PRSice: Polygenic Risk Score software
v1.23

Jack Euesden jack.euesden@kcl.ac.uk Cathryn M. Lewis Paul F. O'Reilly

June 1, 2015

Contents

1 Overview

PRSice (pronounced 'precise') is a software package for calculating, applying, evaluating and plotting the results of polygenic risk scores. PRSice can run at high-resolution to provide the best-fit PRS as well as provide results calculated at broad P-value thresholds, illustrating results corresponding to either, can thin SNPs according to linkage disequilibrium and P-value ("clumping"), handles genotyped and imputed data, can calculate and incorporate ancestry-informative variables, and facilitates the systematic application of PRS across multiple traits.

PRSice is a software package written in R, including wrappers for bash data management scripts and PLINK2 ([1]) to minimise computational time; thus much of its functionality relies entirely on computations written originally by Shaun Purcell in PLINK 1 and Christopher Chang in PLINK 2. PRSice runs as a command-line program, compatible for Unix/Linux/Mac OS, with a variety of user-options and is freely available for download from: `http://PRSice.info`.

2 R packages required

- ggplot2
- plyr
- batch
- fmsb
- gtx

N.B. Users will also need to install dependancies for these

3 Quickstart

As standard, PRSice has four input files and three main outputs:

Inputs

- PRSice.R file: includes R scripts, bash and PLINK2 wrappers
- PLINK2 executable files (Linux and Mac)
- Base data set: GWAS summary results, which the PRS is based on
- Target data set: Raw genotype data of 'target phenotype'

Outputs

- Figures illustrating the model fit of the PRS analyses
- Data on the model-fit of the PRS analyses
- PRS for each individual in the target data (for best-fit PRS as default)

Simple command for running PRSice on toy data provided:

```
R --file=PRSice_v1.23.R -q --args  \
  plink  ./plink_1.9_mac_160914 \
  base TOY_BASE_GWAS.assoc  \
  target TOY_TARGET_DATA \
  slower 0  \
  sinc 0.01 \
  supper 0.5
```

This runs PRSice with PRS calculated at P-value thresholds between $P_T = 0$ and $P_T = 0.5$ at increments of 0.01, where the target phenotype is a binary trait. Ancestry-informative covariates are calculated in the target data set and the first two used to adjust for population structure by default. Clumping is applied by default to thin SNPs according to linkage disequilibrium and P-value (the SNP with smallest P-value in each 250 kb window is retained and all those in LD, $r^2 > 0.1$, with this SNP are removed).

Below is a different run with more non-default options added (see Section 5 for the full list of user-options):

```
R --file=PRSice_v1.23.R -q --args  \
  plink  ./plink_1.9_mac_160914 \
  base TOY_BASE_GWAS.assoc  \
  target TOY_TARGET_DATA \
  slower 0  \
  sinc 0.02 \
  supper 0.3 \
  covary F \
  best.thresh.on.bar F \
  report.individual.scores T
```

This runs PRSice with PRS calculated between $P_T = 0$ and $P_T = 0.3$ at increments of 0.02, where the target phenotype is a binary trait. No covariates are included. Polygenic risk scores for each individual at the best-fit PRS are printed to a file `PRSice_SCORES_AT_BEST-FIT-PRS.txt`. The best-fit PRS bar will not be included in the bar plot.

3.1 Input Data

3.1.1 Base data set

Base (i.e. GWAS) data must be provided as a whitespace delimited file containing association analysis results for SNPs on the base phenotype. Columns with the following header names are essential and must be present: `SNP`, `A1`, `OR` or `BETA`, `P` - containing SNP name (eg. rs number), effect allele (A1), effect size estimate as an odds ratio (binary phenotype) or continuous effect beta (continuous phenotype) and P-value for association. Other fields, not essential for PRSice to run, can be included - chromosome number (`CHR`), base pair position (`BP`), reference allele (`A2`) and standard error (`SE`). Extra columns are allowed, but will be ignored. Please note, many studies label reference allele and effect allele `A1` and `A2` respectively; if this is the case, these column names need to be switched for PRSice to run correctly.

Thus, the first few lines of the base data set may look like the following for a binary trait:

```
SNP         CHR  BP      A1  A2 OR     SE         P
rs3094315   1    752566  A   G  0.9912 0.0229     0.7009
rs3131972   1    752721  A   G  1.007  0.0228     0.769
rs3131971   1    752894  T   C  1.003  0.0232     0.8962
```

or for continuous traits:

```
SNP         CHR  BP      A1  A2 BETA   SE        P
rs3094315   1    752566  A   G  -0.008 0.0229    0.7009
rs3131972   1    752721  A   G  0.0069 0.0228    0.769
rs3131971   1    752894  T   C  0.002  0.0232    0.8962
```

N.B. PRSice currently only supports autosomal SNPs.

