Research Updates
May 21, 2019

Spurious Associations Project

Based on feedback from last week's lab meeting presentation, I am planning to make the following changes to my presentation of our spurious associations project.

1. New collider bias example

Instead of using the attractiveness/talent/acting example, I will instead use the following (more relevant) example from Day et al., (2016). "A robust example of collider bias in a genetic association study."

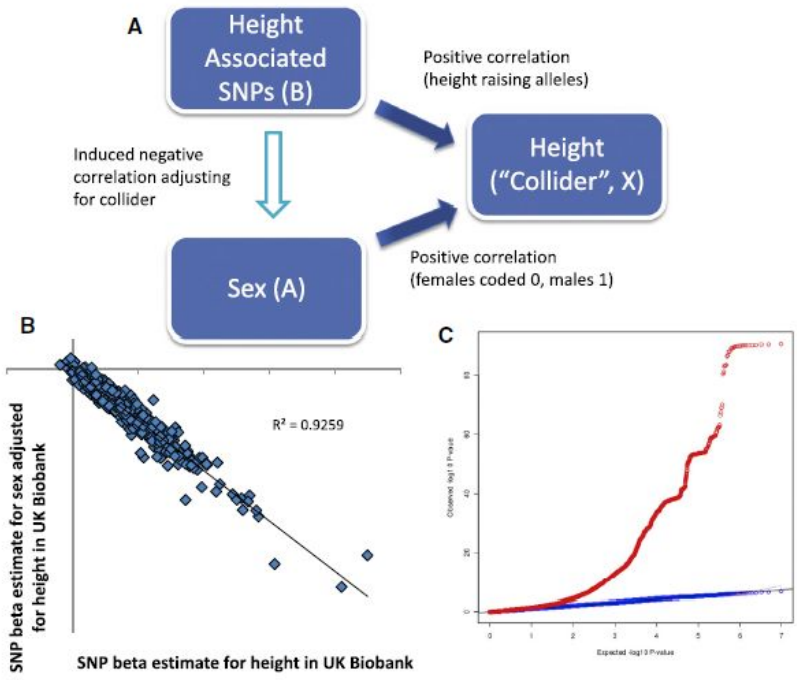
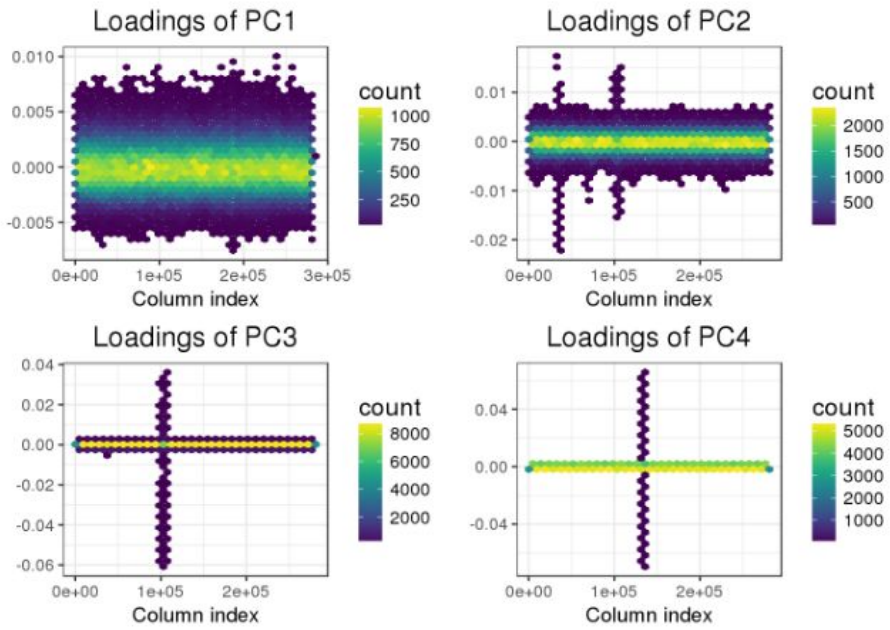


Figure 1. Induced Collider Bias between Genetic Variants, Height, and Sex
(A) Schematic diagram of the scenario in which collider bias can occur between genetic variants, height, and sex.
(B) Spurious autosomal SNP-effect estimates for sex, created by adjusting for height as a covariate, are almost perfectly correlated with SNP-effect estimates for height. In this scenario of collider bias, adjustment for the collider height creates biologically implausible sex associations for the 694 previously identified genome-wide significant autosomal SNPs for height.
(C) A quantile-quantile plot of genome-wide autosomal test statistics for sex ~ SNP (shown in blue) and sex ~ SNP + height (shown in red).

