Polygenic risk score analysis with PRSice-2

To perform polygenic risk score analysis we suggest to use PRSice (http://prsice.info). In this tutorial we provide a step-by-step guide to perform a simple polygenic risk score analysis using PRSice and explain how to interpret the results. The current tutorial is based on Linux. If you use Windows of Mac, or require additional information refer to the PRSice user manual.

Installation of PRSice-2

PRSice-2 is an R program that can be run in Unix, Linux and Mac OS from the command line. The code used in this tutorial used Linux PRSice version 2.2.11 and was tested on R versions 3.2.1 and 3.2.2.

First download and unpack the PRSice software with the following code on the command line.

```
wget https://github.com/choishingwan/PRSice/releases/download/2.2.11/PRSice_linux.nightly.zip
unzip PRSice_linux.nightly.zip
```

Installing required R packages

PRSice can automatically download all required packages, even without administrative right. You can specify the install directory using --dir. For example:

```
Rscript PRSice.R --dir .
```

will install all required packages under the local directory.

Running a polygenic risk score with PRSice

To run a polygenic risk score analysis on the toy data provided with PRSice run the following code, which is tested in R version 3.2.1

For binary traits:

```
Rscript PRSice.R --dir . \
--prsice ./PRSice_linux \
--base TOY_BASE_GWAS.assoc  \
--target TOY_TARGET_DATA \
--thread 1  \
--stat OR \
--binary-target T
```

For quantitative traits: see http://prsice.info

The *base* parameter refers to the file with summary statistics from the *base* sample (also known as discovery or training samples). These summary statistics contain for each genetic variant at least an

effect size and p-value. The *target* parameter refers to the prefix of the files (without file extension) that contain the genotype data in binary plink format (i.e., .bed,.bim,.fam file extensions). However, BGEN format is also supported. The base and target sample are also known as validation or test samples. This target sample should be completely independent from the base sample that was used to compute the summary statistics. Sample overlap across the discovery and target sample will greatly inflate the association between the polygenic risk score and the disease trait.

For the purposes of this tutorial we run the polygenic risk score analysis in PRSice folder where the toy data resides. In more realistic settings the data is likely to be in a separate folder and you can add another line of code with the parameter 'wd data-directory' to change the working directory of PRSice to data-directory. This will also be the directory where PRSice saves its results. Note that `\` characters are used to break up the code in different lines. If you add parameters do not forget to add `\` to all lines except the last one.

If the type of Effect (--stat) or data type (--binary-target) were not specified, PRSice will try to determine these information based on the header of the base file.

Instead of performing a polygenic risk score analysis on all genetic variants it is customary to clump first. In clumping, within each block of correlated SNPs the SNP with the lowest p-value in the discovery set is selected and all other SNPs are ignored in downstream analyses. This clumping procedure is performed by PRSice automatically, but can be adjusted with several clumping parameters. Although many other options exist, we refer to the PRSice user manual for more detailed information about the program.

For simplicity sake, we did not include principal components or covariates in this analyses, however, when conducting your own analyses we strongly recommend to include these.

Interpreting the results

By default, PRSice saves two plots and and several text files. The first plot is `PRSice_BARPLOT_<date>.png` (Figure S1). This plot shows the predictive value (Nagelkerke's R^2) in the target sample of models based on SNPs with p-values below specific thresholds in the base sample. In addition, for each model, a p-value is provided for the null hypothesis that the respective $R^2 = 0$. As Figure S1 shows, a model using SNPs with a p-value up to 0.4463 achieves the highest predictive value in the target sample with a p-value of 4.69493e-18. However as is often the case in polygenic risk scores analysis with relatively small samples, the predictive value is relatively low (Nagelkerke's R^2 around 5%). The text files include the exact R^2 values for each p-value threshold. The second plot is PRSice_HIGH-RES_PLOT_<date> .pgn (Figure S2) shows for many different p-value thresholds the p-value of the predictive effect (R^2) in black together with an aggregated trend line in green.

Both figures show that many SNPs that affect the trait in the base sample can be used to predict the trait in the target sample. Note that the two traits can be either the same or different. If the same trait is used the predictive value is related to the heritability of the trait (as well as the sample size of the base sample). If different traits are analyzed, the predictive value is also related to the genetic overlap between the two traits. Either way, polygenic risk score analysis typically shows that models with lenient p-value thresholds often predict better than models with more stringent thresholds, suggesting that many statistically insignificant SNPs still have predictive value in polygenic traits.

