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

Kelsey E. Grinde 1 \*, Brian L. Browning 2, Sharon R. Browning 1

1 *Department of Biostatistics, University of Washington, Seattle, WA, USA*

2 *Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA, USA*

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.

Other information that will be submitted:

* Abstract track: basic/translational
* [Main topic](https://www.ashg.org/2019meeting/pages/abstract_topics.shtml): statistical genetics and genetic epidemiology
* [Keywords](https://www.ashg.org/2019meeting/pages/abstract_keywords.shtml): genome-wide association, GWAS, population structure, linkage disequilibrium, statistical genetics
* Epstein Trainee Award nomination
  + “The nominator will be asked by email to complete a form verifying the trainee status of the applicant and the work completed by this trainee.”
  + “both the nominator and the trainee applicant must be 2019 ASHG members.”

Formatting notes:

* **Deadline:** Thursday, June 6
* **Word limit:** 2500 characters