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

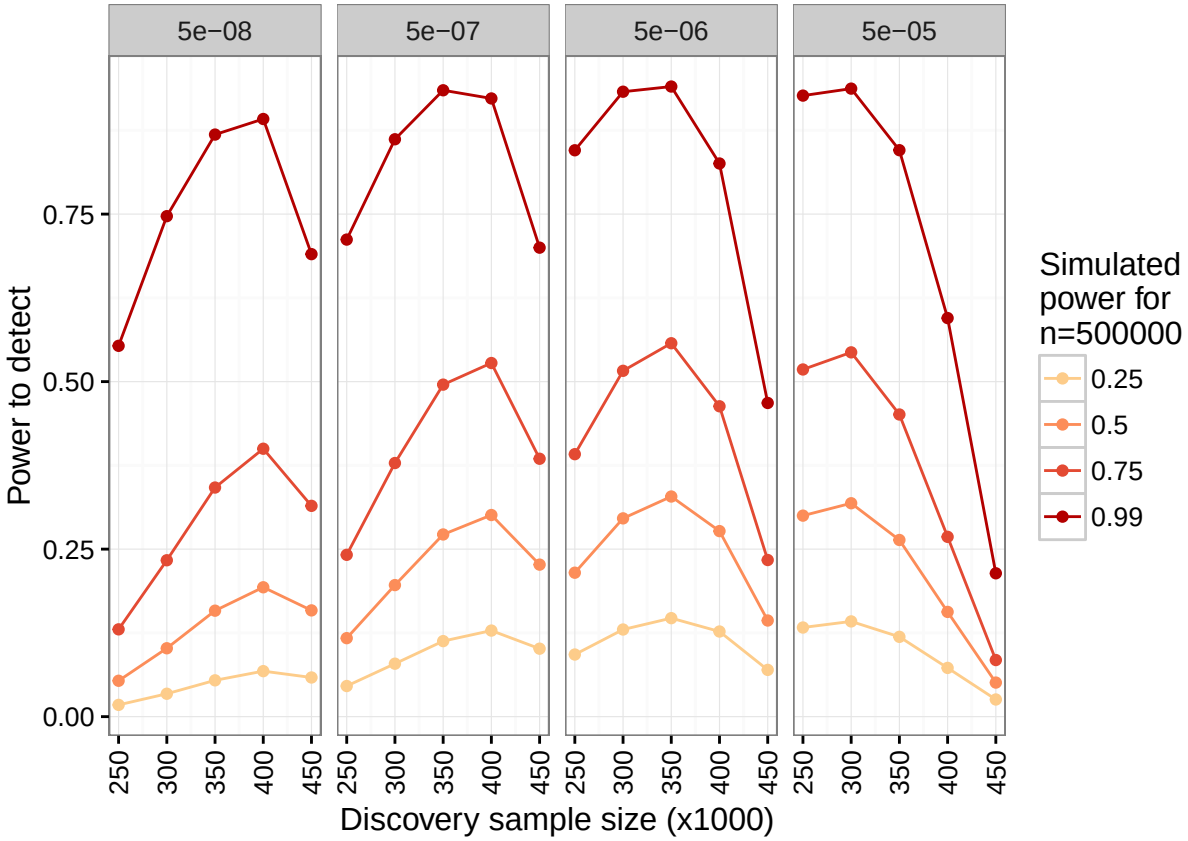
1. The data is split into discovery and replication sets
2. GWAS is performed in discovery. Some threshold is applied to identify SNPs to take forward to the replication data
3. Test for replication of the signals. Correct for multiple testing, e.g. if the discovery threshold was 5e-5, and there are 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

