1 Target Journal

American Journal of Human Genetics

2 Title

Adjusting for principal components can induce spurious associations in genome-wide association studies

3 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

4 Correspondence

kgrinde@macalester.edu

5 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.

6 Introduction

GWAS in Admixed Populations

Adjusting for Population Structure

Choosing PCs

Outline rest of paper

7 Material and Methods

7.1 GWAS Regression Framework

Introduce notation, describe general set of methods used for running GWAS and adjusting for population structure

7.2 Adjusting for Population Structure

Introduce notation, describe methods used for inferring and adjusting for global ancestry proportions, describe methods used for computing and adjusting for principal components

Do we need a new (more specific) term other than *Population Structure*?

7.2.1 Global Ancestry Proportions

7.2.2 Principal Component Analysis

7.3 Simulations

Describe simulation study verifying theoretical work: simulating genotypes, simulating traits, evaluation

7.4 Application to 1000 Genomes (?) Sequence Data

Describe data: sequencing methods, populations used, simulating traits, evaluation

8 Results

8.1 Need to Adjust

Simulation and mathematical results

8.2 Inferring Population Structure

Highly correlated with each other

Some PCs are highly correlated with genotypes in small regions

This could also go after we show spurious assoc results: here's the problem ¿ here's where we think the problem is coming from

8.3 FWER

Do they have similar FWER? (Manhattan plots for one or two traits, overall summary of rejection rates)

Is it appropriate to use the same significance threshold for both?

8.4 Spurious Associations

Simulation (Manhattan plots for one or two traits, overall summary of rejection rates) and mathematical results demonstrating spurious associations when adjusting for certain types of PCs

8.5 Power

Do they have similar power?

9 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

10 Appendices

10.1 Regions Removed Prior to PCA

10.2 Mathematical Derivations

11 Supplemental Data

12 Declaration of Interests

The authors declare no competing interests.

13 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.

14 Web Resources

15 Data and Code Availability

16 References

17 Figure Titles and Legends

18 Tables