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] Sharon R. Browning[3]

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. Department of Biostatistics, University of Washington, Seattle, WA, 98195, USA

* kgrinde@macalester.edu

Abstract

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[1,2,3,4,5]. 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[6,7,8,9]. 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)[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 (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.

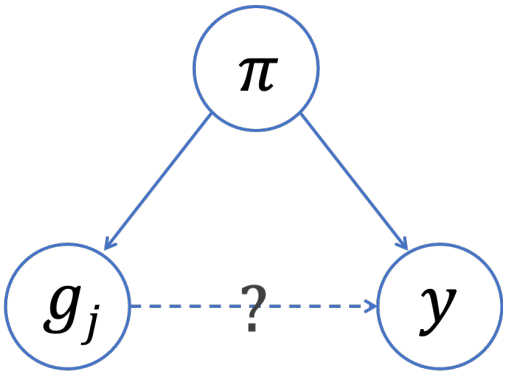

Figure 1: Global ancestry ($\boldsymbol{\pi}$) confounds the association between the genotype at position j ($\mathbf{g}_j$) and the phenotype of interest ($\mathbf{y}$) if ancestry is associated with both the genotype (e.g., the allele frequencies differ across the ancestral populations) and the phenotype (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[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 currently used extensively throughout the literature, with global ancestry inferred using either model-based ancestry inference methods (e.g., `frappe`[21], `STRUCTURE`[22], `ADMIXTURE`[23]) or principal component analysis (e.g., `EIGENSTRAT`[7], `SNPRelate`[24], `PC-AiR`[25]).

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[26]. To infer global ancestry using PCA, we preform 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), $\mathbf{u}_1, \mathbf{u}_2, \ldots$ tend to reflect global ancestry[27,28]. 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[27,7], examining inflation factors[29,5] and/or the proportion of variance explained by each PC[30,29,5], comparing PCs to self-reported race/ethnicity[5], and keeping PCs that are significantly associated with the trait[31,32]. Typically, the number of PCs selected is on the order of 1–10[33], but in practice it is not uncommon to see applications in which more PCs are used 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" [34] at the cost of only "subtle" decreases in power [35].

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 [27,9,36,25], array artifacts or other data quality issues [27,7,9,37], and/or small regions of the genome with unusual patterns of linkage disequilibrium (LD) [27,7,38,39,40,9,37,41,42,36,43]. 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` [44] to remove variants that are in "high" LD (e.g., pairwise-correlation $r^2 > 0.2$) with nearby variants [38,45,26,46,47,48,37,42,36,49,25,29,50,5,32], and/or excluding regions of the genome that are known to have extensive, long-ranging, or otherwise unusual patterns of LD [38,45,26,40,48,37,30,5]. 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. To conclude, we provide

Chr	Start (bp)	End (bp)	References
chr1	48000000	52060567	48,40,37
chr2	85941853	100500000	48,40,37
chr2	129600000	140000000	40,26,37,30,5,51
chr2	182882739	190000000	48,40,37
chr3	47500000	50000000	48,40,37
chr3	83500000	87000000	48,40,37
chr3	89000000	97500000	40,37
chr3	163100000	164900000	51
chr5	44000000	51500000	45,48,40,37
chr5	98000000	100500000	40,37
chr5	129000000	132000000	48,40,37
chr5	135500000	138500000	40,37
chr6	23800000	39000000	45,48,40,26,37,30,5,51
chr6	57000000	64000000	48,40,37
chr6	140000000	142500000	48,40,37
chr7	55000000	66193285	48,40,37
chr8	6300000	13500000	45,48,40,26,39,37,30,5,51
chr8	43000000	50000000	48,40,37
chr8	112000000	115000000	48,40,37
chr10	37000000	43000000	48,40,37
chr11	45000000	57000000	45,40,37
chr11	87500000	90500000	48,40,37
chr12	33000000	40000000	48,40,37
chr12	109500000	112021663	40,37
chr14	46600000	47500000	51
chr17	37800000	42000000	26,5
chr20	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/.

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

2 Material and Methods

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 focuses in particular on 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 data include 3,406 African American individuals and the freeze 5b COPDGene data include 8,742 African American and European American individuals.