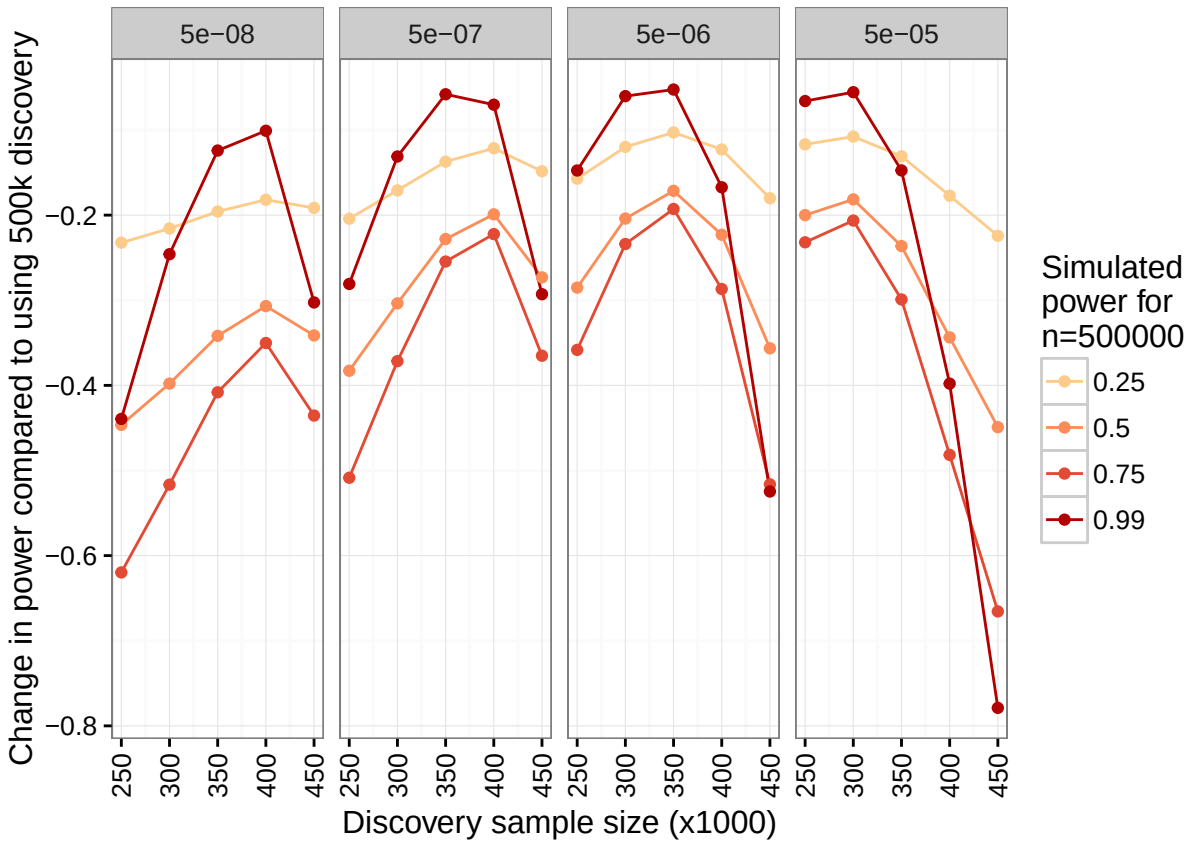
- The sample split can be e.g. 250k discovery 250k replication, or some other proportions
- The discovery threshold can be 5e-8 or something more relaxed
- The effect size can be simulated such that for a full 500k sample the power is 0.25, 0.5, 0.75, 0.99

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 - columns of graphs are for different discovery thresholds



Compared against what the power for just doing a 500k sample - columns of graphs are for different discovery thresholds

