions as a covariate. In this case, we see a perfect correspondence between the observed, expected, and true effect sizes for both the causal and unlinked neutral variants. This confirms, as our theory suggests, that models that appropriately adjust for ancestral heterogeneity will yield unbiased estimates of variant effect sizes (i.e., $E[\hat{\beta}_j] = \beta_j$).

Figure S10 illustrates the impact of adjusting for a principal component that captures

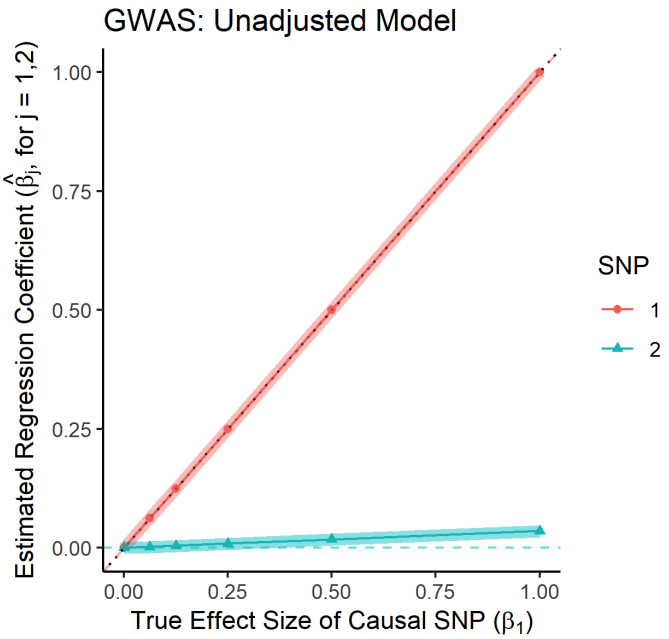

Figure S8: Comparison of observed, expected, and true effect sizes from unadjusted GWAS models applied to simulated data. Observed effect sizes are represented by the thin solid lines with points (red with dots = SNP 1, blue with triangles = SNP 2). Expected effect sizes from Section S4.2 are represented by the wider and faintly colored solid lines (red = SNP 1, blue = SNP 2). True effect sizes are represented by the dashed lines (red = SNP 1, blue = SNP 2). The $y = x$ line is also provided for reference (dotted black line).

local genomic features (in this case, the genotype of variants 1 and 2) instead of global ancestry. Specifically, the second PC was generated according to the equation $z_1 g_{i1} + z_2 g_{i2} + N(0, 0.25^2)$ for scalars z_1, z_2. Although the effect sizes estimates for variant 1 are unbiased across all simulation settings, confirming our theoretical result that $E[\hat{\beta}_1] = \beta_1$ (see Section S4.2), we see that the observed and expected effect sizes for SNP 2 often deviate from the truth. The only situations in which effect size estimates for this neutral unlinked variant are unbiased (i.e., $E[\hat{\beta}_2] = \beta_2 = 0$) are when the PC is not actually affected by the causal variant (i.e., $z_1 = 0$) or when the PC is not affected by the neutral variant (i.e., $z_2 = 0$). In other words, as long as the extraneous PC captures the genotypes of both the causal variant and the unlinked neutral variant, we see that effect sizes are biased away from zero at this neutral variant. The magnitude of this bias increases with the strength of the relationship between the variants and the PC (i.e., as we increase z_1 and/or z_2).