2.2 TOPMed Sequencing and Quality Control

For TOPMed freezes 4 and 5b, high coverage ($\approx$ 30X) whole genome sequencing was performed by several sequencing centers, with, for the most part, all samples for a given study sequenced at the same center. Variant discovery, and genotype calling was performed by the TOPMed Informatics Resources Center (IRC) using the `GotCloud` pipeline[52]. Quality control (QC) was performed by a combination of the sequencing centers, IRC, and TOPMed Data Coordinating Center (DCC), and only those samples and variants that passed this QC are included in the VCF files that can be downloaded from dbGaP. De-

tails on TOPMed QC methods are available in Taliun et al.[53] and on the TOPMed website: https://www.nhlbiwgs.org/data-sets.

2.3 Additional Filtering and Quality Control

After downloading the JHS and COPDGene data from dbGaP, we performed additional rounds of filtering/QC before proceeding with further analyses.

New outline:

- TOPMed data
- QC
  - rare variants [... move this to QC step? ...]; [... citations ...]
  - missing rates [... STILL NEED TO IMPLEMENT THIS!! ...]
  - relatives; [... citations ...]
  - non-admixed individuals
- inferring ancestry using PCA
  - software
  - types of pruning/filtering considered
  - plots we look at (loadings, screeplots, parallel coordinates, etc.)
- inferring ancestry using ADMIXTURE
  - motivation for comparison
  - recommended pruning/filtering
- simulation study
  - traits

  - models
  - evaluation
- software and data availability
  - dbgap
  - github

2.4 Adjusting for ancestral heterogeneity in genome-wide association studies

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 ($j = 1, \ldots, m$):

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

where $\mathbf{z}_i$ is a vector of additional covariates (e.g., potential confounding variables) that we want to include in the model. This linear regression model can be replaced with a logistic regression model in the case of a binary trait (e.g., disease status). In either case, we test for an association between the trait and genotype by testing the null hypothesis $H_0 : \beta_j = 0$ at each position $j = 1, \ldots, m$.

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 of principal component analysis.

2.5 Simulation study using TOPMed whole genome sequence data

Are we sure we want to use TOPMed? Or should we switch back to WHI?

We implement a simulation study using whole genome sequence data from the Trans-Omics for Precision Medicine (TOPMed) program to compare different approaches for adjusting for ancestral heterogeneity and explore the impact of different variant-level filtering choices, particular with respect to linkage disequlibrium.

2.5.1 TOPMed whole genome sequence data

TOPMed is [... describe TOPMed ...]. Whole genome sequence data for contributing TOPMed studies is available on dbGaP. We focus on two such studies, the Jackson Heart Study (JHS) (accession number: phs000964) and the Chronic Obstructive Pulmonary Disease Genetic Epidemiology Study (COPDGene) (accession number: phs000951).

- describe TOPMed
- describe sequencing methods
- how many samples in each study

2.5.2 Quality control

We process the sequence data for quality control, keeping only those variants that are bi-allelic, have a minor allele count of at least one, and pass standard variant filters (Mendelian or duplicate genotype discordance $< 3/5\%$, Hardy-Weinberg Equilibrium p-value $> 1 \times 10^{-6}$, etc.). [... Add filter for missing rates! ...] After all exclusions, we are left with ??? and ??? variants in JHS and COPDGene, respectively.

After variant-level filtering, we then use the iterative procedure suggested by Conomos et al.[5] and implemented in the TOPMed analysis pipeline to identify a subset of ?? mutually

unrelated individuals. [... more details about iterative procedure; e.g., sing a kinship threshold of 0.044 ← double-check (i.e., excluding first, second, and third degree relatives) ...] Next, we run an unsupervised `ADMIXTURE` analysis with $K = 2$ and $K = 3$ and plot inferred global ancestry proportions to identify admixed (African American) and non-admixed (European) individuals. [... more details about thresholds used ...] After sample-level filtering, ??? and ??? individuals remain in JHS and COPDGene, respectively.

QC for JHS (dbgap accession phs000964):

- filtering
  - bi-allelic SNPs
  - minor allele count at least 1
  - pass variant filters (in VCFs that were downloaded from dbgap): overlaps with SNP, overlaps with indel, overlaps with VNTR, failed SVM filter, high (3/5% or more) mendelian or duplicate genotype discordance, excess heterozygosity with HWE p-value ¡ 1e-6
- merge the two subsets (cg1 and cg3)
- convert from VCF to GDS
- remove close relatives
  - run king (LD r threshold: 0.32, LD window size: 10, MAF threshold: 0.01, exclude PCA corr: TRUE, build: hg19)
  - run PC-AiR
  - run PCRelate
  - run PC-AiR again
  - run PCRelate again

