Sociogenomics & Polygenic Scores

PDHP begins our 2021 workshop series on March 16th, with a workshop entitled Sociogenomics & Polygenic Scores, co-presented by Ben Domingue of Stanford University’s Graduate School of Education and Erin Ware of the University of Michigan Population Neurodevelopment & Genetics Group. This half-day workshop is geared toward data analysts interested in combining social science and genetic analysis, and will provide information on the recent history of sociogenomics and a novel approach for examining gene-by-environment interactions, as well as hands-on practice with state-of-art techniques in the field (including creating polygenic scores from simulated plink data using a high-performance computing environment).

Topics include:

- Recent history of sociogenomics
- A novel approach for examining gene-by-environment interactions
- Hands-on introduction to high-performance computing and genetic data types
- Computation of polygenic scores using PRSice2 software

Please make sure you go to the PDHP website and

Download this document

Download Remote Desktop Connection file

Getting onto the lab computer

- Go to this google sheet and claim a login name
- https://docs.google.com/spreadsheets/d/1ID--Ch_kTRTvN0sTGvU5_W0KX6svJi5ewitNYH9sWN-k/edit?usp=sharing

This workshop is funded with support from the Population Dynamics Branch of the Eunice Kennedy Shriver National Institute of Child Health and Human Development. Grant support should be acknowledged in all publications using grant # P2CHD041028.
Contents

Sociogenomics & Polygenic Scores ................................................................. 1
Remote desktop connection ........................................................................ 3
A very quick intro to high power computing .............................................. 6
  How does command line work? ............................................................... 6
  Syntax (or structure) ................................................................................ 6
  Linux quirks ............................................................................................ 6
  Getting onto PuTTY ................................................................................ 7
Navigating and viewing folder contents .................................................... 9
Viewing files ............................................................................................ 9
The business of polygenic scores .............................................................. 10
Target dataset (your genetic data) ............................................................. 10
  Genotyped vs Imputed .......................................................................... 10
  Quality control (QC) steps ..................................................................... 10
Base data (summary statistics) ................................................................... 11
  Step 1: Identify a large, replicated GWAS from which to base your SNP weights ... 11
  Step 2: Obtain/download summary statistics from GWAS ....................... 12
  Step 3: Formatting your downloaded results ........................................... 13
Constructing the polygenic score ............................................................. 15
  Step 4: Decide what method/program you are going to use to create your PGS and create it. ... 15
Finally! A hands-on example .................................................................. 16
  Step 1: Identify a large, replicated GWAS from which to base your SNP weights ... 16
  Step 2: Obtain/download summary statistics from GWAS ....................... 16
  Step 3: Formatting your downloaded results ........................................... 17
  Step 4: Decide what method/program you are going to use to create your PGS and create it. ... 19
  Test run: getting everything going, and looking at some output .................. 20
  First Run: Let’s add some options to evaluate our phenotype file ............... 23
  Second Run: Only running the bar-chart levels, request all bar chart level PGSs, add quintile plot ............................................................... 28
  Third Run: Use the clumping defaults to see how the score changes .......... 30
Step 5: What do you do after you have a polygenic score? .......................... 32
Some reviews and commentaries – polygenic scores and precision medicine ... 36
PRSice options ....................................................................................... 37
Remote desktop connection
Why? To control the environment and make things go easier for lab. All programs are loaded, same workspace

You need to download the 2021-03-PDHP_Sociogenomics.rdp from the PDHP website. You'll have to unzip it

DO NOT GO ROGUE – USE THE .rdp FILE TO CONNECT!!

On a Mac:
If you don’t already have Microsoft Remote Desktop, open the App Store and type “Remote Desktop” in the search bar. Find Microsoft Remote Desktop 10 and install

On a PC:
Right click and edit. Under “User name:” add your assigned lab number:
   
isr\train##

*Go to this google sheet and claim a login name

Right click on your downloaded file and change some settings, if you want full screen

Double click it to connect

When you get to the credentials window, type in the password:
   
   Wrk*t0dAy031621 (those are all zeros)
Here is an optional setting that would be nice to have (remote desktop in full screen)

Once you hit connect, these are the screens you should see:
Since you populated your user name on the main Remote Desktop Connect screen, this should be updated and you’ll need to put in your password here: hp426@ISR

And you’re connected!!
A very quick intro to high power computing

How does command line work?