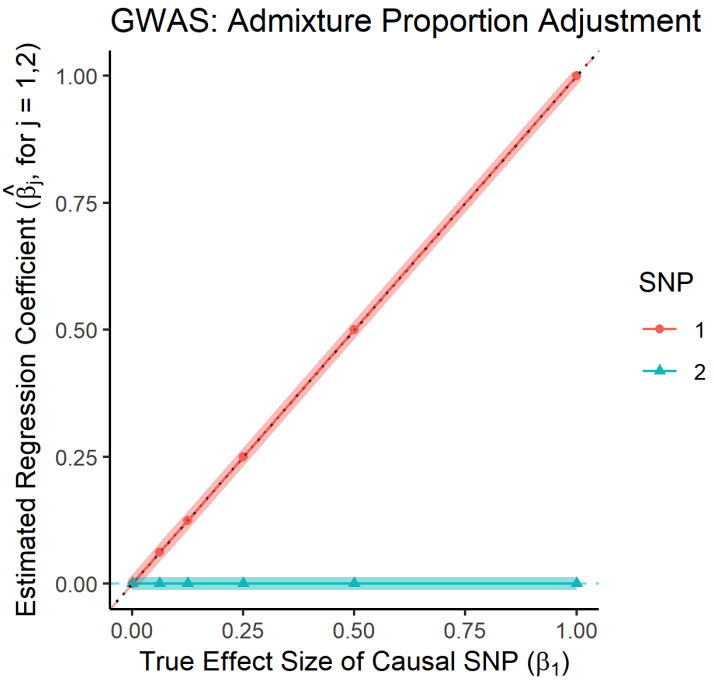

Figure S9: Comparison of observed, expected, and true effect sizes from GWAS models adjusting for admixture proportions in simulated data. Observed effect sizes are represented by the thin solid lines with points (red with dots = SNP 1, blue with triangles = SNP 2). Expected effect sizes from Section S4.2 are represented by the wider and faintly colored solid lines (red = SNP 1, blue = SNP 2). True effect sizes are represented by the dashed lines (red = SNP 1, blue = SNP 2). The $y = x$ line is also provided for reference (dotted black line).