Strand flips are automatically detected and accounted for.

3.1.2 Target Dataset

The target data set must be supplied in PLINK binary format, with the extensions `.bed` `.bim` `.fam` - where bed contains compressed genotype data. Missing phenotype data can be coded as NA, or -9 for binary traits.

3.1.3 PLINK2 Executable FIle

If the PLINK2 executable file is not in the working directory then the path to it must be given.

3.2 Outputs

3.2.1 Figures

- A bar plot named `PRSice_BARPLOT_[date].png` , where [date] is today's date, is generated, displaying the model fit of the PRS at different broad P-value thresholds (fig 1a).
- A high-resolution plot named `PRSice_HIGH-RES_PLOT_[date].png` , where [date] is today's date, displaying the model fit of PRS calculated at a large number P-value thresholds. A green line connects points showing the model fit at the broad P-value thresholds used in the corresponding bar plot (fig 1b).
- Quantile plots may also be optionally drawn to illustrate the effect of increasing score on predicted risk of phenotype. These will be named `PRSice_QUANTILES_PLOT.png` .

3.2.2 PRS model-fit

A file containing the PRS model fit across thresholds is named `PRSice_RAW_RESULTS_DATA.txt`; this is stored as threshold, P-value, variance in target phenotype explained, r^2, and number of SNPs at this threshold.

3.2.3 Scores for each Individual

A file containing PRS for each individual at the best-fit PRS named
`PRSice_SCORES_AT_BEST-FIT-PRS.txt` (by default) or
`PRSice_SCORES_AT_ALL_THRESHOLDS.txt` (as required).

Please note, as PRSice runs, a large number of temporary files are generated. If you have a small amount of storage space (or you use very high resolution on a very large data set), this could cause PRSice to fail. In this case, you may wish to reduce the number of thresholds used, or run PRSice on a larger disk.

N.B. PRSice also supports multiple base data sets and multiple phenotypes for target data, and draws heatmaps to describe these results. For more details on the options used to implement this, see below.

3.3 Summary-Summary Statistic Based Analysis

While the main function of PRSice is to apply polygenic risk scores that use effect size estimates from summary statistic data in a target data set containing raw genotype data, there is also the option to perform a summary-summary statistic analysis that exploits the gtx R package by Toby Johnson (gtx). An example command is shown below:

```
R --file=PRSice_v1.23.R -q --args  \
  plink  ./plink_1.9_mac_160914 \
  base TOY_BASE_GWAS.assoc  \
  target TOY_TARGET_GWAS.assoc \
  sumsum T \
  size.targ 2000 \
  slower 0  \
  sinc 0.02 \
  supper 0.3 \
  clump.snps F
```

This command includes the option to use genotype data for clumping by LD. See below for more details on the summary-summary analysis option.

4 User Options

Essential Arguments

- `target`: Location of the PLINK data used for the target data set. This must be supplied without the file's extension (i.e. `bed bim fam`)
- `base`: Location and full name of the GWAS results for the base data set - see the note above on column naming.

PLINK

- `plink`: The location and name of the binary file for executing plink. N.B. this is to be the more recent version of PLINK, '1.9' if the target data set is in genotype format, and the older version, '1.07' if dosage data is used. Default value is NA.

Target Data set Phenotype

- `pheno.file`: Location of the file containing external phenotype data, if it is not coded in the genotype file. This file must have two columns, individual ID and phenotype, with no header line. Missing data is coded NA, and NA and -9 for binary traits. If NA, phenotypes will be extracted from the genotype data. Default value is NA
- `binary.target`: If T, phenotype in target data is assumed to be binary. If F, phenotype in target data is assumed to be continuous. Default value is T.

Target Data set Format

- `geno.is.ped`: If `geno.is.ped T`, target data set will be converted from pedigree format (`.ped, .map`) to binary format before analysis begins. Default value is F
- `geno.as.list`: If `geno.as.list T`, the argument passed to target is treated as a general path to multiple genotype datasets, one per chromosome. In this case, every instance of the string `CHRNUM` will be substituted for values between 1-22 in order to analyse all chromosomes. Default value is F.