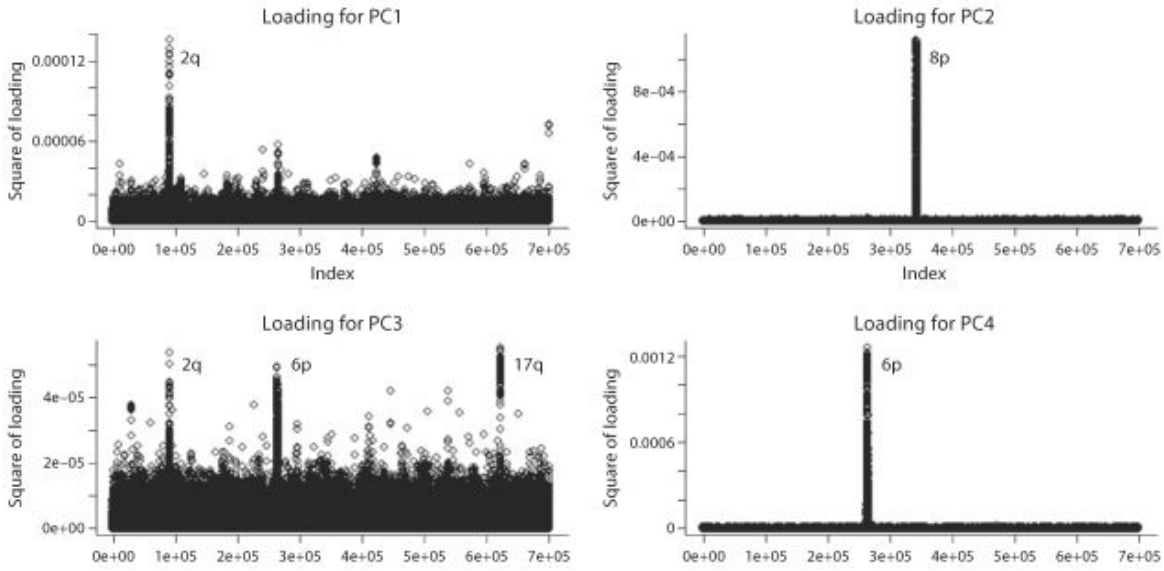
2. Examples of PCs capturing small regions of the genome in Europeans

To provide additional context regarding existing knowledge of the potential for PCs to capture small regions of the genome, I will add one of the following examples from studies in European populations.

Option 1: from the GitHub tutorial: [Prive (2017). "How to capture population structure with PCA?"]



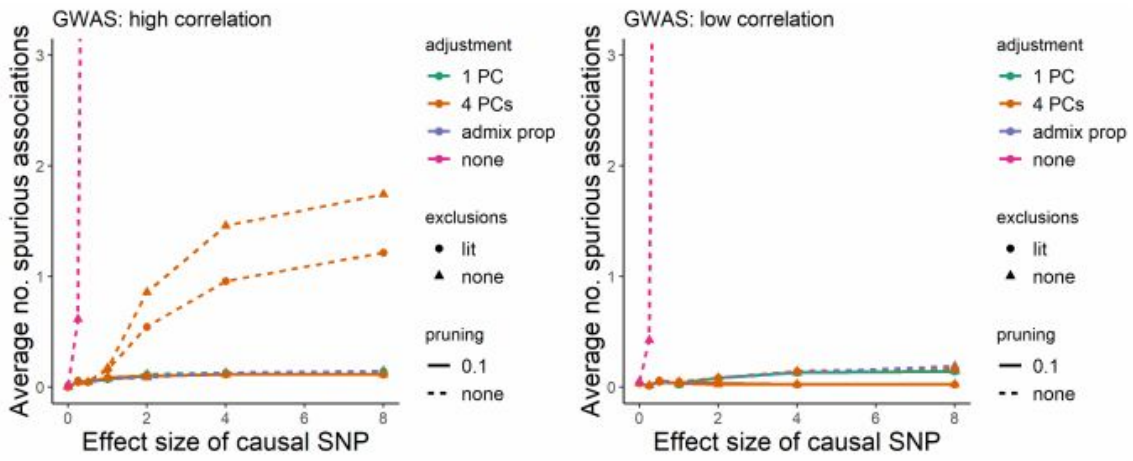
Option 2: from [Zou et al., (2010). "Quantification of Population Structure Using Correlated SNPs by Shrinkage Principal Components"]