<table>
<thead>
<tr>
<th>Command line pros/cons</th>
<th>GUI pros/cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Fast!</td>
<td>- Mouse/keyboard makes things slow</td>
</tr>
<tr>
<td>+ Minimal system resources</td>
<td>- A lot of resources due to icons/fonts/video/mouse/etc</td>
</tr>
<tr>
<td>+ Once you learn command line, doesn’t change much</td>
<td>- New GUIs come out and you relearn everything</td>
</tr>
<tr>
<td>- A lot of memorizing code which can be difficult</td>
<td>+ Point and click is fairly simple</td>
</tr>
<tr>
<td>- Multitasking takes a bit of effort and requires some parallel computing skills</td>
<td>+ Multitasking is fairly easy with multiple windows</td>
</tr>
</tbody>
</table>

**Syntax (or structure)**

```bash
$ command -flag argument
```

Example

```bash
$ ls -l *.txt
```

the command "ls" is executed with the command line flag "-l" and all files in the current directory ending with ".txt" as arguments.

**Linux quirks**

Command line: *directory*   GUI: *folder*

* is the symbol for ‘wildcard’

Case sensitive!

If you’re typing out a file (or path), “Tab” will auto-complete the name

Cannot have carriage returns (enters) or extra spaces between flags/arguments/commands

In other words, your code must all be one-liners. Sometimes it’s hard to type this/see the commands. You can break a line with a \ and an enter without interrupting the code

Copy-Paste as you know it will not work. To copy in Linux, just highlight text. To Paste, just right click
Getting onto PuTTY

Double click on the putty icon (or find putty.exe on your computer)
You will be presented with the following screen:

Host Name (or IP address): genomics.ad.isr.umich.edu
Port: 22
Connection type: SSH

This is a saved profile and you will be using!

Click Open and you will be presented with the following screen:

TIPS: Command line is case sensitive, be careful what you type!
Type in your user name and hit enter

```
login as:  
```

Type in your password. *NOTE, you will not be able to see characters or *****

You have now logged onto the main node for the ISR servers (see below).

**Home screen:**

```
login as: ebakshis
```

---

NOTICE TO USERS

This computer system is owned by the Institute for Social Research at the University of Michigan. It is for authorized use only. System usage is carefully monitored for compliance.

By your use of these resources, you agree to abide by Proper Use of Information Resources, Information Technology, and Networks at the University of Michigan (SPG 601.07), in addition to all relevant state and federal laws. [http://spg.umich.edu/policy/601.07](http://spg.umich.edu/policy/601.07)

---

```
Last login: Tue Feb 23 13:33:49 2021 from dhcp-21-032.private.isr.umich.edu
```

```
ebakshis@genomics ~$  
```
Navigating and viewing folder contents

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>ls</code></td>
<td>List commands: lists all files in a directory.</td>
</tr>
<tr>
<td><code>ls -l</code></td>
<td>List commands: lists all files in a directory WITH details</td>
</tr>
<tr>
<td><code>cd &lt;absolute/relative path&gt;</code></td>
<td>Change directory</td>
</tr>
<tr>
<td><code>cd ..</code></td>
<td>Move up a directory</td>
</tr>
<tr>
<td><code>cd ../..</code></td>
<td>Move up three directories</td>
</tr>
<tr>
<td><code>cd ~</code></td>
<td>Go directly home!</td>
</tr>
<tr>
<td><code>pwd</code></td>
<td>List what directory you are in (print working directory)</td>
</tr>
</tbody>
</table>

A **relative** path is defined from where you are currently located: i.e. `../newfolder/`

An **absolute** path is defined from the home directory: i.e. `/home/users/newfolder/`

