Split biobank into discovery and replication?

Gibran Hemani

2016-08-05

All code to run these analyses are here: https://github.com/explodecomputer/biobank_gwas_power

Objective

- Test the difference in power if we split 500k Biobank samples into discovery and replication sets
- Test the difference in bias for detected signals

Calculating power for sample sizes of 500000 and significance thresholds of 5e-8 leads to numerical instability using standard packages, so these simulations are based on an approximation for estimating the effect sizes for specific power levels, and then Monte Carlo simulations to check they make sense.

NOTE: These simulations assume a real effect. An obvious utility of having discovery and replication samples is to improve robustness, but that is not being assessed here.

One stage approach

Perform GWAS using all 500k samples using a threshold of 5e-8.

Two stage approach

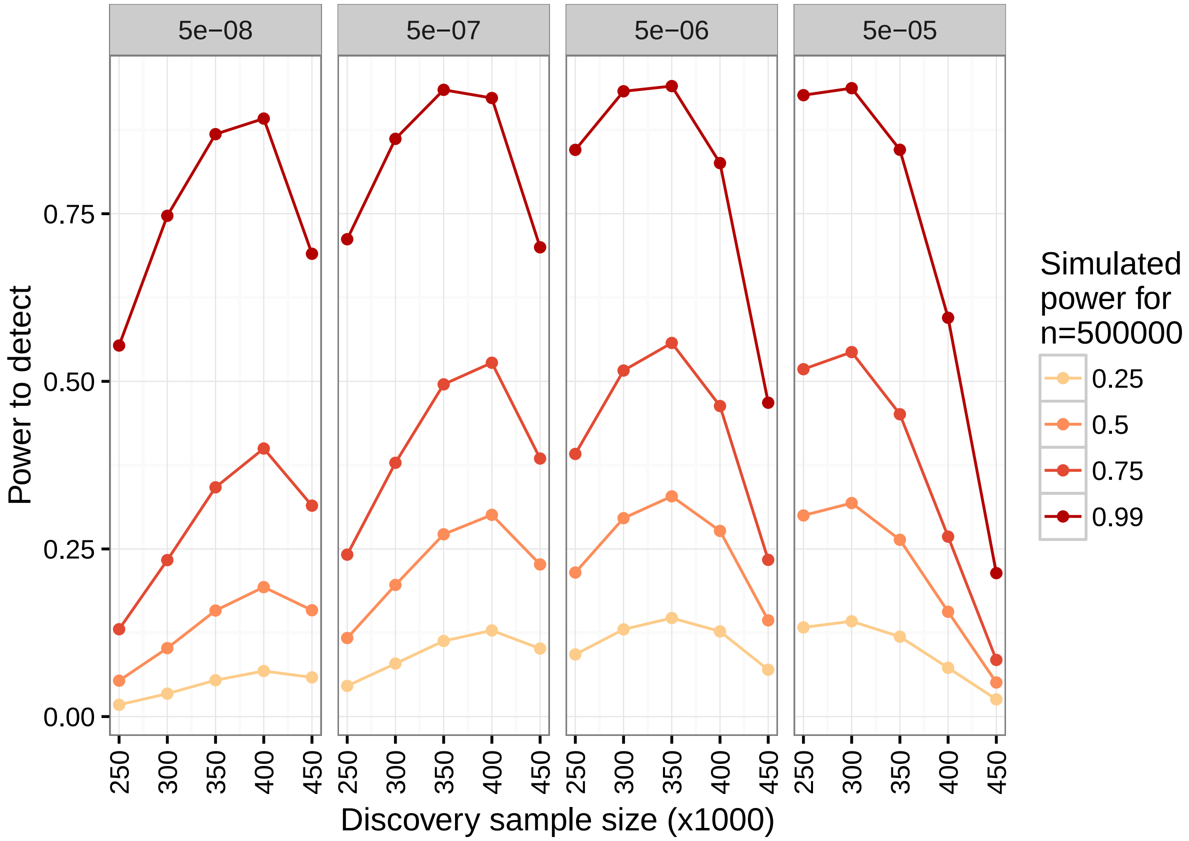
1. The data is split into discovery and replication sets
2. GWAS is performed in discovery. Some threshold is applied to identify SNPs to take forward to the replication data
3. Test for replication of the signals. Correct for multiple testing, e.g. if the discovery threshold was 5e-5, and there are 1 million independent regions in the genome, assume that 50 false positives would be identified, so a replication threshold of 0.05 / 50 = 0.001.
4. To be significant, the SNP has to pass the thresholds in both (2) and (3).