Covariates

- `covary`: If T, covariates are used when testing model fit of polygenic score on phenotype. If F, covariates are not used. Default value is T
- `user.covariate.file`: Location of file containing custom covariates. This must have column names individual ID (headed `IID`) and then covariates. If `covary T` but no user covariate file is supplied, ancestry-informative dimensions will be generated automatically from the data. Default value is NA

- `covariates`: Names of covariates to adjust for. If `covary T`, but `user.covariate.file NA`, these are the number of ancestry informative covariates to calculate and use, denoted `C1,C2,C3` etc. Default value is C1,C2
- `ancestry.dim`: If `covary T`, but `user.covariate.file NA`, ancestry informative dimensions are generated automatically. This specifies the method that should be used. Users can select PCA or MDS. Default value is MDS.

Figures

- `ggfig`: If T, ggplot2 will be used to generate figures. If F, base graphics will be used to generate figures. Default value is T
- `barchart.levels`: Thresholds which should be plotted on the bar chart. Default value is 0.001,0.05,0.1,0.2,0.3,0.4,0.5
- `barpalatte`: Colour palette used to colour bars on barplot, if `ggfig T` and `bar.col.is.pval F`. Default value is YlOrRd
- `best.thresh.on.bar`: If `fastscore F` and `best.thresh.on.bar T`, the most predictive threshold from high-resolution scoring will be identified automatically and added to the barchart. If `best.thresh.on.bar F`, only thresholds listed in barchart.levels will be plotted. Default value is F if `fastscore T`. If `fastscore F`, which is default, default value is T
- `scatter.R2`: If `scatter.R2 T`, variance explained will be used as y axis on the high-resolution plot. If `scatter.R2 F`, $-\log_{10}(P)$ will be used as y axis on high-resolution plot. If `fastscore T`, a high-resolution plot won't be generated. Default value is F
- `figname`: An optional prefix for figures and text-files. Default value is PRSice
- `bar.col.is.pval`: If `ggfig T`, and `bar.col.is.pval T`, bars are coloured by association with phenotype. If `bar.col.is.pval F`, bars are coloured by P-value threshold. Default value is F
- `bar.col.is.pval.lowcol`: If `bar.col.is.pval T`, `bar.col.is.pval.lowcol` will be used to colour the poorest predicting thresholds. Default value is dodgerblue
- `bar.col.is.pval.highcol`: If `bar.col.is.pval T`, `bar.col.is.pval.highcol` will be used to colour the best predicting thresholds. Default value is firebrick.

Clumping

- `clump.snps`: If `clump.snps T`, base SNPs will be clumped to remove linkage disequilibrium. If `clump.snps F`, base SNPs will not be clumped. This is currently unsupported with dosage data. Default value is T
- `clump.p1`: The clumping threshold, P-value, for index SNPs. Default value is 1

- `clump.p2`: The clumping threshold, P-value, for clumped SNPs. Default value is 1
- `clump.r2`: The LD threshold, r^2, for clumping. Default value is 0.1
- `clump.kb`: The distance threshold for clumping, in Kb. Default value is 250.

Pruning

- `prune.snps`: If `prune.snps T`, LD will be stripped using pruning, which is agnostic to base P-value. If `prune.snps F`, pruning will not be used. N.B. pruning and clumping can not both be used. Unless `clump.snps` is set to F, this will override `prune.snps T`. Default value is F
- `prune.kb.wind`: Window size for pruning, in Kb. Default value is 50
- `prune.kb.step`: Increment for sliding window, for pruning, in Kb. Default value is 2
- `prune.kb.r2`: Pairwise LD for pruning, as r^2. Default value is 0.8.

High-Resolution Scoring

- `slower`: The lower bound at which polygenic risk score is calculated at, as P_T. Default value is 0.0001
- `supper`: The upper bound at which polygenic risk score is calculated at, as P_T. Default value is 0.5
- `sinc`: The increment size between `slower` and `supper`. Polygenic risk scores will be calculated at each increment between the two bounds. Default value is 0.00005
- `fastscore`: If `fastscore T`, scores will only be calculated at the thresholds specified by `barchart.levels`. If `fastscore F`, scores will be calculated at incrementes of `sinc` between `slower` and `supper`. This is high-resolution scoring. Default value is F
- `mend.score`: If `mend.score T`, the first n SNPs will be added from the base data set, sorted by P-value, one by one in order to verify that there are no individual loci of large effect influencing target phenotype. If `mend.score F`, these will not be added. Default value is F
- `mend.score.len`: If `mend.score T`, a number of SNPs will be added one by one to calculate PRS. `mend.score.len` sets the number of SNPs to add before thresholds are used. Default value is 100
- `score.at.1`: If `score.at.1 T`, an extra threshold will be added at $P_T = 1$. This allows users to compare a polygenic risk score calculated with all SNPs to the next-best score. If `score.at.1 F`, this will not be calculated. Default value is F.

