1 Title

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

2 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

3 Correspondence

kgrinde@macalester.edu

4 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.

5 Introduction

6 Material and Methods

7 Results

8 Discussion

9 Appendices

10 Supplemental Data

11 Declaration of Interests

The authors declare no competing interests.

12 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.

13 Web Resources

14 Data and Code Availability

15 References

16 Figure Titles and Legends

17 Tables