Theoretical results

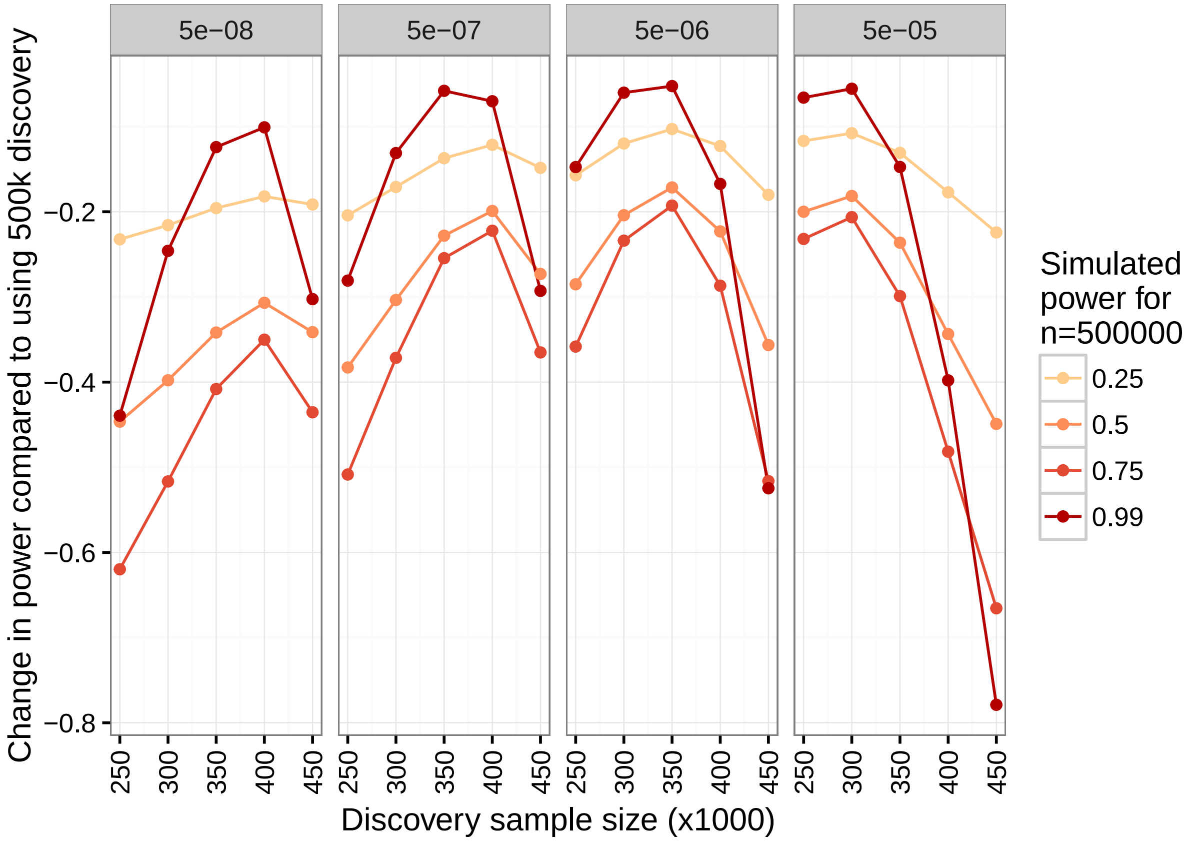
- The sample split can be e.g. 250k discovery 250k replication, or some other proportions
- The discovery threshold can be 5e-8 or something more relaxed
- The effect size can be simulated such that for a full 500k sample the power is 0.25, 0.5, 0.75, 0.99

Given this range of effect sizes, how do different sample splits and discovery thresholds measure up against just doing a single GWAS using all 500k samples?