Dosage Data

- `dosage`: If `dosage T`, PRSice will expect to read in dosage data. If `dosage F`, PRSice will expect to read in genotype data. Default value is dosage F
- `dosage.format`: The format of dosage files can be specified to use default values for some of the options below. Currently only 'gen' is supported by this option. Default value is gen
- `dos.skip0`: The number of columns in dosage file before column containing SNP names. Default value is 1
- `dos.skip1`: The number of columns in dosage file between SNP names and A1. Default value is 1
- `dos.coding`: The value that dosage probabilities sum to per variant per individual. Default value is 1
- `dos.format`: Number of columns dosage data occumpies per variant per individual. Default value is 3
- `dos.sep.fam`: The path to an external fam file containing individual-level data for the dosage file. Default value is NA
- `dos.fam.is.samp`: If `dos.fam.is.samp T`, the fam file supplied will be automatically converted from gtool's 'sample' format before being read in. If `dos.fam.is.samp F`, the fam file will be read in like a normal 6-column PLINK fam file. Default value is F
- `dos.impute2`: If `dos.impute2 T`, dosage data will be read in using impute2 format defaults. If `dos.impute2 F`, these must be set manually. Default value is F
- `dos.path.to.lists`: If there are dosage files for each chromosome, with a regular pattern in which `CHRNUM` is a value between 1 and 22 for each dosage file, `dos.path.to.lists` can be set to a single string, e.g. `/path/to/chr_CHRNUM/chr_CHRNUM.dat`. Default value is NA
- `dos.list.file`: If dosage file names are listed in a separate file, these will be read in automatically from the external file. These file names must have one line per file. Default value is NA.

Multiple Phenotypes

- `multiple.target.phenotypes`: If `multiple.target.phenotypes T`, multiple columns will be read from the external phenotype file supplied using `pheno.file`. If this is used, the file `pheno.file` must have column names. The first column must be `ID`. The other columns must be names of phenotypes. If `multiple.target.phenotypes F`, only the first two columns from the phenotype file will be used. Default value is F

- `target.phenotypes`: A list of names of target phenotypes to be read from a phenotype file, if `multiple.target.phenotypes T`. These must be separated by commas. Polygenic risk score for base phenotype will be regressed on each of these in turn. Default value is NA

- `target.phenotypes.binary`: A vector of logical T and F's separated by commas. This determines whether the phenotypes listed by `target.phenotypes` are binary (i.e. case-control) or quantitative (i.e. continuous) traits. This must have the same number of items as `target.phenotypes`. NA, -9 and empty values will be treated as missing values for binary phenotypes, *binary phenotypes must be coded 0,1* NA and empty values will be treated as a missing value for quantitative traits. Default value is NA

- `multiple.base.phenotypes`: If `multiple.base.phenotypes T`, the argument passed to base will be interpreted as a general file path, in which the string `PHEN.NAME` will be evaluated as the names of different base phenotypes. Default value is F

- `base.phenotypes.names`: A vector of base phenotype names. These will be substituted for `PHEN.NAME` in the argument passed to base, and used to read multiple GWAS results in series.

Summary-Summary Statistic Based Analysis

- `sumsum`: If `sumsum T`, PRSice will expect both target and base datasets to be GWAS results files, and will use the method of Toby Johnson ([2] as implemented in the R package gtx) to evaluate evidence for shared genetic aetiology between base and target phenotypes. If using `sumsum T`, a number of changes need to be made to input formats - see below. If `sumsum F`, polygenic risk scores will be calculated, using genotype data, as decribed above. Default value is F

- `clump.ref`: If `sumsum T`, input data can be clumped to ensure linkage equilibrium. This is performed on the base data set, and is elected using `clump.snps T`. `clump.ref` is used to select genotype data to be used to clump the base data set. Default value is NA

- `size.targ`: The sample size of target data set. Default value is NA.

Quantile Plots

- `quantiles`: If `quantiles T`, a plot will be produced displaying association between target phenotype and PRS at the most predictive threshold, divided into a number of quantiles, as shown in figure 1c. If `quantiles F`, this plot will not be produced. (N.B. these will be adjusted for covariates unless `covary F`). Default value is F

- `num.quantiles`: Number of quantiles to split PRS into when producing quantile plot. Default value is 5

- `quant.ref`: Reference quantile to use when produing quantile plot. Default value is 3.