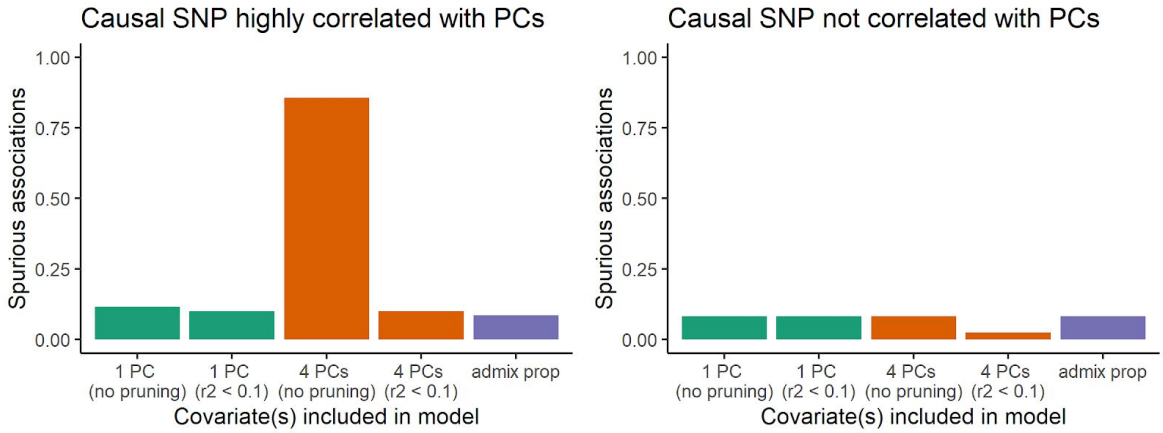
3. New ways to present simulation results

To show that spurious associations can arise in a single simulation setting (i.e., for a single choice of causal SNP), I will continue to present side-by-side Manhattan plots.

To compare the rate of spurious across all simulation settings, I had been using figures like this:



However, these figures are difficult to interpret, in part because there is a lot of information being conveyed. One alternative is to focus on a single choice of causal SNP effect size (e.g., $\beta_1 = 2$) and reduce the number of models being compared (e.g., just focus on the impact of LD pruning):



Question to discuss: Are these new figures easier to process? Any other ideas on how to present results that might be even clearer?

4. Other questions regarding spurious associations project

We discussed the possibility of motivating our choice of causal SNP effect sizes by pointing to published associations between HLA and traits such as diabetes.

- Most of the associations reported in the GWAS catalog for HLA are from studies with binary traits, so the effect size is reported as an odds ratio. However, there are a few examples that report $\beta \approx 1$ and p-values smaller than 1×10^{-100} (see here).
- Would it be reasonable, then, to point to these studies as motivation for using an effect size of 1?

For our manuscript, I am working on showing that spurious associations can replicate across studies.

- Does this need to be finished in time for my final exam, or to include in the final draft of my dissertation? Or is enough to say in the dissertation/final exam that this is future work?

Administrative Updates

1. Plan for upcoming meetings

May 28: discuss final exam slides (could also do this on May 29 after statgen and popgen presentations)

June 4: discuss dissertation draft

2. Conference abstracts

ASHG:

- Aim to submit spurious associations project as late-breaking abstract
- Last year, the late-breaking deadline was in early September

IGES: (deadline is May 27)

- The spurious associations project is not ready to submit, so I've been thinking of other projects that I could present.
- One idea: the need to adjust for global ancestry in admixture mapping studies + the impact that adjusting for global ancestry has on the correlation of admixture mapping test statistics [and, as a result, the multiple testing correction]