**Viewing files**

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>wc -l &lt;filename&gt;</code></td>
<td>Search for how many rows are in a file (includes headers)</td>
</tr>
<tr>
<td><code>head &lt;filename&gt;</code></td>
<td>Show me the first 10 lines of a file</td>
</tr>
<tr>
<td><code>tail &lt;filename&gt;</code></td>
<td>Show me the last 10 lines of a file</td>
</tr>
<tr>
<td><code>more &lt;filename&gt;</code></td>
<td>Show me the whole file, one screen at a time</td>
</tr>
<tr>
<td><code>less &lt;filename&gt;</code></td>
<td>Show me the whole file, allows backward scroll</td>
</tr>
<tr>
<td><code>grep &lt;word&gt; &lt;file&gt;</code></td>
<td>Essentially a Ctrl F (find) function</td>
</tr>
</tbody>
</table>

When using `more/less`: `<enter>` will advance the preview, `q` will quit the preview

When using `head/tail`: `-n ##` will specify how many lines to view
The business of polygenic scores

Is it polygenic score (PGS), polygenic risk score (PRS, or PGRS), genetic risk score (GRS), polygenic index (PGI)?

Honestly, it’s up to you. The words “score” and “index” are loaded. People will always think a high score/index is good and a low score/index is bad. My two cents: I don’t always work with outcomes where it is “risky” to have a high value (is being tall risky? Is having a high education risky? versus other phenotypes where this is more obvious: having a high HbA1c, high BMI, etc. may be actually risky), and the direction of the polygenic score is dependent on the calculation. I choose to use polygenic score (PGS).

Target dataset (your genetic data)

<table>
<thead>
<tr>
<th>.bed/.bim/.fam files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotyped?</td>
</tr>
<tr>
<td>Quality control steps</td>
</tr>
<tr>
<td>File types</td>
</tr>
<tr>
<td>.bed/.bim/.fam .ped/.map .gen .gprobs .bim</td>
</tr>
<tr>
<td>plink.bed (binary file, genotype information)</td>
</tr>
<tr>
<td>plink.fam (family and individual ID, paternal and maternal ID, sex, phenotype)</td>
</tr>
<tr>
<td>plink.bim (CHR, SNP, BP, CM: allele names (2))</td>
</tr>
</tbody>
</table>

Genotyped vs Imputed

Which should you choose? Up to you! Not much gain to using imputed data, especially if you do heavy pruning/clumping. Will likely have more variants in your score with imputed data. As GWAS moves into the future, they are requesting contributing studies to impute to the Haplotype Reference Consortium (HRC) – if you use HRC imputed data, you will have a great deal of overlap. Programs are flexible and can handle genotyped and/or imputed.

Quality control (QC) steps

A great resource for genomic quality control


A tutorial for genomic quality control

https://choishingwan.github.io/PRS-Tutorial/target/

Sample QC tasks include checking for:

1. discordant sex information (Assumptions: you have both self-identified sex and X/Y chrom)
2. Individual missingness
3. heterozygosity scores
4. relatedness

SNP QC tasks include checking:

1. minor allele frequencies
2. SNP missingness
3. differential missingness (Assumptions: Case-control status has been specified in the .fam file)
4. Hardy Weinberg Equilibrium deviations

Formatting of your data

1. Make sure the alleles are as you want them to be
   a. (Major allele? Minor allele? Aligned to some consortia? Alphabetical?)
2. Platform-specific names SNP_id (kgp) converted to rs_id
3. Make sure the strand is aligned
4. Make sure the build is updated
5. Ambiguous SNPs
6. Mismatching SNPs: N.B. Most PRS software will perform strand-flipping automatically, thus this step is usually not required.

Base data (summary statistics)

Step 1: Identify a large, replicated GWAS from which to base your SNP weights

- Identify from genomic literature (PubMed, Google scholar, Nature publications, etc.)
- GWAS catalog (https://www.ebi.ac.uk/gwas/)
Step 2: Obtain/download summary statistics from GWAS

You need to choose which summary statistics you are downloading.