Miscellaneous

- `wd`: The folder which is to be used to store output data, temporary files and figures. Default value is ./ , i.e. current directory
- `print.time`: If `print.time T`, running time will be printed at end of output. If `print.time F`, running time will not be printed. Default value is T
- `cleanup`: If `cleanup T`, all temprorary files will be removed at the end of the analysis. If `cleanup F`, these will be left in the working directory. Default value is T
- `remove.mhc`: If `remove.mhc T`. the MHC region between 26 and 33Mb on chromosome 6 will be removed when calculating polygenic risk scores. If `remove.mhc F`, this region will not be removed. Default value is F. N.B., not valid for dosage data
- `for.meta`: If `for.meta T`, coefficients and standard errors are reported for each regression model, into the file `PRSICE_RAW_RESULTS_DATA.txt`. This allows meta-analysis of scores across different target data sets. If `for.meta F`, these extra columns are not reported. Default value is F
- `report.individual.scores`: If `report.individual.scores T`, a file called `PRSice_SCORES_AT_ALL_THRESHOLDS.txt` or `PRSice_SCORES_AT_BEST-FIT-PRS.txt` (see below) will be produced containing every individual's polygenic risk score at every threshold (or just the most predictive threshold) and written to the working directory. N.B., this file may be very large depending on the number of thresholds used. If `report.individual.scores F`, this file will not be written. Default value is T
- `report.best.score.only`: If `report.best.score.only F`, and `report.individual.scores T`, polygenic risk scores for all individuals at all thresholds will be written to a file called `PRSice_SCORES_AT_ALL_THRESHOLDS.txt`. If `report.best.score.only F`, scores for every individual at the most predictive threshold will be written to a file called `PRSice_SCORES_AT_BEST-FIT-PRS.txt`. Default value is T
- `plink.silent`: if `plink.silent T`, PLINK will not output to the terminal whilst PRSice is running. If `plink.silent F`, PLINK's outputs will print to the terminal as PRSice runs. Default value is T
- `no.regression`: if `no.regression T`, phenotype will not be regressed on polygenic risk scores, but they will be calculated and printed to a file, if `report.individual.scores T`. If `no.regression F`, phenotype will be regressed on polygenic risk scores. Default value is F

- `allow.no.sex`: If `allow.no.sex F`, individuals with missing sex data in the genotype file will be given missing phenotype data and excluded from regression models. If `allow.no.sex T`, individuals with missing sex data but non-missing genotype data will be included in regression models. Default value is F
- `debug.mode`: If `debug.mode T`, more text including warning and error messages from bash, R and PLINK will be reported to the terminal. If `debug.mode F`, these will be suppressed. Default value is F.

5 Detail on command line options

5.1 Target Data Set Format

If genotype data is stored chromosome by chromosome, this data can be analysed without merging, by using the option `geno.as.list T`. If this is the case, the string supplied to `target` must contain the characters `CHRNUM` - this will be iteratively replaced with chromosome numbers from 1 to 22 and scores calculated across the genome.

5.2 Target Data Set phenotype

By default, the software identifies binary phenotype, coded either 1,2 or 0,1 from a column in the `ped` or `fam` file. If the phenotype data stored in this file is continous, the option `binary.target F` must be used. If binary phenotype data is stored 1,2 then -9 and 0 will be coded as missing. If it is coded 0,1 then -9 will be coded as missing. NA will also be coded as missing.

If the phenotype data is stored in an external file, this file must have two columns, individual ID and phenotype, with no header line. This is specified with the option `pheno.file "/path/to/pheno.file"`.

5.3 Covariates

By default, the software calculates two ancestry informative dimensions using Multi-dimensional Scaling (MDS) and uses these as covariates when predicting target data set phenotype using polygenic score. Using covariates altogether can be disabled using `covary F`.

A different number of ancestry informative dimensions can be used by specifying `covariates C1,C2,C3`, for example to use three covariates. PCA or MDS can be used for calculating ancestry informative dimensions, using `ancestry.dim "PCA`" or `ancesrtry.dim "MDS"` respectively.

N.B. if covariates are automatically generated, output files should be inspected to check for outliers etc.

External covariates can also be used. These are added using the option `user.covariate.file "path/to/file.covary"`. This file must have a first column headed IID containing individual ID. The remaining columns must be covariates with a header line. The covariates to use from this file must be specified based on the column name, using the option `covariates name1,name2` etc.