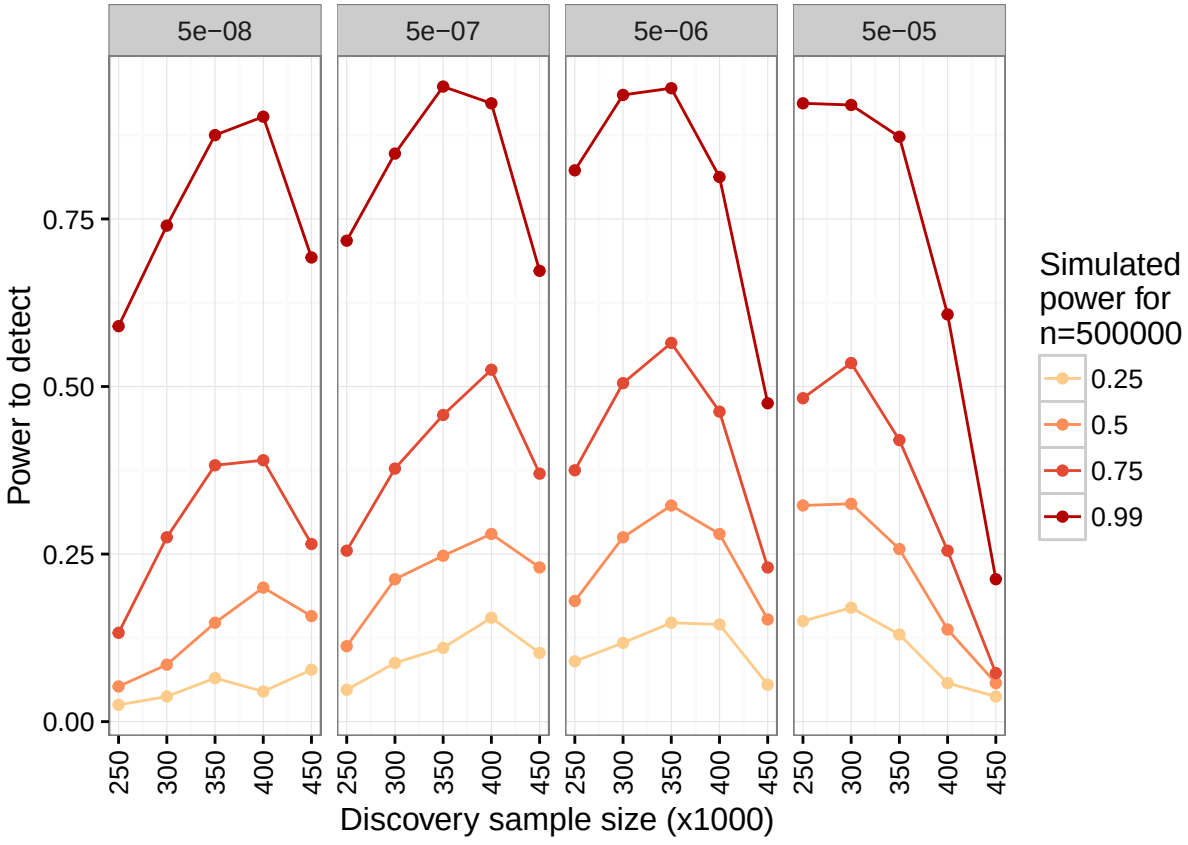


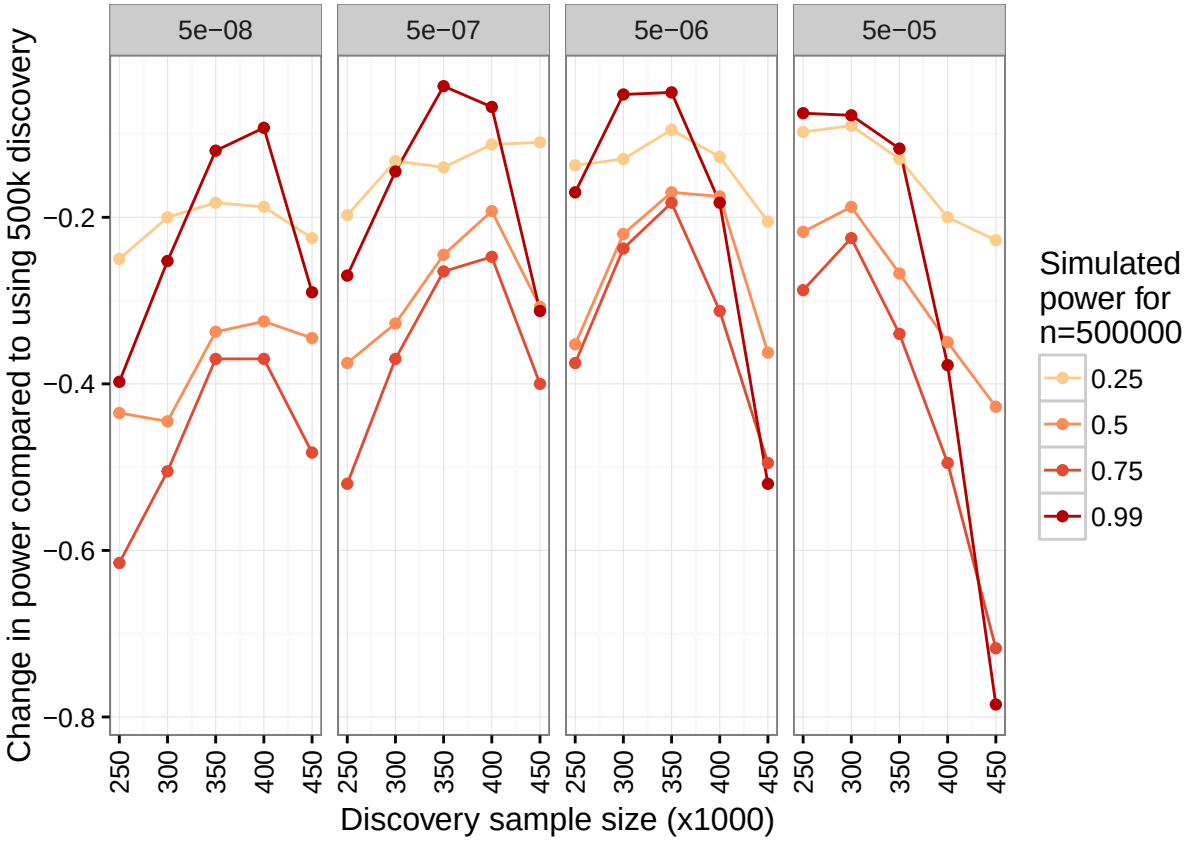
Summary

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
−0.2
−0.4
−0.6
−0.8
250
300
350
400
450
Discovery sample size (x1000)
Simulated power for n=500000
0.25
0.5
0.75
0.99