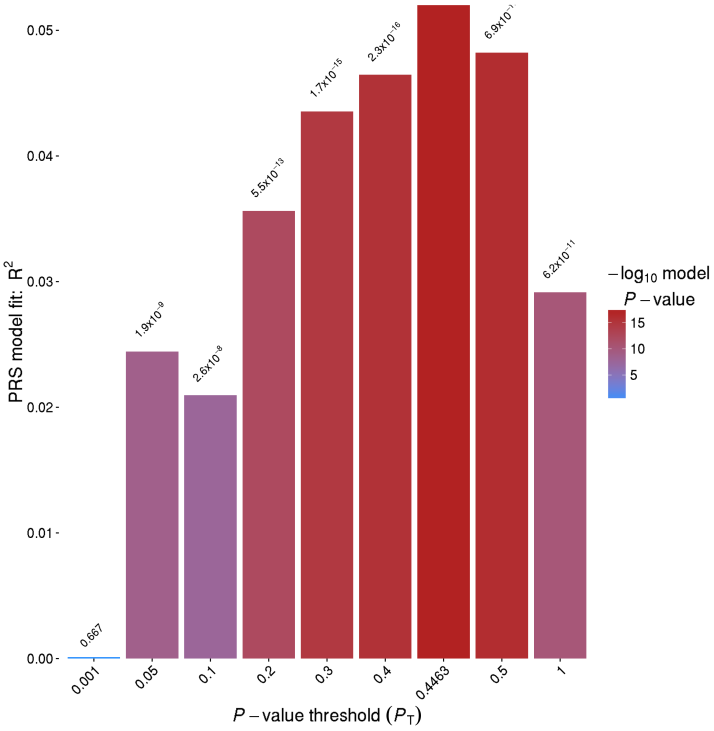

Figure S1. Default PRSice barplot. Nagelkerke R2 and p-value as a function p-value threshold in discovery sample.

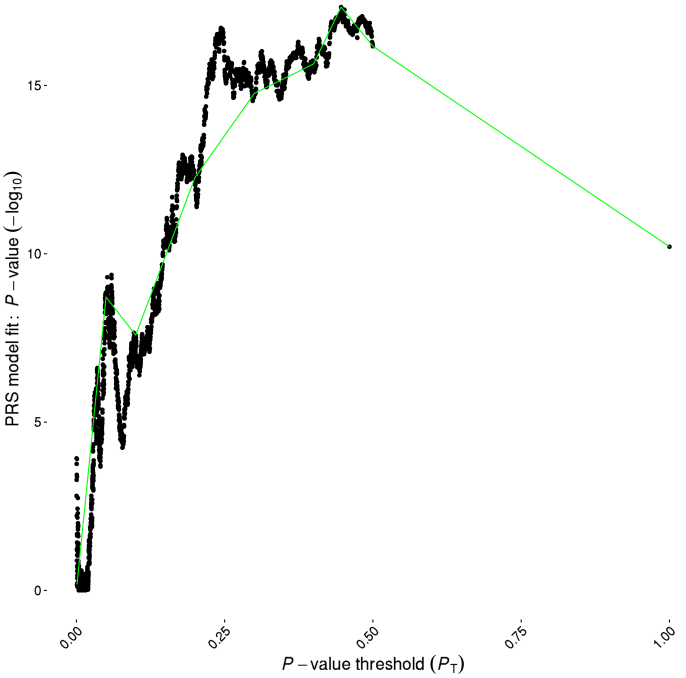

Figure S2. Default PRSice high resolution plot. P-value as a function of predictive p-value (black) together with a trend line (green).

Conclusion

In this tutorial we discussed how to perform a simple polygenic risk score analysis using the PRSice script and how to interpret its results. When PLINK genotype target files are available, PRSice provides a relatively easy way of performing polygenic risk score analysis. As mentioned before, PRSice offers many additional options to adjust the risk score analysis, including adding covariates and additional principal components and adjusting clumping parameters. We therefore recommend reading the user manual of PRSice to perform a polygenic risk score analysis optimal to the research question at hand.