5.4 Graphical Parameters

The default output is a barchart and a high-resolution plot (fig 1a) using ggplot2. Base graphics can be used instead using `ggfig F` - N.B., plotting with base graphics allows fewer options. The levels of P_T to use in the barchart are specified using `barchart.levels 0.1,0.2,0.3` , for example. If `fastscore F`, the most predictive polygenic score can be identified from the high-resolution plot and added to the barchart using `best.thresh.on.bar T`. If `ggfig T`, the user can specify the colour scheme used for the barchart using `barpalatte`. If high-resolution scoring is used, the y axis of the high-resolution plot is $-\log_{10}(P)$ by default - this can be switched to R^2 (or Nagelkerke's pseudo R^2 for binary target phenotypes) using `scatter.R2 T`

The extension name for all output files, which is `PRSice` by default, can be altered using `figname`.

If `ggfig T`, barplots are coloured by the predictive ability of that score on phenotype, $-\log_{10}(P)$, as a default. The colours of the barplot can be set by selecting the low and high colours to use for gradients using `bar.col.is.pval.lowcol` and `bar.col.is.pval.highcol`. Bars can be coloured by P-value threshold instead, using `bar.col.is.pval F`. In this case, `barpalatte` is used.

5.5 Linkage Disequilibrium

By default, base SNPs are clumped using base P-values and LD data from the target dataset, in order to obtain SNPs in linkage equilibrium. This option is disabled when using dosage data.

5.5.1 Clumping

Clumping may be disabled using `clump.snps F`.

Predefined clumping parameters can be changed by updating the `clump.p1` `clump.p2` `clump.r2` and `clump.kb` arguments.

5.5.2 Pruning

An alternative method for obtaining SNPs in linkage equilibrium is LD-informed pruning. The LD structure of the target data set is used to obtain SNPs in linkage equilibrium in the base dataset, and is agnostic of base P-value. This option is enabled using `prune.snps T`. N.B. if this is used, `clump.snps` must be set to `F`, as the two methods are mutually exclusive. If both are set to `T`, clumping will be used by default.

Pruning takes three arguments: the window size in Kb, the step size in Kb, and the r^2 to prune to within each window. These arguments are `prune.kb.wind`, `prune.kb.step`, and `prune.kb.r2` respectively, and default to 50, 2 and 0.8.

5.6 Dosage Data

Dosage data is not used by default and requires a number of defaults to run correctly. If using dosage format data, this must be specified using `dosage T`

Individual-level data must be provided using `dos.sep.fam "/path/to/file.fam"`. By defalt this must be in the format of a PLINK .fam file. Users may want to use the `.sample` files which are generated by impute2 - if so, this must be specified with `dos.fam.is.samp T` - then the input file can be reformatted correctly.

A target data set can be provided in three ways.

- A single file containing dosage data - this can be specified using `target "/path/to/dosages.dos"`
- A single filepath, where a certain part varies for dosage files across the 22 chromosomes. This is specified using `dos.path.to.lists "/path/to/chr_CHRNUM/chr_CHRNUM.impute"`. `CHRNUM` will be substituted for numbers 1-22 in order to read in autosome data.
- A file that contains the full paths and names of each dosage file, specified using `dos.list.file "/path/to/list_of_dos_files"`.

Other dosage options, indicating the format of the input data, are specified using `dos.skip0 dos.skip1 dos.coding` and `dos.format`. These options can be set to their defaults for reading in output from impute2 using `dos.impute2 T`.

5.7 High-Resolution Scoring

If `fastscore T`, polygenic scores will only be calculated at the levels specified by `barchart.levels`. If `fastscore F`, scores every few thresholds between a lower and upper bound will be calculated. By default, these are increments of 0.00005 from 0.0001 to 0.5 for genotype data. We recommend changing these to increments of 0.001 from 0.001 to 0.5 for dosage data in order to offset the increased running time for dosage files.

The size of these increments, and the upper and lower bounds used, can be set using the `sinc`, `slower` and `supper` options respectively.
`fastscore` allows a second figure to be produced, a high-resolution plot showing the result of high-resolution scoring, and allows users to add an extra bar to the barchart to indicate the most predictive threshold, using `best.thresh.on.bar T`.

5.8 Multiple Phenotypes

PRSice supports the analysis of pairwise comparisons between multiple base and multiple target phenotypes. Multiple target phenotypes must be stored in a single external phenotype file, specified using `pheno.file`. If `multiple.target.phenotypes T`, this file

must have column names ID and then phenotype names. The option to use multiple target phenotypes is selected using `multiple.target.phenotypes T`. Target phenotypes are selected from this file using `target.phenotypes`, a vector of names. Whether these are binary or quantitative traits must be specified using `target.phenotypes.binary`, a vector of logical TRUE's and FALSE's. `target.phenotypes.binary` must have the same number of items as `target.phenotypes`.

