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. Department of Medicine, Division of Medical Genetics, 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). Although it has been shown that the top principal components (PCs) typically reflect population structure, deciding exactly how many PCs must be included as covariates 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, we show that adjusting for extraneous PCs can induce spurious associations as a result of the phenomenon known as collider bias. Through both analytic results and application to 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 spurious associations can arise when regression models adjust for PCs that 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 issues of collider bias can be avoided entirely in these data, and the rate of spurious associations appropriately controlled, 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.

1 Introduction

Historically, admixed populations such as African Americans and Hispanics/Latinos have 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]. Ancestral heterogeneity is an example of population structure, which is known to pose challenges for genetic association studies: In particular, it can lead to spurious associations in GWAS unless appropriately addressd.

- define population structure (use *ancestral heterogeneity* instead??) and cite papers showing it exists in admixed populations
- why we need to adjust
- discuss methods for adjusting for population structure, with particular focus on PCA; mention challenges in choosing P and deciding how to pre-process
- outline rest of paper

2 Material and Methods

2.1 Regression models for genome-wide association studies

- introduce notation (trait, genotype, covariates)
- describe regression framework used for running GWAS (linear regression model, but mention that logistic can be used for binary traits)

2.2 Inferring and adjusting for ancestral heterogeneity

remind that we're focusing on fixed effect adjustment

2.2.1 Model-based global ancestry inference

- introduce notation
- direct (e.g., ADMIXTURE) vs indirect (e.g., average local ancestry)
- adding to regression model

2.2.2 Principal component analysis

- define
- cite existing programs
- cite papers that show top PC(s) capture genetic ancestry
- choosing P
- adding to regression model

2.2.3 Variant- and sample-level filtering

recommended filtering before model-based GAI and/or PCA:

- 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

2.3 Simulation study using TOPMed whole genome sequence data

brief intro

2.3.1 TOPMed whole genome sequence data

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

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
- Suggested diagnostics

5 Appendices

5.1 Regions Removed Prior to PCA

5.2 Mathematical Derivations

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 repo: 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.

Figure Titles and Legends

Tables