This graph just show the power for the split sample approach:



Compared against what the power for just doing a 500k sample:

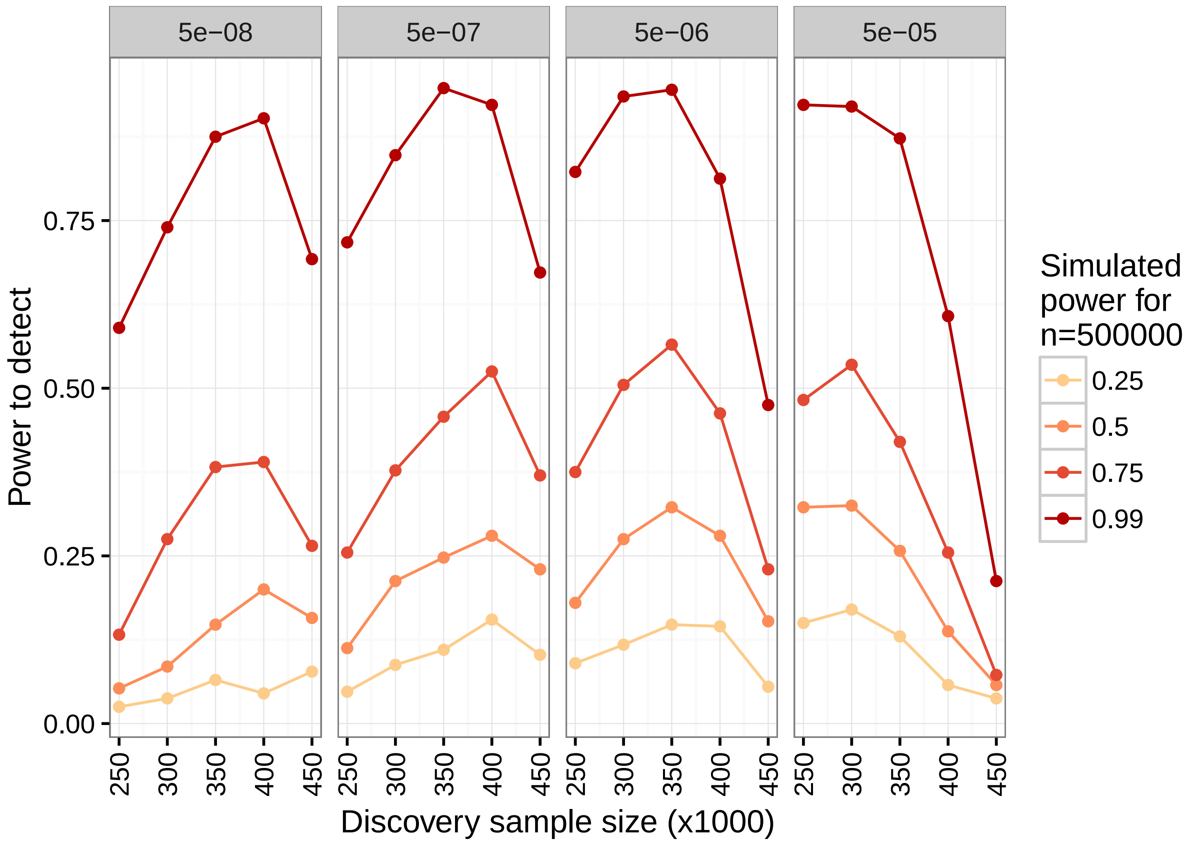


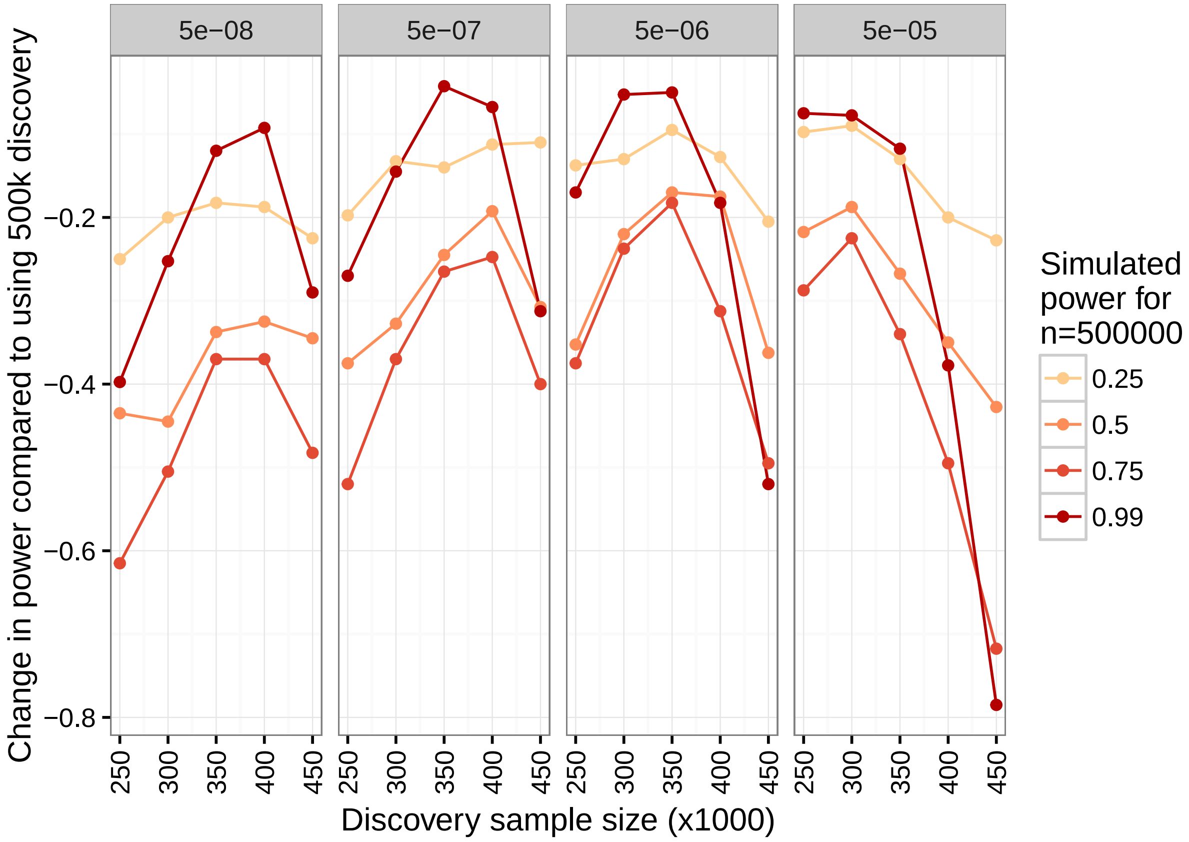
Summary

It is clear that doing the full 500k GWAS without splitting will be more powerful in all scenarios. If we were to split the data then something like 350k discovery vs 150k replication, with a discovery threshold of 5e-6 would be the most powerful approach for a range of effect sizes.

Simulation results

The same plots are shown below for data that were simulated to the same parameters:


5e−08
5e−07
5e−06
5e−05
Power to detect
0.00
0.25
0.50
0.75
250
300
350
400
450
Discovery sample size (x1000)
Simulated power for n=500000
0.25
0.5
0.75
0.99


5e−08
5e−07
5e−06
5e−05
Change in power compared to using 500k discovery
−0.8
−0.6
−0.4
−0.2
250
300
350
400
450
Discovery sample size (x1000)
Simulated power for n=500000
0.25
0.5
0.75
0.99