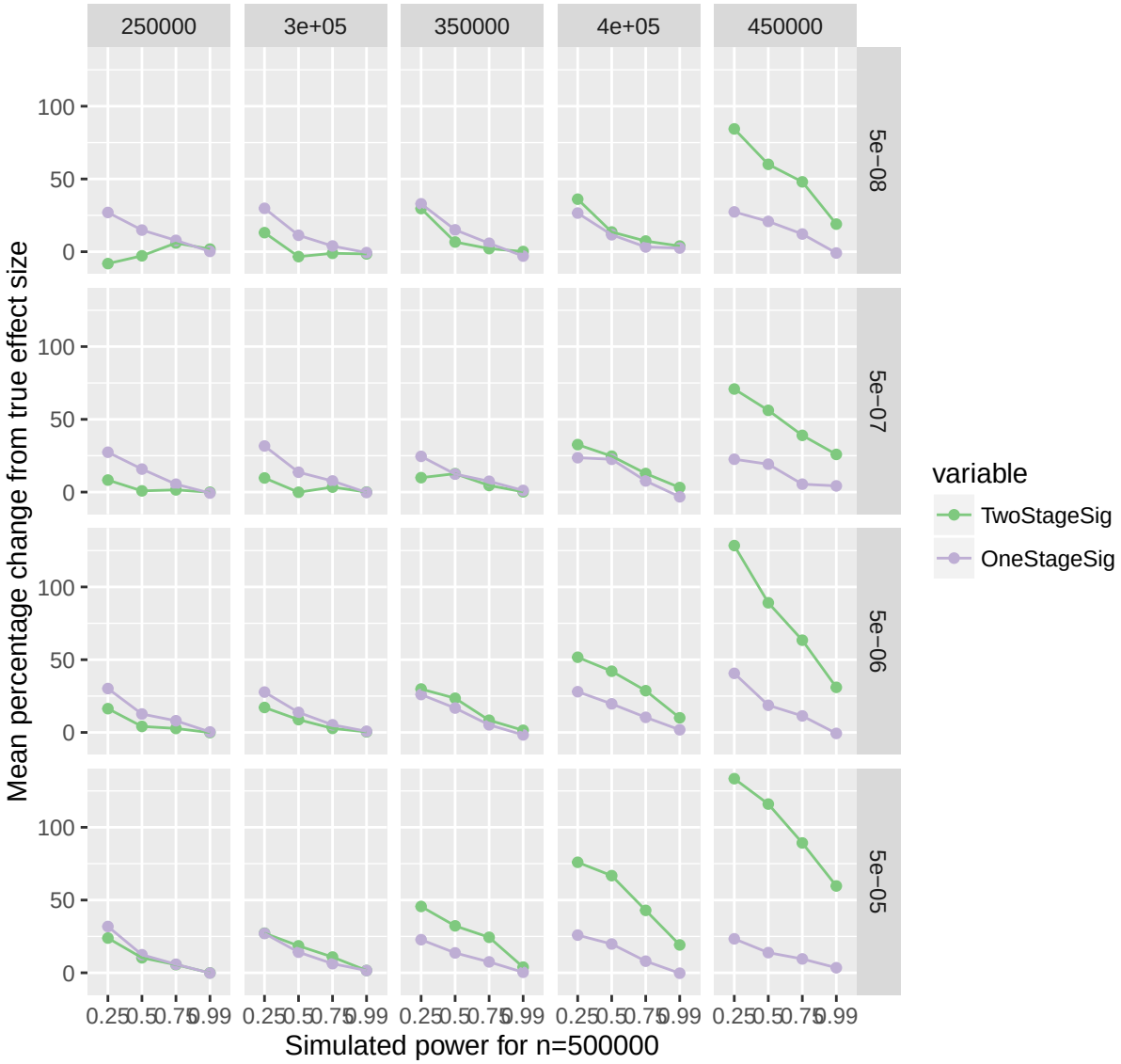
Bias in effect sizes

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 - columns of graphs are for different discovery sample sizes, e.g. from 250k discovery 250k replication to 450k discovery 50k replication. Rows of boxes are for different discovery thresholds.



Note that in almost all cases where power is below 0.99 there is bias for each method. Having a small replication sample is a problem for TwoStages.

The only way to avoid bias is to have a TwoStage design where the discovery and replication sizes are each 250k, and the discovery threshold is 5e-8. However, this may still incur bias under the case of multiple true positives, because the replication threshold will be higher.

For the case where discovery sample size is 350k (replication 150k), and discovery threshold is 5e-6, the bias is fairly similar between One stage and Two stage approaches.

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

If the replication sample was used to **just estimate the effect size** and not to test for significance in the replication stage then this would not be an issue - we would get no bias in effect size estimation. But this would just lead to substantially reduced effects.

We should probably continue with using the full 500k to do a One stage analysis. For things like MR analysis, we will be suffering from inaccurate estimates if we don't account for winner's curse in the exposure, e.g by re-estimating the effect size in an independent sample.