- LD hub (hosted by the Broad Institute) [http://ldsc.broadinstitute.org/gwashare/](http://ldsc.broadinstitute.org/gwashare/)
  - Many summary statistics, identified by ancestry, links to PMID

- GWAS catalog!
  - ftp site to download
  - There are ~3500 available summary statistics from published data
  - There are ~4500 available summary statistics from unpublished/prepublished data

- Individual consortium

<table>
<thead>
<tr>
<th>Consortium</th>
<th>Full consortium name</th>
<th>Summary statistics link</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALSKP</td>
<td></td>
<td><a href="http://alskp.org/informational/data">http://alskp.org/informational/data</a></td>
</tr>
<tr>
<td>CKDGen</td>
<td>Chronic Kidney Disease Genetics Consortium</td>
<td><a href="http://ckdgen.imbi.uni-freiburg.de">http://ckdgen.imbi.uni-freiburg.de</a></td>
</tr>
</tbody>
</table>
Contact the authors of an article
- Mixed results… “Our analyst left… the results are on our old server… why do you want them”. Fair amount of ghosting
- May need to write an analysis plan and have it approved by the consortium
- If the study you’re analyzing is IN the summary statistics that are available, you may need to request they re-run their meta-analysis WITHOUT your study [Social Science Genomic Analysis Consortium is good at this]
- Nice to have the power of a contributing study behind you, but not always possible and doesn’t always work

Step 3: Formatting your downloaded results

Huge heterogeneity in what is available in each summary statistic file

Buyer beware. READ THE README FILE!!!

At a MINIMUM: SNP, effect (beta/OR), P-value, effect allele
Preferred labels: SNP, BETA/OR, P, A1

Some examples of headers:

- snp effect_allele other_allele maf effect stderr pvalue
- SNPID CHR POS A1 A2 Freq_HapMap Zscore Pvalue
- Chromosome Position MarkerName Effect_allele Non_Effect_allele Beta SE Pvalue
- Marker Chrom Pos Allele1 Allele2 Ncases Ncontrols GC_Pvalue Overall
- SNP IDCHR BP Allele1 Allele2 Freq1 Effect StdErr P.value TotalN
- SNP CHR BP A1 A2 OR SE P INFO EUR_FRQ
- rsID, allele1, allele2, freqA1, beta, se, pval, N
- Marker Chr Position PValue OR(MinAllele) LowerOR UpperOR Alleles(Maj>Min)
- Chr Position Allele1 Allele2 Freq1 Pvalue EffN

Some programs to create polygenic scores (e.g. LDpred) require much more information from the GWAS that is the sources of the weights (e.g. non-coded allele, standard error, base position, etc.)

Note on Genome Build

Many freshly downloaded GWAS summary stats also only contain SNP ID and not chromosome number or base pair positions. In some cases, some old GWASs before 2012 use HG18 (NCBI B36) for base pair positions. For these type of data, one has to match the SNP IDs against the reference panel legend to find out the chromosome number and base pair positions.

Note on Sample Size

Note that some GWASs report the total sample size, which includes samples both in the discovery stage, and samples in the replication stage. However, it’s often the case that sample size of the discovery GWAS stage is the one that matches the data.

Constructing the polygenic score

Step 4: Decide what method/program you are going to use to create your PGS and create it.

Consider this carefully and report all decisions in your methods.

Common programs:

PLINK: not a dedicated PRS software, however, you can perform every required steps of the clumping/thresholding approach. Breaks down the processes which are involved in computing the polygenic scores that are generally “black box” style with polygenic score programs.

Required: QCed summary statistics, genotype/imputed data

Optional: covariate (+ genetic principal components), phenotype files

PRSice2: a dedicated polygenic score program that wraps R and plink functions. Includes some easy-to-forget QC steps. Implements the clumping and thresholding method.

Required: QCed summary statistics, genotype/imputed data

Optional: covariate (+ genetic principal components), phenotype files

LDpred-2: an R package that uses a Bayesian approach to polygenic scoring

Required: QCed summary statistics (with additional columns), genotype/imputed data, reference genome file that contains the LD structure matching your target data ancestry

Optional: covariate (+ genetic principal components), phenotype files

lassosum: a dedicated polygenic score program which is an R package that uses penalized regression (LASSO).

Required: QCed summary statistics, genotype/imputed data

Optional: covariate (+ genetic principal components), phenotype files
Finally! A hands-on example

Assuming you have QCed genetic data for your target sample, in plink format (.bed/.bim/.fam) and the study is not in the GWAS you’re using for your base data…

**Research question**: What is the association between a polygenic score for BMI based off the GIANT BMI and UK BioBank meta-analysis summary statistics and measured BMI in the PDHP data set?