- find African Americans
  - run stricter LD pruning (MAF = 0.01, window size = 0.5, rsq = 0.01, regions = TRUE, build 37)
    - List of regions stored here: `/projects/browning/brwnlab/kelsey/spurious_assoc/highl`
  - convert GDS to BED
  - run ADMIXTURE (K = 2 and K = 3)
  - plot proportions
  - exclude 40 people inferred to be 100% European; left with 1888

QC for COPDGene (phs000951):

- filtering
  - bi-allelic SNPs
  - minor allele count at least 1
  - pass filtering (from GDS annotation info, I inferred this to include: variant located in centromeric region, variant failed SVM filter, mendelian or duplicate genotype discordance is high (3/5% or more), excess heterozygosity in chrX in males, excess heterozygosity with HWE p-value ¡ 1e-6)
- convert from VCF to GDS
- remove close relatives
  - run king (LD r threshod = 0.32, LD window size = 10, MAF threshodl = 0.01, exclude PCA corr = TRUE, build = hg38); regions = see table below
  - run PCAiR
  - run PCRelate

name	chrom	start.base	end.base	comment
2q21	2	129883530	140283530	LCT
HLA	6	24092021	38892022	includes MHC
8p23	8	6612592	13455629	inversion
17q21.31	17	40546474	44644684	inversion

Table 2: TOPMed hg19 high corr regions

name	chrom	start.base	end.base	comment
2q21	2	129125957	139525961	LCT
HLA	6	24091793	38924246	includes MHC
8p23	8	6755071	13598120	inversion
17q21.31	17	42394456	46567318	inversion

Table 3: TOPMed hg38 high corr regions

  - run PCAiR again

- find African Americans
  - run stricter LD pruning (exclude PCA corr regions = TRUE, build = hg38, LD R threshold = 0.1, LD window size = 10, MAF threshold = 0.01); regions = `/projects/browning/brwnlab/kelsey/spurious_assoc/highLD_regions/`
  - convert from GDS to PLINK
  - run ADMIXTURE with K = 2 and K = 3
  - plot proportions and use cut-off of 30% to identify (and then remove) Europeans : Parker et al. 2014 "Admixture mapping identifies a quantitative trait locus associated with..." `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4190160/` (reduced from 8406 to 2676)

2.5.3 Genetic ancestry inference

- ADMIXTURE
  - JHS: ran with both K = 2 and K = 3
  - COPDGene: ran with both K = 2 and K = 3

  - unsupervised for both

- PCA using SNPRelate

- 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)

2.5.4 Evaluating population structure adjustment approaches

- plot PCs vs global ancestry proportions
- plot SNP loadings
- simulating traits (effect sizes, choice of causal SNPs)
  - find loading peaks from "naive" approach
  - simulate trait that is beta * x + rnorm(0, 1), where beta = 1 or 2 and x = genotype at one of the peaks
- running GWAS
  - for each of 188*2 simulated phenotypes
  - for each set of PCs
  - including 1, 4, or 10 PCs
- defining spurious associations

3 Results

3.1 Ancestral heterogeneity in TOPMed African American samples

- quickly summarize ancestral heterogeneity (barplots of ADMIXTURE proportions)

3.2 Confirming the importance of adjusting for population structure

- show an example manhattan plot with no adjustment
- compare average number of spurious associations
- tie in theoretical results

3.3 Comparing different approaches for adjusting for population structure

Part 1: how does FWER compare?

- manhattan plots for one or two simulated traits
- overall summary of rejection rates
- is it appropriate to use same significance threshold for all?

Part 2: how does rate of spurious associations compare? (and alpha-adjusted spurious assoc?)

- manhattan plots for one or two traits
- overall summary of rejection rates

Part 3: why is this happening?

- are admixture proportions and PCs capturing similar information?
  - correlation between PCs and admixture proportions (PC1 highly correlated with admix prop)
  - correlation between PCs and genotypes (without pruning, later PCs highly correlated with genotypes in small regions)
- mathematical results

4 Discussion

Global ancestry = confounder

- Summarize conditions under which global ancestry is a confounder
- Relate to current understanding in literature

Be careful with PCs!

- Summarize conditions under which PCs can be problematic
- Relate to current understanding in literature (what have others shown can happen if you include a PC that captures local LD? how are things different in admixed populations, where LD is more extensive?)
- Relate to concept of collider bias
- Suggested diagnostics

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 Vilhjálmsson, 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.

Figure Titles and Legends

Tables