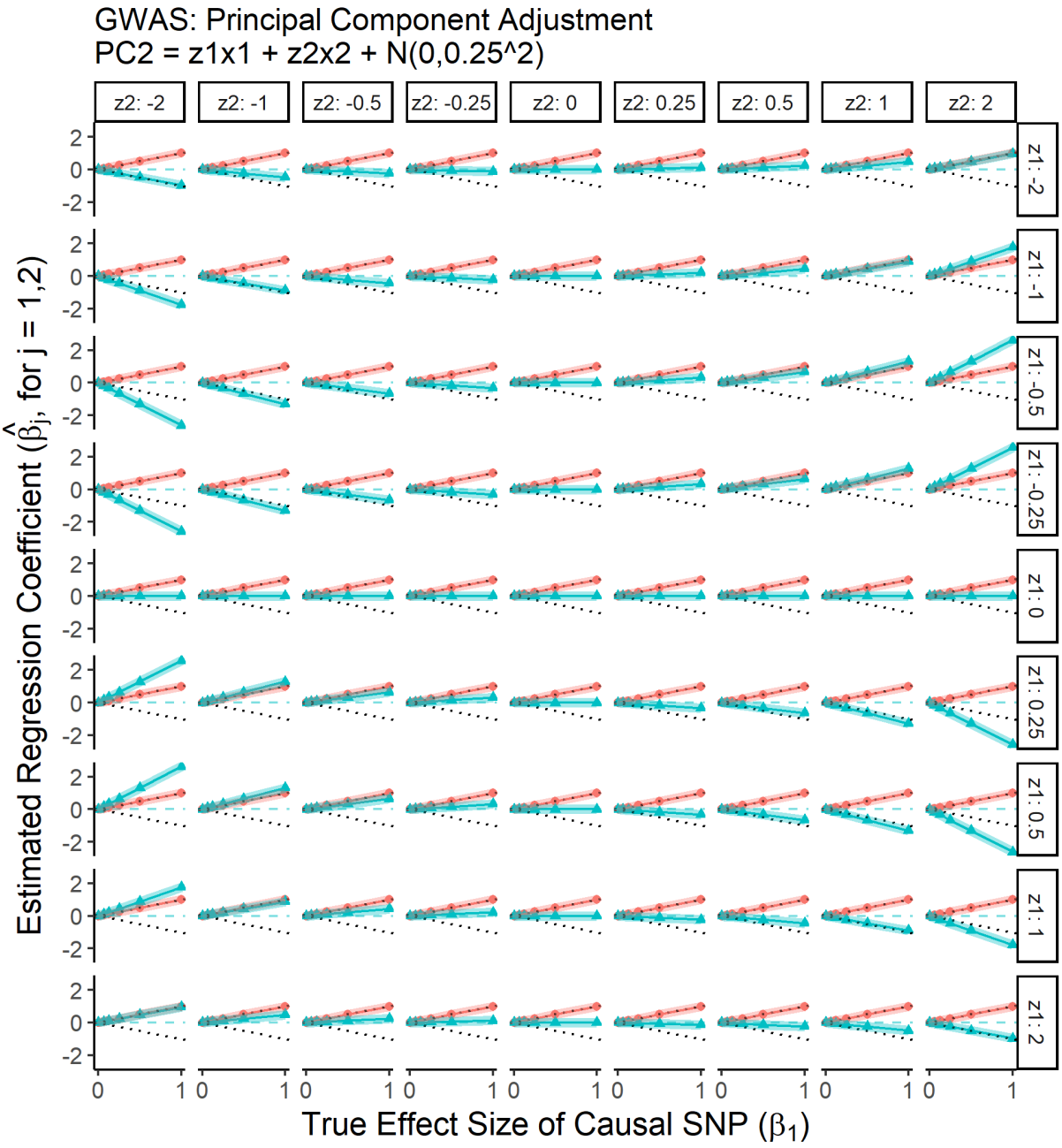

Figure S10: Comparison of observed, expected, and true effect sizes from GWAS models adjusting for two principal components. The first PC captures global ancestry, but the second PC was generated according to the equation $z_1 g_{i1} + z_2 g_{i2} + N(0, 0.25^2)$, changing the scalars z_1, z_2 in each panel. Observed effect sizes are represented by the thin solid lines with points (red with dots = SNP 1, blue with triangles = SNP 2). Expected effect sizes are represented by the wider and faintly colored solid lines (red = SNP 1, blue = SNP 2). True effect sizes are represented by the dashed lines (red = SNP 1, blue = SNP 2). The $y = x$ line is also provided for reference (dotted black line).

S5 Additional Simulations using TOPMed Data

We also performed simulations using whole genome sequence data from the Trans-Omics for Precision Medicine Project to further connect our theoretical results to the findings, presented in the main paper, from our simulation study using WHI SHARe genotype data. For each individual, we simulated a quantitative trait that depended only on their genotype at a single variant on chromosome 8. We also constructed two "fake" principal components. The first PC was set equal to the estimated African admixture proportion (see Methods). The second PC was generated such that it depended on the genotype at this same causal variant, as well as another variant on chromosome 6. Figure S11 shows that we see similar patterns in the correlation between this PC and genotypes as we did with real PCs (calculated without strict LD-based pruning) in WHI SHARe, TOPMed JHS, and TOPMed COPDGene African Americans. These PC-genotype correlation plots clearly show—as we know to be true, by design—that the second PC is driven by variants on two chromosomes (6 and 8) rather than detecting genome-wide ancestry.

We then investigated the impact of including this extraneous PC in GWAS models. Our results again mirror the patterns observed in our WHI SHARe simulations. Figure S12 presents Manhattan plots from a single simulation replicate, comparing results from a model that adjusted for just the first principal component (top panel) versus a model that adjusted for both PCs (bottom panel). In both cases, we see a genome-wide significant association on chromosome 8 — the location of the true causal variant. However, in the case of the model adjusting for two PCs, we also see a spurious association on chromosome 6 — the location of the second variant that contributes to the second principal component. The second PC plays the role of a collider variable in this setting, and adjusting for it has induced a spurious association. It is worth noting that the quantile-quantile (QQ) plots and inflation factors for these two analyses are indistinguishable (Figure S13), so those tools alone are not sufficient for detecting this issue of collider bias caused by including PCs that capture multiple local genomic features, rather than genetic ancestry.