**Step 1: Identify a large, replicated GWAS from which to base your SNP weights**


**Step 2: Obtain/download summary statistics from GWAS**

![GIANT consortium data files](https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files#BMI_and_Height_GIANT_and_UK_BioBank_Meta-analysis_Summary_Statistics)
Which file do I pick? Depends on what you want to do… We specifically want the GIANT+UKBiobank summary stats for BMI (Locke et al), so we will download the first one.

Clicking on that link will download a file called:

```
Meta-analysis_Locke_et_al+UKBiobank_2018_UPDATED.txt.gz
```

**Step 3: Formatting your downloaded results**

You would need to unzip it [Already done!!]

```
$ gunzip -d Meta-analysis_Locke_et_al+UKBiobank_2018_UPDATED.txt.gz
```

Then take a peek at what’s in the unzipped file [Already done!!]

```
$ head Meta-analysis_Locke_et_al+UKBiobank_2018_UPDATED.txt
```

How many lines are in this file? [Already done!!]

```
$ wc -l Meta-analysis_Locke_et_al+UKBiobank_2018_UPDATED.txt
2336270 - 1 = 2336269
```
What do all the headers mean?! (Look at the ReadMe file)

## This describes the columns of the summary statistics generated in Yengo et al. (2018)

## Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry

----------------------
Columns description
----------------------

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>RS ID</td>
</tr>
<tr>
<td>CHR</td>
<td>Chromosome</td>
</tr>
<tr>
<td>POS</td>
<td>Physical position (Genome build hg19)</td>
</tr>
<tr>
<td>Tested_Allele</td>
<td>Allele corresponding to the effect size (BETA/BETA COJO)</td>
</tr>
<tr>
<td>Other_Allele</td>
<td>Other allele</td>
</tr>
<tr>
<td>Freq_Tested_Allele_in_HRS</td>
<td>Frequency of the tested allele in the Health and Retirement Study (from 8,552 unrelated participants).</td>
</tr>
<tr>
<td>BETA</td>
<td>Marginal SNP effect size.</td>
</tr>
<tr>
<td>SE</td>
<td>Standard error of the effect size.</td>
</tr>
<tr>
<td>P</td>
<td>P-value measuring the significance of the marginal effect.</td>
</tr>
<tr>
<td>N</td>
<td>Sample size.</td>
</tr>
</tbody>
</table>


We need to change “Tested_Allele” and “Other_Allele” to be “A1” and “A2”. One way to do this quickly is to use a Linux “sed” function. It is exactly like find-replace. [Already done!!]

## This will replace “Tested_Allele” with “A1” anywhere in the document and OVERWRITE the original file (-i)

```
$ sed -i 's/Tested_Allele/A1/g' Meta-analysis_Locke_et_al+UKBiobank_2018_UPDATED.txt
```

## This will replace “Other_Allele” with “A2” anywhere in the document and write to a new file (>) called Yengo_meta_analysis_summarystats_BMI.txt

```
$ sed 's/Other_Allele/A2/g' Meta-analysis_Locke_et_al+UKBiobank_2018_UPDATED.txt > Yengo_meta_analysis_summarystats_BMI.txt
```
Your turn:

In your terminal window, let’s change directory into our PDHP_lab folder

```
$ cd /home/train##/PDHP_Lab/data
```

What files are in there?

```
$ ls /home/train##/PDHP_Lab/data
```

Take a look at your summary stats file:

```
$ head Yengo_meta_analysis_summarystats_BMI.txt
```

```
[ebakshis@genomics data]$ head Yengo_meta_analysis_summarystats_BMI.txt
SNP CHR POS A1 A2 Freq_A1_in_HRS BETA SE P N
rs1000000 12 126890980 A G 0.2219 1e-04 0.0021 0.96 689928
rs1000002 1 183635768 T C 0.5086 -1e-04 0.0016 0.94 783519
rs1000003 3 95733906 T G 0.5817 -0.0047 0.0018 0.0072 676691
rs1000004 4 98342907 A G 0.8404 0.0029 0.0024 0.23 690549
rs1000005 4 38924330 A G 0.2516 6e-04 0.002 0.75 691768
rs1000006 4 165621955 T G 0.8555 6e-04 0.0025 0.81 689797
rs1000007 2 5254744 C G 0.149 -8e-04 0.0025 0.74 691547
rs1000008 2 237752054 T C 0.7218 0.0024 0.0019 0.2 688538
```