Multiple base phenotypes are specified using `multiple.base.phenotypes T`. If `multiple.base.phenotypes T`, the names of phenotyoes supplied to `base.phenotypes.names` will be substituted for `PHEN.NAME` in the argument passed to `base`. For example, if `base /path/to/PHEN.NAME.assoc` and `base.phenotype.names DIS1,DIS2`, then two files, `/path/to/DIS1.assoc` and `/path/to/DIS2.assoc` would be read in as base data sets.

If both multiple target and multiple base phenotypes are supplied, PRSice will perform pairwise tests of every base phenotype predicting every target phenotype at every threshold, and save the most predictive threshold from each comparison. For more detail on the outputs generated when using multiple base and target phenotypes see below.

5.9 Summary-Summary Statistic Based Analysis

PRSice can use GWAS summary data in both the base and target data sets to evaluate evidence for shared genetic aetiology, using the method of Johnson et al, as implemented in gtx ([2]). This is enabled using `sumsum T`.

If `sumsum T`, `target` must specify the path to a file containing GWAS summary results. This *must* contain columns labelled `SNP`, `SE`, `A1`, `A2`, for marker ID, standard error, effect allele, non-effect allele respectively. Users must also include a column for effect size, either `OR` for binary phenotypes or `BETA` for quantitative traits. The sample size of the target data set must also be set using `size.targ`.

The base data set *must* contain columns labelled `SNP`, `P`, `A1`, `A2` for marker ID, P-value, effect allele and non-effect allele respectively.

A sample of genotype data, e.g. HapMap (available here), can be used to clump the base GWAS data before testing for evidence for shared genetic aetiology. This is selected using `clump.ref`, where the path provided is the name of three plink binary files (i.e. .bed, .bim, .fam). N.B. The authors of gtx recommend using more stringent clumping parameters; these must be set manually. The authors recommend `clump.p1 0.5 clump.p2 0.5 clump.kb 300 clump.r2 0.05`.

When using `sumsum T`, many other options are no longer valid. The following options are still valid:

- `plink.silent`
- `print.time`
- `cleanup`
- `debug.mode`
- `slower`
- `sinc`
- `supper`
- `ggfig`
- `target`
- `base`
- `clump.ref`
- `clump.p1`
- `clump.p2`
- `clump.r2`
- `clump.kb`
- `plink`
- `binary.target`
- `size.targ`
- `figname`
- `bar.col.is.pval.lowcol`
- `bar.col.is.pval.highcol`
- `barchart.levels`

5.10 Miscellaneous

Running time is printed by default - this can be disabled using `print.time F`. Profile lists and supporting files are also removed from the working directory by default. This can be disabled using `cleanup F`.

By default all intermediate and output files are printed to the current directory. This can be changed using `wd "path/to/wd/"`.

The MHC region on chromosome 6 (26-33Mb) is frequently omitted from polygenic risk scores, as the long-range linkage disequilibrium in this region makes linkage equilibrium difficult to obtain. This region can be removed using `remove.mhc T`

The option `for.meta T` reports coefficients and standard errors for each regression model in the output file `PRSice_RAW_RESULTS_DATA.txt`, allowing meta-analysis across target data sets to be performed, at a given threshold.

6 Outputs

The script produces a large amount of supporting data. Two figures are generated - a barplot and a high-resolution plot of polygenic score threshold, P_T, plotted against model-fit across thresholds. The high-resolution plot depicts model-fit as $-\log_{10}(P)$. The barplot depicts model-fit as improvement in model-fit, as Nagelkerke's pseudo R^2, provided by adding polygenic score as a predictor to the model. The barplot appends the Wald test P-value for polygenic score above each bar.

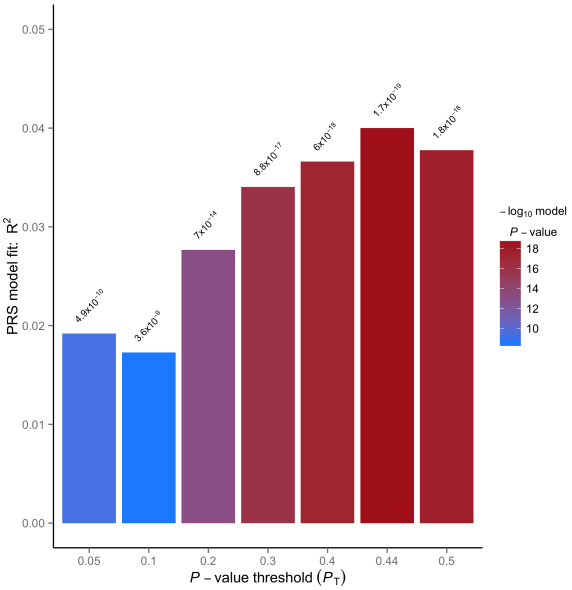
The high-resolution plot shows a green line that connects points corresponding to the model-fit at the broad P-value thresholds considered in the bar plot. This demonstrates the performance of high-resolution thresholds in comparison to the more traditional candidate thresholds.

