Influence of UL BiLEVE participants on UK Biobank GWAS

Gibran Hemani

2016-07-22

Problem:

> Ascertainment on smoking will induce spurious association between genetic variants that affect smoking and any traits that "cause" smoking, and traits that are genetically or environmentally related to smoking. The extent of bias induced will depend on the phenotypic correlation between smoking and the variable analysed. For variables that do not show strong correlations with smoking, then this bias will be small. - Dave Evans

The model is:

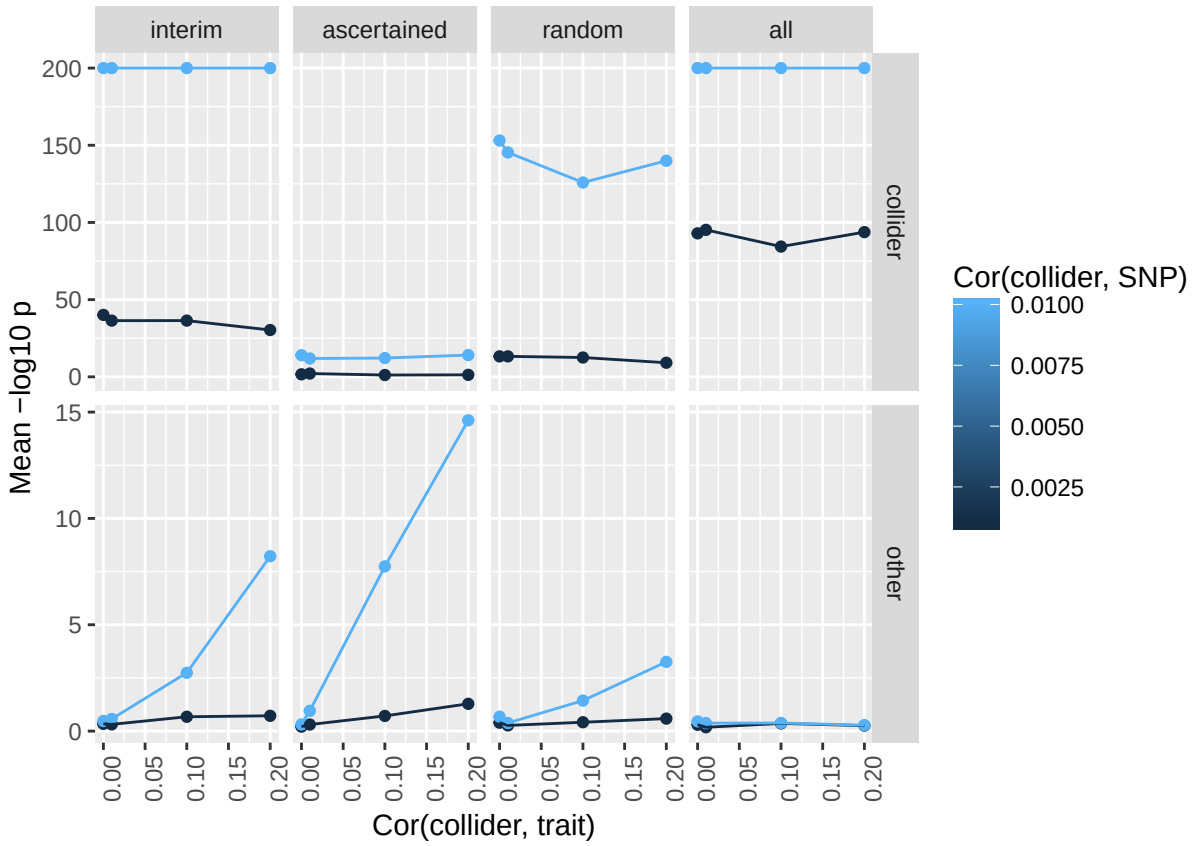
```
snp -> smoking <- trait
g -> collider_trait <- other_trait
```

The total sample size is 500000. The proportion that are UL BiLEVE is 0.1 (around 50000). The interim sample size is 150000 - so 100000 are sampled randomly from the remaining 'non-UL BiLEVE' samples. Simulate the following parameters:

	sim	cor_collider_other	cor_collider_snp
1	1	0.00	0.001
11	1	0.01	0.001
21	1	0.10	0.001
31	1	0.20	0.001
41	1	0.00	0.010
51	1	0.01	0.010
61	1	0.10	0.010
71	1	0.20	0.010

To ascertain samples for UL BiLEVE, take the individuals with the top 10% of values in the collider trait. **This may be too extreme.**

Perform 10 simulations for each row. The results are shown below:



In this graph:

```
all = all 500000 samples in biobank
interim = 150000 interim samples including UL BiLEVE
ascertained = only the UL BiLEVE samples
random = only the non-UL BiLEVE samples
```

These are the average p-values between the SNP and the collider trait (e.g. smoking, top row) and the SNP and the other trait (e.g. correlated with smoking, bottom row).

Conclusions:

- There is no inflation in the for the entire biobank sample.
- The interim sample will have inflation
- Excluding the ascertained sample will still have inflation, but not as much