How many SNPs are in this file?

```
$ wc -l Yengo_meta_analysis_summarystats_BMI.txt
```

```
[ebakshis@genomics data]$ wc -l Yengo_meta_analysis_summarystats_BMI.txt
723219 Yengo_meta_analysis_summarystats_BMI.txt
```

We were expecting 2336269, but there are 723218 SNPs.

**Note**, to reduce the size of the file and thus the computation load for this lab, I have also filtered out any SNPs from the base file that were not in our target file**

**Step 4: Decide what method/program you are going to use to create your PGS and create it.**

We will be using PRSice-2: [https://www.prsice.info/step_by_step/](https://www.prsice.info/step_by_step/)

Nice things PRSice-2 does for you:
• Align alleles (if the effect allele in the base file is A (alt G) and in the target data is G (alt A), automatically flips to make them the same)
• Strand flips are automatically detected and accounted for.

There are a LOT of options to specify. See PRSice options at the end of this document

---

**Test run: getting everything going, and looking at some output**

```
Rscript PRSice.R \
--prsice PRSice_linux \ 
--target pdhp_geno \ 
--base Yengo_meta_analysis_summarystats_BMI.txt \ 
--out test \ 
--binary-target F \ 
--extract test.valid \ 
--no-regress \ 
--no-clump
```

**Syntax explained**

Rscript  
Tells Linux I’m going to be running R code
--prsice  
Where is the PRSice_linux file located
--target  
Where are the target .bim/.bed/.fam files located
--base  
Where is the formatted base file located
--out  
What should the program append at the beginning
--binary-target  
T/F (looking for BETA or OR)
--extract  
If there are any funky (tri-allelic) SNPs, don’t include them and use this as a list for the good SNPs
--no-regress  
Don’t try and perform any regressions
--no-clump  
Don’t clump or prune the data

**Log output**

PRSice 2.3.3 (2020-08-05)
https://github.com/choishingwan/PRSice
(C) 2016-2020 Shing Wan (Sam) Choi and Paul F. O'Reilly
GNU General Public License v3
If you use PRSice in any published work, please cite:
Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score Software for Biobank-Scale Data. GigaScience 8, no. 7 (July 1, 2019)
2021-03-11 12:39:48

20
./PRSice_linux \
   --a1 A1 \
   --a2 A2 \
   --bar-levels 0.001,0.05,0.1,0.2,0.3,0.4,0.5,1 \
   --base Yengo_meta_analysis_summarystats_BMI.txt \
   --binary-target F" \
   --chr CHR \
   --extract test.valid \
   --interval 5e-05 \
   --lower 5e-08 \
   --no-clump \
   --no-regress \
   --num-auto 22 \
   --out test \
   --pvalue P \
   --seed 1804366554 \
   --snp SNP \
   --stat BETA \
   --target pdhp_gen0 \
   --thread 1 \
   --upper 0.5

Initializing Genotype file: pdhp_gen0 (bed)

Start processing Yengo_meta_analysis_summarystats_BMI
===============================================

SNP extraction/exclusion list contains 5 columns, will assume
first column contains the SNP ID

Base file: Yengo_meta_analysis_summarystats_BMI.txt
Header of file is:
SNP CHR POS A1 A2 Freq_A1_in_HRS BETA SE P N

Reading 100.00%
723218 variant(s) observed in base file, with:
14109 variant(s) excluded based on user input
709109 total variant(s) included from base file

Loading Genotype info from target
===============================================

80 people (28 male(s), 52 female(s)) observed
80 founder(s) included

14108 variant(s) not found in previous data
1 variant(s) with mismatch information
709108 variant(s) included

Start calculating the scores
**NOTE** 14109 variant(s) excluded based on user input – these are variants that were excluded because of either ambiguous SNPs or triallelic variants

Files that are output:

- **test.all_score**
  - File with every calculated PGS + FID and IID
  - FID and IID, 242 different scores, pT (5e-08 to 0.5, by 5e-05) and at bar levels of 0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1

- **test.log**
  - Saved version of the PRSice output

- **test.mismatch**
  - Listing of variants with the same name and different information (CHR/BP)

- **test.prsice**
  - Number of SNPs at different p-value thresholds in the base data set

## To view the files

```
$ head Run1.log
$ head Run1.mismatch
$ head test.prsice
```
First Run: Let’s add some options to evaluate our phenotype file

Rscript PRSice.R \
--prsice PRSice_linux \
--target pdhp_geno \
--base Yengo_meta_analysis_summarystats_BMI.txt \
--out Run1 \
--binary-target F \
--extract test.valid \
--pheno-file pdhp_phenocov.txt \
--pheno-col BMI \
--cov-file pdhp_phenocov.txt \
--cov-col sex,age,age2,@PC[1-5] \
--no-clump

Syntax explained

Rscript          Tells Linux I’m going to be running R code
--prsice        Where is the PRSice_linux file located
--target        Where are the target .bim/.bed/.fam files located
--base          Where is the formatted base file located
--out           What should the program append at the beginning
--binary-target T/F (looking for BETA or OR)
--extract       If there are any funky (tri-allelic) SNPs, don’t include them and use this as a list for the good SNPs
--pheno-file    What is the location and name of the phenotype file
--pheno-col     What is the title of the phenotype column
--cov-file      What is the location and name of the covariate file
--cov-col       Which columns do you want to include for covariates - no space, separated by commas
--no-clump      Don’t clump or prune the data
Log output

PRSice 2.3.3 (2020-08-05)
...

./PRSSice_linux \
   --a1 A1 \
   --a2 A2 \
   --bar-levels 0.001,0.05,0.1,0.2,0.3,0.4,0.5,1 \n   --base Yengo_meta_analysis_summarystats_BMI.txt \n   --binary-target F\n   --chr CHR \n   --cov pdhp_phenocov.txt \n   --cov-col sex,age,age2,@PC[1-5] \n   --extract test.valid \n   --interval 5e-05 \n   --lower 5e-08 \n   --no-clump \n   --num-auto 22 \n   --out Run1 \n   --pheno pdhp_phenocov.txt \n   --pheno-col BMI \n   --pvalue P \n   --seed 1168992417 \n   --snp SNP \n   --stat BETA \n   --target pdhp_genoc \n   --thread 1 \n   --upper 0.5

Initializing Genotype file: pdhp_genoc (bed)

Start processing Yengo_meta_analysis_summarystats_BMI
==============================================

SNP extraction/exclusion list contains 5 columns, will assume first column contains the SNP ID

Base file: Yengo_meta_analysis_summarystats_BMI.txt
Header of file is:
SNP CHR POS A1 A2 Freq_A1_in_HRS BETA SE P N

Reading 100.00%
723218 variant(s) observed in base file, with:
14109 variant(s) excluded based on user input
709109 total variant(s) included from base file
Loading Genotype info from target

80 people (28 male(s), 52 female(s)) observed
80 founder(s) included

14108 variant(s) not found in previous data
1 variant(s) with mismatch information
709108 variant(s) included

Phenotype file: pdhp_phenocov.txt
Column Name of Sample ID: FID+IID
Note: If the phenotype file does not contain a header, the column name will be displayed as the Sample ID which is expected.

There are a total of 1 phenotype to process

Processing the 1 th phenotype

BMI is a continuous phenotype
80 sample(s) with valid phenotype

Processing the covariate file: pdhp_phenocov.txt

Include Covariates:
Name    Missing Number of levels
age     0    -
sex     0    -
age2    0    -
PC1     0    -
PC2     0    -
PC3     0    -
PC4     0    -
PC5     0    -

After reading the covariate file, 80 sample(s) included in the analysis

Start Processing
Processing 100.00%
There are 1 region(s) with p-value between 0.1 and 1e-5 (may not be significant).

