

1 Title

2 Adjusting for principal components can induce spurious associations in genome-wide associ-
3 ation studies

4 2 Authors and Affiliations

5 Kelsey E. Grinde^{1*}, Brian L. Browning², Sharon R. Browning³

6 1. Department of Mathematics, Statistics, and Computer Science, Macalester College,
7 Saint Paul, MN, 55105, USA

8 2. Department of Medicine, Division of Medical Genetics, University of Washington, Seat-
9 tle, WA, 98195, USA

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

11 3 Correspondence

12 kgrinde@macalester.edu

13 4 Abstract

14 Principal component analysis (PCA) is widely used to control for population structure in
15 genome-wide association studies (GWAS). Although it has been shown that the top prin-
16 cipal components (PCs) typically reflect population structure, deciding exactly how many
17 PCs must be included as covariates in GWAS regression models can be challenging. Often
18 researchers will err on the side of including more PCs than may be actually necessary in
19 order to ensure that population structure is fully captured. However, we show that ad-
20 justing 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.

35 **5 Introduction**

36 **6 Material and Methods**

37 **7 Results**

38 **8 Discussion**

39 **9 Appendices**

40 **10 Supplemental Data**

41 **11 Declaration of Interests**

42 The authors declare no competing interests.

43 **12 Acknowledgments**

44 K.E.G. was supported by the National Science Foundation Graduate Research Fellowship
45 Program under grant no. DGE-1256082. Any opinions, findings, and conclusions or rec-
46 ommendations expressed in this material are those of the author(s) and do not necessarily
47 reflect the views of the National Science Foundation.

48	13	Web Resources
49	14	Data and Code Availability
50	15	References
51	16	Figure Titles and Legends
52	17	Tables