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

Admixed populations such as African Americans and Hispanics/Latinos have historically been vastly underrepresented in genome-wide association studies (GWAS)[1,2,3,4,5,6]. Although this underrepresentation has many causes, some authors have cited the statistical challenges posed by ancestrally heterogeneous populations as a possible contributing factor[1,2,3]. [... define global ancestry ...] Considerable variability of inferred global ancestry, or *ancestral heterogeneity*, has been observed in many studies of African American and Hispanic/Latino populations[7,8,9,10,11]. It has been widely documented that heterogeneous global ancestry, along with other types of population structure, can lead to spurious associations in genome-wide association studies[12,13,14,15]. These spurious associations arise due to the fact that global ancestry can confound the association between genotypes and a phenotype of interest, particularly when genetic variants are more frequent in some ancestral populations than in others and global ancestry has an effect on the trait through, for example, environmental differences across ancestral groups.

A number of methods for detecting and controlling for ancestral heterogeneity in genetic association studies have been proposed. Early approaches included restricting analyses to subsets of ancestrally homogeneous individuals[16], performing a genome-wide correction for test statistic inflation due to ancestral heterogeneity via *genomic control*[12], and using family-based designs[17]. More recently, approaches based on mixed models have been proposed[18,19,20]. These mixed model approaches use 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. However, 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[13,21]. 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., `ADMIXTURE`[22]) or unsupervised dimension reduction techniques (e.g., principal component analysis (PCA)[13]).

Model-based approaches for global ancestry inference model the probability of observed genotypes given unobserved ancestry and allele frequencies in each ancestral population[23,24,22,25]. Most often, these approaches are used to estimate *admixture proportions* $\hat{\boldsymbol{\pi}}_i = \begin{pmatrix} \hat{\pi}_{i1} & \dots & \pi_{iK} \end{pmatrix}^\top$ for individuals $i = 1, \dots, n$, where $\hat{\pi}_{ik}$ is the estimated proportion of genetic material inherited by individual i from ancestral population k. Once estimated, $\hat{\boldsymbol{\pi}}$ can then easily be included as a covariate in GWAS models to adjust for ancestral heterogeneity. One of the challenges of using these model-based approaches to infer global ancestry is that the number of ancestral populations, K, usually needs to be pre-specified. In addition, some of these model-based approaches are *supervised*, requiring reference panel data from each ancestral population of interest to estimate allele frequencies. Furthermore, ancestry inference is typically conducted at a continental level (e.g., African versus European, rather than South European versus North European), so finer levels of population structure could be missed; recent efforts have considered global ancestry inference on a sub-continental scale[25,26].

Principal component analysis (PCA), on the other hand, 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[27]. To infer global ancestry using PCA, we preform an eigenvalue decomposition of the genetic relationship matrix (GRM) $\hat{\boldsymbol{\Psi}} = \frac{1}{m}\mathbf{X}\mathbf{X}^\top$, where $\mathbf{X}$ is the $n \times m$ matrix of standardized gentoypes for n individuals at m single nucleotide variants (SNVs). The top eigenvectors, or *principal components* (PCs) of $\hat{\boldsymbol{\Psi}}$ tend to reflect global ancestry[28,29], so adjusting for PCs can be an effective approach for controlling for ancestral heterogeneity in genetic association studies[13]. In practice, however, determining the number of PCs needed to capture global ancestry can be difficult. Furthermore, it has been shown that PCs can sometimes capture features other than global ancestry, such as relatedness across samples[28,30], data quality issues[13,31], and/or small regions of the genome with unusual patterns of linkage disequilibrium (LD)[32,33]. To address this last issue, some authors have suggested running PCA on a reduced subset of SNVs, after first removing regions of

the genome that are known to have high or long-range LD[33] and/or performing LD pruning[34,35]. However, these suggestions are not universally implemented, and the downstream implications of adjusting for PCs that capture features other than global ancestry are not fully understood.

In this paper, we investigate the impact of ancestral heterogeneity on genome-wide association studies in admixed populations. Through both simulation studies and analytic results, we provide new insight into when genetic association studies must adjust for global ancestry. In addition, we compare two approaches for adjusting for global ancestry, using model-based estimates of admixture proportions or principal components, and show that using PCs can actually induce spurious associations in GWAS. To conclude, we provide suggestions regarding best practice for appropriately controlling for ancestral heterogeneity in genetic association studies in admixed populations.

2 Material and Methods

2.1 Regression models for 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. [... mention that logistic can be used for binary traits? ...]

We test for an association between the trait and genotype at each position by testing the null hypothesis $H_0 : \beta_j = 0$.

