Supplementary Materials: Quantifying utility of PRS models using synthetic genetic data to guide population-level screening of human diseases

Anubhav Kaphle anubhavkaphle@gmail.com

August 28, 2021

1 Supplementary information

Here, I describe how to use the two tools I described in the previous document.

1.1 G-WIZ

paper source: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0220215

The tool is written is R and available for download from https://github.com/jonaspatronjp/GWIZ-Rscript/.
To run the tool I have a set of scripts that just needs to be run in R giving correct filepaths.

We will use script name *GWAS_call_batch.R* to run the analysis. We can just run each of these steps in R and we will get the output. I have setup everything and removed some bugs in the orginal software. So, I expect a smooth run for this.

The important things to note in the script are:

- The PRS summary data should be inside the folder with name "Data" but this can be changed from the script.
- The folder structure is critical so please do not change anything in the script.

Table 1 shows how the PRS files should look to be able to run GWIZ. The column names are important.

The final results from the analysis will be saved in the folder *resampled Data*. It is a plain-text file so can be used to do any analysis on it.

phenotype	dataset	accession	control_size	case_size	risk_allele_freq	OR	model
n/a	n/a	rs10752881	50000	1001	0.76	1.07	recessive
n/a	n/a	rs6691170	50000	1001	0.31	1.06	recessive
n/a	n/a	rs10936599	50000	1001	0.19	1.04	recessive
n/a	n/a	rs1321311	50000	1001	0.3	1.1	recessive

Table 1: Example structure of the PRS file that G-WIZ readily reads without much fuss.It should be a comma-separated file. A basic shell script can be used to make it easier to convert any structure of PRS file to this structure.

1.2 PLINK 1.9

Next, I will describe how PLINK can be used to generate those datasets:

PLINK v1.9 download link: `https://www.cog-genomics.org/plink/`

Once downloaded, please follow these commands to run plink: The command generates genetic data using information provided in *snps.sims* file, 100K cases and controls, with disease prevalence 0.1 (can be obtained from epidemiological studies).

```
~/tools/plink --simulate snps.sims --make-bed --out \\
plink_file --simulate-ncases 100000 --simulate-ncontrols 100000 \\
--simulate-prevalence 0.1
```

This is how *snps.sims* files should look like:

```
100000  null      0.00 1.00  1.00 1.00
 100     disease   0.00 1.00  2.00 mult
```

Description of the columns are as follows, more info here `https://zzz.bwh.harvard.edu/plink/simulate.shtml`.

```
Number of SNPs in this set
Label of this set of SNPs
Lower allele frequency range
Upper allele frequency range
Odds ratio for disease, heterozygote
Odds ratio for disease, homozyygote (or "mult")
```

PLINK outputs results in *bed* and *bim* files. These are binary files which cannot be read into python without much coding. So, to recode these files into numericals and output results in plain text file, use the following command or versions of the command:

```
~/tools/plink --bfile data --recodeAD
```

More details about coding here `https://zzz.bwh.harvard.edu/plink/dataman.shtml`. You can choose SNP coding based on what kind of genetic model is assumed for the snps. If additive model, then we just sum up the number of effect allele in the genotype, if dominance model is used we give 1 to genotype containing at least 1 of the effect alleles.