Documentation to the L1outPRS program

Jurg Ott, Rockefeller University, New York 1 Jan 2025

This outline documents a simple implementation of a polygenic risk score (PRS) for GWAS data, that is, case-control data genotyped at a number of genetic variants, SNPs. It makes use of the `--score` function in *plink* [1-3], based on an allelic association test, `--assoc`. A PRS is computed for each case and control individual, with cases on average showing higher scores than controls. Classification (phenotype prediction) is performed by calling an individual a "case" if the individual's score is above the 95th percentile of control scores, and otherwise is called a "control". The raw classification is followed by cross-validation, that is, phenotype prediction is carried out for individuals not used for the classification procedure. The specific type of cross-validation implemented here is the Leave-one-out method [4], hence the program name, L1outPRS. The Pascal source code was compiled with the *fpc* package in Ubuntu (https://www.freepascal.org/).

The L1outPRS program will compute PRSs for the best 5, 10, 20, …, 50,000 variants, where "best" refers to the ordering of variants selected by command line parameter #2 (see below). For each such selection, predicted phenotypes will be computed.

Once a prediction has been made for each individual, the following parameters are in general use:

Known phenotype	Predict "case"	Predict "control"
case	*a*	*b*
control	*c*	*d*

The letters, *a*, *b*, *c*, and *d*, indicate numbers of individuals, for example, *a* = number of cases predicted to be cases. Sensitivity, or power, of the test to identify cases is given by $a/(a + b)$; positive predictive value, PPV = $a/(a + c)$; negative predictive value, NPV = $d/(b + d)$; odds ratio, OR = $(a \times d)/(b \times c)$; prediction accuracy, ACC = $(a + d)/(a + b + c + d)$.

Included in this package is a case-control dataset, AMDHK, on wet age-related macular degeneration (AMD), collected in Hong Kong and published in 2006 [5]. The significant SNP in that publication is `rs10490924` on chromosome 10.

The focus here is on prediction, not statistical significance of a case-control association test [6]. There is often no clear distinction between prediction and significance. Even in excellent textbooks [4], we find "Two useful summaries of predictive power are Sensitivity …, Specificity". However, sensitivity and specificity are properties of a statistical test and have more to do with statistical significance than prediction. For phenotype prediction, a stated aim of PRSs, a PRS is best based on highly predictive SNPs than significant SNPs [7].

L1outPRS program

This program creates a polygenic risk score (PRS) for single variants as implemented in *plink 1.9* with the allelic scoring function, `--score`. Because L1outPRS uses the Linux `sort` function (which is very different from `sort` in Windows), it runs properly only in Linux, for example, Ubuntu (Linux executable = `L1outPRS`). The *plink* executable must reside in the program path as `/usr/local/bin/plink19`. Download it from https://www.cog-genomics.org/plink/ .

Command line parameters are as follows:

1) The *plink* fileset name, without ".map" and ".ped", for example, AMDHKg. A minor allele frequency of at least 0.01 (preferably 0.05) should have been applied.

2) A number, 1 or 2, to indicate whether the variants should be ordered by the *p*-value (by 1 – *p*, actually) [1] or by the odds ratio [2].
3) A number, 1 or 2, to use prediction accuracy, ACC [1], or the positive predictive value, PPV [2], as the score variable.

An example of a command-line job submission is as follows:

```
jurg@Lap2023:~$ L1outPRS AMDHKg 2 2
```

Preferably start such a job in the foreground to see potential error messages and verify that everything is going well. If so, the same job may be submitted to run in the background:

```
jurg@Lap2023:~$ (L1outPRS AMDHKg 2 2 /dev/null)&
```

A report file, `L1outPRS.AMDHKg-2-2.rpt`, will be created in a directory above the running directory and will contain information on numbers of replicates processed.

Currently the following maximum program constants apply:

- 10,000 individuals

References

1. Chang, C.C. *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. DOI: 10.1186/s13742-015-0047-8
2. Chang, C.C. (2020) Data management and summary statistics with PLINK. *Methods Mol Biol* 2020, 49-65. DOI: 10.1007/978-1-0716-0199-0_3
3. Choi, S.W. *et al.* (2020) Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* 15, 2759-2772. DOI: 10.1038/s41596-020-0353-1
4. Agresti, A. (2019) *An introduction to categorical data analysis* (*Wiley series in probability and statistics*, 3rd edn), Wiley
5. Dewan, A. *et al.* (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* 314, 989-992. DOI: 10.1126/science.1133807
6. Lo, A. *et al.* (2015) Why significant variables aren't automatically good predictors. *Proc Natl Acad Sci U S A* 112, 13892-13897. DOI: 10.1073/pnas.1518285112
7. Wang, G. *et al.* (2024) Prediction mapping and polygenic risk scores in humans. *In preparation*,