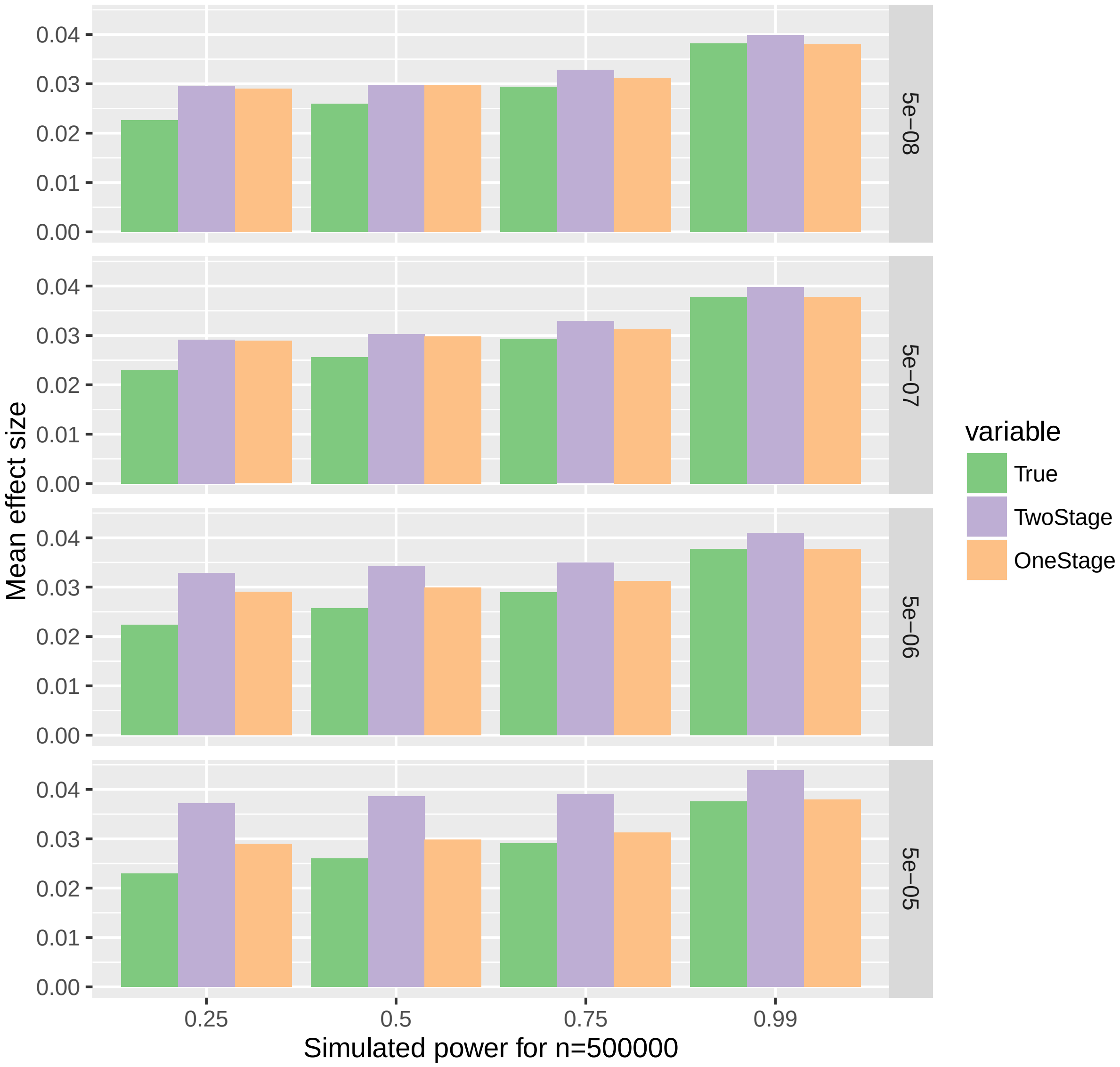
Bias in effect sizes

Winner's curse applies to GWAS, and is higher when power is lower. A very large effect that is not close to the significance threshold will not be influenced by winner's curse, but SNPs that are hovering around the threshold are much more likely to be overestimated effect sizes. Having a replication dataset might alleviate this problem…

Using the simulation data, calculated

- the true simulated effect
- the one stage effect size - **just the effect size for SNPs that were significant**
- the two stage effect size - the effect size obtained from the replication sample **if the SNP was significant in the discovery and the replication**

Results:



Note that in almost all cases the One stage and Two stage approaches both bias the discovered effects upwards. However, in all cases the bias in the Two stage approach is higher.

Summary

Winner's curse is a greater issue in the two stage approach than the one stage approach, probably because:

- power is lower
- there is still some thresholding being applied