The third output file contains polygenic scores for each individual at each threshold. This file is called `PRSice_SCORES_AT_ALL_THRESHOLDS.txt`. The P-value thresholds used for each column are specified in the column headings, and the first column is individuals' IDs.

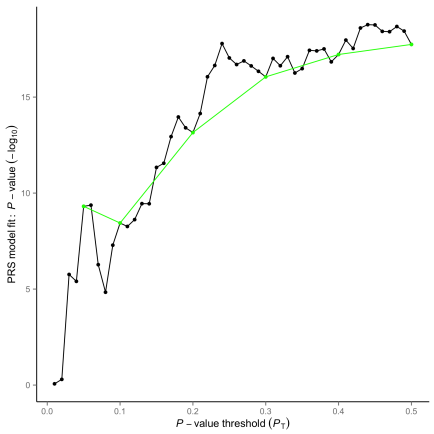
The fourth output file contains all the raw data used to generate the high-resolution plot. This is named `PRSice_RAW_RESULTS_DATA.txt` and is written to the working directory. It is in the following format:

```
thresh   p.out    r2.out       nsnps
1e-04    0.9138   44.8535e-06  46
0.0001   0.8689   11.1294e-05  47
2e-04    0.8689   11.1294e-05  51
0.0003   0.8689   11.1294e-05  67
```

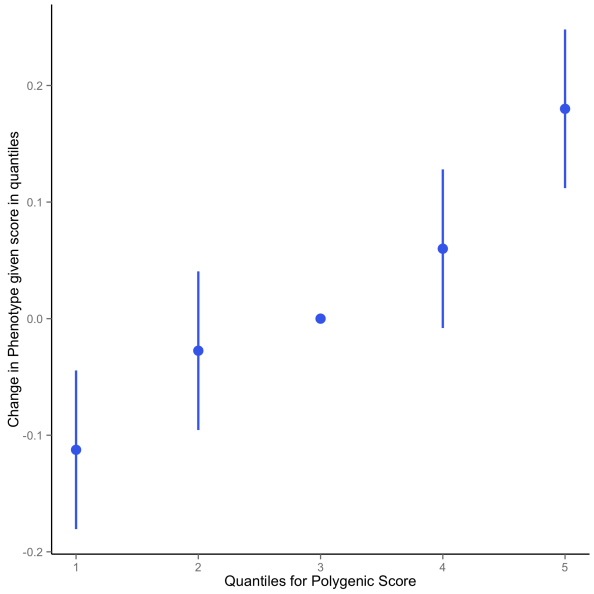
If multiple base and target phenotypes are used, additional files are generated. The file `PRSice_ALL_BEST_THRESHOLDS_BASE_AND_TARGET.txt` contains the P-value for the most predictive model between each pair of phenotypes in a matrix. This is also displayed visually in a heatmap called `PRSice_HEATMAP.png`. N.B. we reccomend using `ggfig T` for this, as heatmaps produced using base graphics are less readable.



(a) Bar plot generated by PRSice as default, demonstrating PRS-model fit across a small number of thresholds, P_T.



(b) High-resolution plot generated by PRSice as default, showing PRS model-fit over a large number of thresholds, P_T.



(c) Quantiles plot generated using `quantiles T` showing the effect of increasing score on risk of disease, at the most predictive threshold, P_T.

Figure 1: Figures generated by PRSice run at high-resolution. The best-fit bar, added to the bar plot, is calculated from a high-resolution run, and so is not available if the high resolution option is not used. The quantiles plot demonstrates that increasing PRS is associated with increasing odds of phenotype.

7 Acknowledgements

We would like to thank all beta testers for their help and invaluable suggestions:

- Jonathan Coleman
- Simone de Jong
- Eva Krapohl
- Niamh Mullins
- Stuart Newman (computing and cluster support)
- Robert Power
- Adam Socrates

References

[1] C. C. Chang, C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *ArXiv e-prints*, October 2014.

[2] Toby Johnson. *gtx: Genetics ToolboX*, 2013. R package version 0.0.8.