2.2 Inferring and adjusting for ancestral heterogeneity

To adjust for ancestral heterogeneity in genome-wide association studies in admixed populations, we include inferred global ancestry as a potential confounder in our regression models. We infer global ancestry using one of two techniques: model-based global ancestry inference of principal component analysis.

Note: some of the following is a bit redundant with the Introduction section; will likely want to trim here or in Intro

2.2.1 Model-based global ancestry inference

Various model-based approaches exist for estimating global ancestry proportions, also known as *admixture proportions*, in admixed populations. We represent global ancestry proportions via the vector $\boldsymbol{\pi}_i = \begin{pmatrix} \pi_{i1} & \dots & \pi_{iK} \end{pmatrix}^\top$, where π_{ik} denotes the genome-wide proportion of genetic material inherited by individual i from ancestral population k and $\sum_{k=1}^{K} \pi_{ik} = 1$. Note that the total number of ancestral populations, K, typically must be pre-specified, and the definition of global ancestry is typically restricted to the autosomes. Admixture proportions can be estimated directly using a program such as `ADMIXTURE`[22], or by calculating the genome-wide average local ancestry (i.e., $\hat{\pi}_{ik} = \frac{1}{2m} \sum_{j=1}^{m} a_{ijk}$), where local ancestry a_{ijk}—the number of alleles (0, 1, or 2) inherited by individual i from ancestral population k at position j—was first inferred using a program such as `RFMix`[36]. To adjust for ancestral heterogeneity, we include $K - 1$ of these estimated admixture proportions as covariates in our GWAS regression models:

$$E[y_i \mid g_{ij}, \hat{\boldsymbol{\pi}}_i] = \beta_0 + \beta_j g_{ij} + \beta_{\pi,1} \hat{\pi}_{i,1} + \cdots + \beta_{\pi,K-1} \hat{\pi}_{i,K-1}.$$

2.2.2 Principal component analysis

Principal component analysis (PCA) is an unsupervised dimension-reduction technique that is widely used for inferring population structure in genetic studies, with a number of software programs available for running PCA on genotype or sequence data (e.g., `EIGENSTRAT`[13], `SNPRelate`[37], `PC-Air`[30]). To run 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., $\mathbf{XX}^\top = \mathbf{UD}^2\mathbf{U}^\top$). The top principal components $(\mathbf{u}_1, \mathbf{u}_2, \dots)$ typically capture global ancestry[28,29]. To adjust for ancestral heterogeneity, we choose some number of principal components, P, needed to capture global ancestry (typically $1 \leq P << n$) and include those PCs as covariates in our GWAS regression models:

$$E[y_i \mid g_{ij}, \hat{\boldsymbol{\pi}}_i] = \beta_0 + \beta_j g_{ij} + \beta_{u1} u_{i1} + \cdots + \beta_{uP} u_{iP}.$$

A number of techniques have been proposed for selecting the number of PCs, P, including formal significance tests based on Tracy-Widom theory[28,13], examining the proportion of variance explained by each PC[38], comparing PCs to self-reported ancestry[11], and/or keeping PCs that are significantly associated with the trait[39,40].

2.2.3 Variant- and sample-level filtering

Depending on the chosen technique and software, it is often recommended that filtering be performed at the variant and/or sample level prior to inferring global ancestry.

- MAF (Jenn Kirk)
- relatives (Matt Conomos)
- high missing rates (SNPs and people)
- high LD regions (refer to appendix, GitHub page with lists for different builds)
- LD pruning

[... Comment on both PCA and model-based approaches (e.g., ADMIXTURE, RFMix, other ancestry inference software) ...]

2.3 Simulation study using TOPMed whole genome sequence data

- Need to decide if we're using TOPMed or WHI data
- brief intro

2.3.1 TOPMed whole genome sequence data

- describe sequencing methods
- which samples we used
- dbGap accession
- QC
- removing relatives

2.3.2 Genetic ancestry inference

- ADMIXTURE
- PCA
- what filtering was performed, and how many variants left after filtering

2.3.3 Evaluating population structure adjustment approaches

- simulating traits (effect sizes, choice of causal SNPs)
- running GWAS
- 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
- 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] Need, A. C. and Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. Trends in Genetics *25*, 489–494.

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

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

[4] Morales, J., Welter, D., Bowler, E. H., Cerezo, M., Harris, L. W., McMahon, A. C., Hall, P., Junkins, H. A., Milano, A., Hastings, E. *et al.* (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biology *19*, 21.

[5] Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The missing diversity in human genetic studies. Cell *177*, 26–31.

[6] Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics *51*, 584–591.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

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

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

[27] 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.

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

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

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

Figure Titles and Legends

Tables