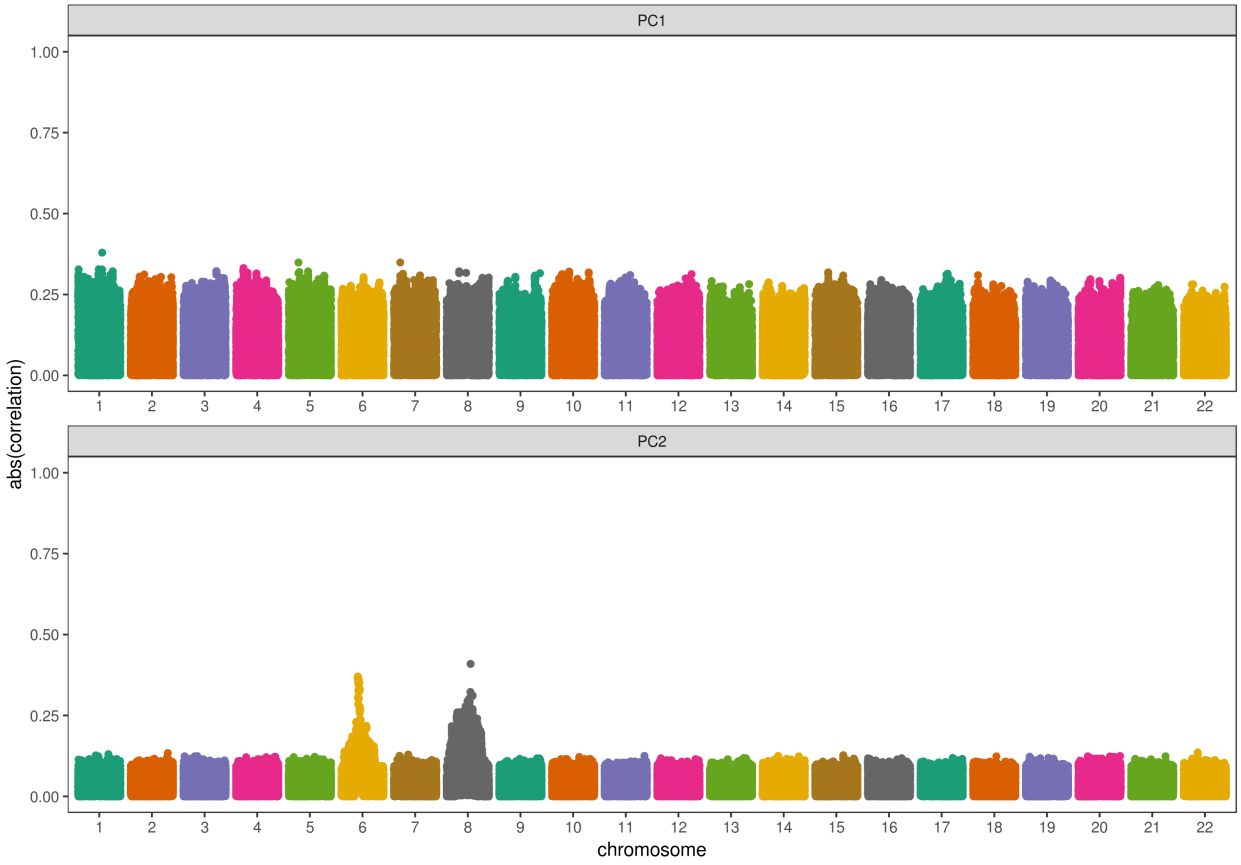

Figure S11: Correlation between fake PCs (i.e., PCs that were constructed such that the first captures genetic ancestry but the second captures genotype at two variants on chromosomes 6 and 8) in TOPMed JHS African Americans. Each panel plots the absolute value of the correlation between principal components 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). Peaks in this plot indicate that a variant has a larger *loading*, i.e., a larger contribution to that principal component.

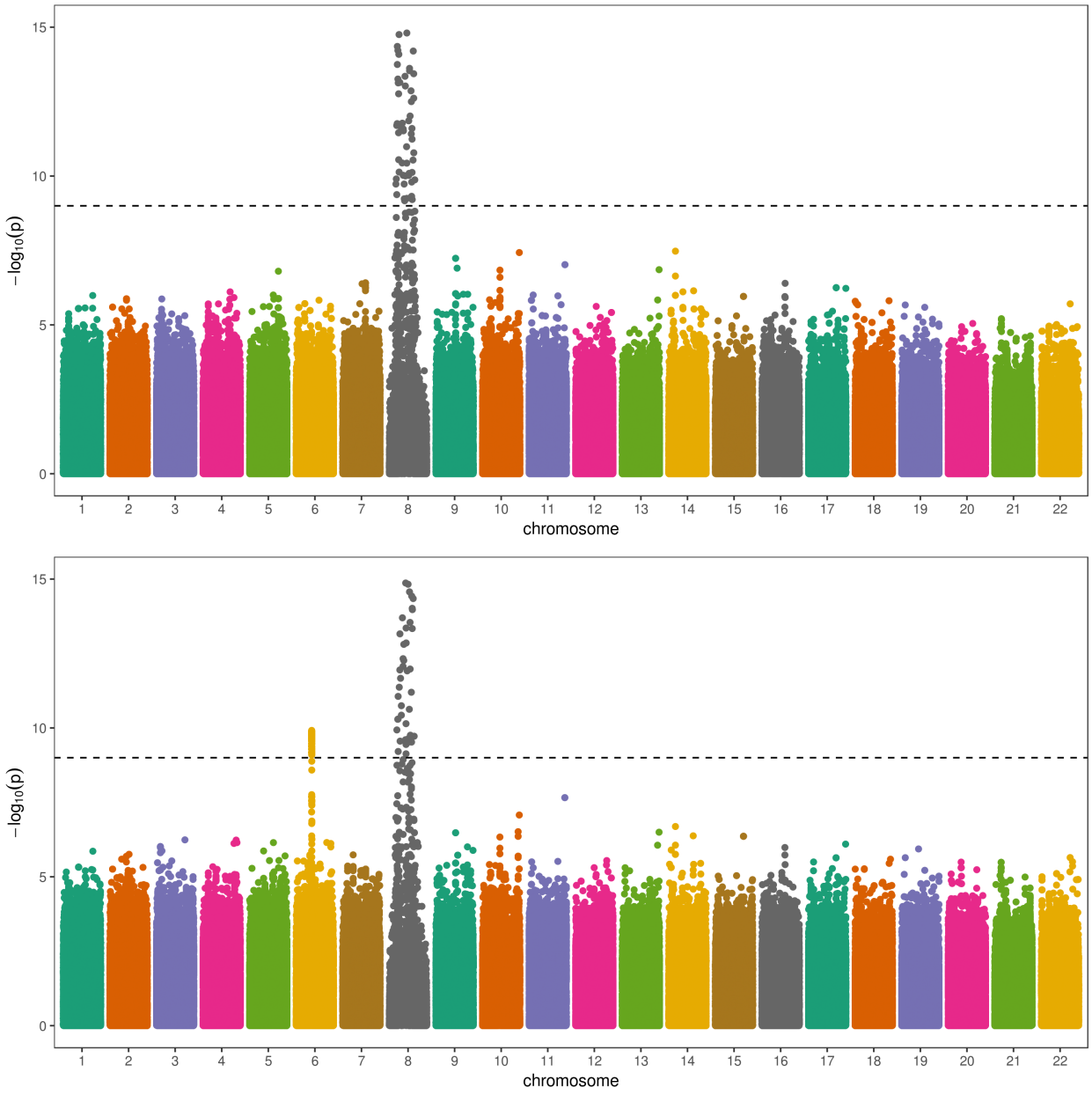

Figure S12: Manhattan plots for GWAS models adjusting for fake PCs (i.e., PCs that were constructed such that the first captures genetic ancestry but the second captures genotype at two variants on chromosomes 6 and 8) in TOPMed JHS African Americans. The top panel presents results from a model adjusting for only the first PC and the bottom panel presents results from a model adjusting for two PCs. In this simulation setting, there is only one causal variant, located on chromosome 8.

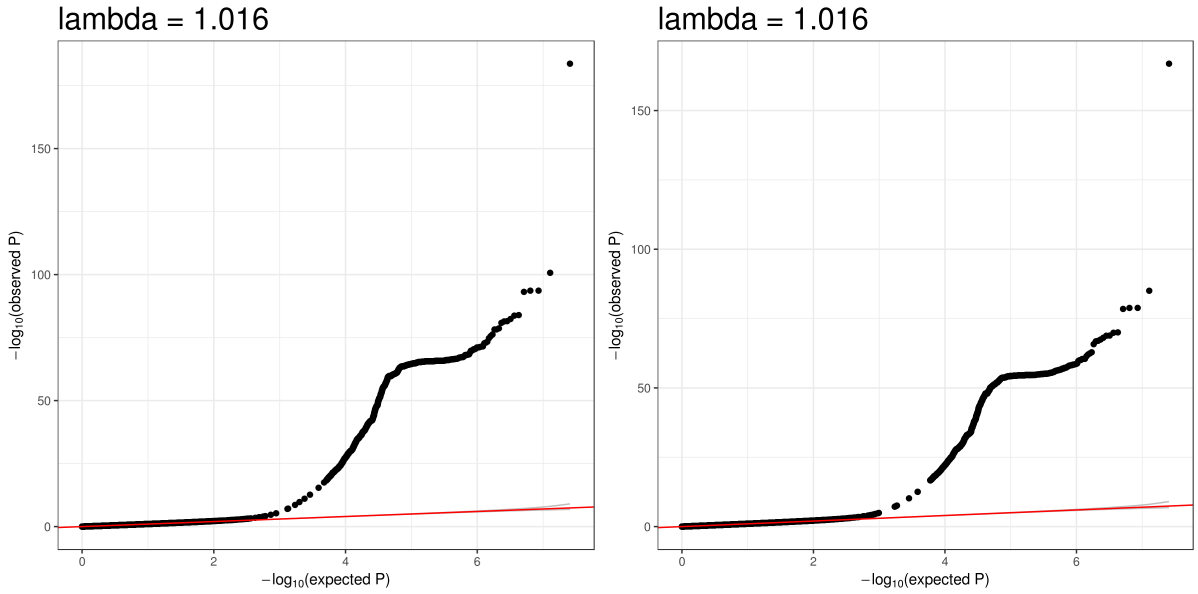

Figure S13: Quantile-quantile (QQ) plots and inflation factors (λ) for GWAS models adjusting for fake PCs (i.e., PCs that were constructed such that the first captures genetic ancestry but the second captures genotype at two variants on chromosomes 6 and 8) in TOPMed JHS African Americans. The left panel presents results from a model adjusting for only the first PC and the right panel presents results from a model adjusting for two PCs. The two plots are indistinguishable.

Supplemental References

[1] 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.

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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.