Begin plotting
Current Rscript version = 2.3.3
Files that are output:

**Run1.best**  
File with only the “best” PGS, FID and IID

```
FID IID In_Regression PRS
pdhp1 pdhp1 Yes  -2.02513774e-05
pdhp2 pdhp2 Yes  -2.09882783e-05
pdhp3 pdhp3 Yes  -1.88086778e-05
pdhp4 pdhp4 Yes  -3.5446279e-05
```

**Run1.log**  
Saved version of the PRSice output

```
PRSice 2.3.3 (2020-08-05)
https://github.com/choishingwan/PRSice
(C) 2016-2020 Shing Wan (Sam) Choi and Paul F. O'Reilly
GNU General Public License v3
If you use PRSice in any published work, please cite:
Choi SW, O'Reilly PF.
PRSice-2: Polygenic Risk Score Software for Biobank-Scale Data.
GigaScience 8, no. 7 (July 1, 2019)
2021-03-11 12:39:48
```

**Run1.mismatch**  
Listing of variants with the same name and different information (CHR/BP)

<table>
<thead>
<tr>
<th>File_Type</th>
<th>RS_ID</th>
<th>CHR_Target</th>
<th>CHR_File</th>
<th>BP_Target</th>
<th>BP_FileA</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_Target</td>
<td>A1_Value</td>
<td>A2_Value</td>
<td>A2_Value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>rs2192400</td>
<td>0</td>
<td>0</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

**Run1.prsice**  
Number of SNPs at different p-value thresholds in the base data set **PLUS** R2 value, p-value for association, beta coefficient and standard error for the effect of PGS on outcome

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Set</th>
<th>Threshold</th>
<th>R2</th>
<th>P</th>
<th>Coefficient</th>
<th>Standard.Error</th>
<th>Num_SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Base</td>
<td>5e-08</td>
<td>0.00860833</td>
<td>0.393671</td>
<td>3142.99</td>
<td>3664.83</td>
<td>10469</td>
<td></td>
</tr>
<tr>
<td>- Base</td>
<td>5.005e-05</td>
<td>0.0174835</td>
<td>0.22272</td>
<td>9490.71</td>
<td>7723.07</td>
<td>28930</td>
<td></td>
</tr>
<tr>
<td>- Base</td>
<td>0.00010005</td>
<td>0.0179673</td>
<td>0.216351</td>
<td>10617.1</td>
<td>8520.01</td>
<td>33152</td>
<td></td>
</tr>
<tr>
<td>- Base</td>
<td>0.00015005</td>
<td>0.0191264</td>
<td>0.201936</td>
<td>11689.9</td>
<td>9085.71</td>
<td>35738</td>
<td></td>
</tr>
</tbody>
</table>

**Run1.summary**  
Summary information for the “best” polygenic score including the threshold, PRS R2, full R2, null R2, coefficient, standard error, P-value and number of SNPs in the score

```
## To view the files

```bash
$ head Run1.best
$ head Run1.log
$ head Run1.mismatch
$ head Run1.prsice
```
A bar plot of incremental R2 at different p-value thresholds (specified by bar-levels), colored by log10 model p-value with the p-value written above the bar.

A scatter plot of the p-value threshold and –log10 p-value for the model fit with a trend line.

These new plots may be helpful for choosing a p-value threshold.

BUT – remember you’ve done 242 tests… and choosing the best pT for your data may not be the same in other studies, so the choice may not be replicable.
Second Run: Only running the bar-chart levels, request all bar chart level PGSs, add quintile plot

Rscript PRSice.R \
--prsice PRSice_linux \
--target pdhp_geno \
--base Yengo_meta_analysis_summarystats_BMI.txt \
--out Run2 \
--binary-target F \
--fastscore \
--print-snp \
--quantile 10 \
--extract test.valid \
--all-score \
--pheno-file pdhp_phenocov.txt \
--pheno-col BMI \
--cov-file pdhp_phenocov.txt \
--cov-col sex, age, age2, @PC[1-5] \
--no-clump

Syntax explained

Same as above except ➔

--fastscore Only calculates scores for the bar-chart levels 0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1
--print-snp Creates a file that lists all the SNPs used in the best score
--quantile Creates a quantile plot for the best score in whatever number of specified quantiles
--all-score Creates an output file for all scores specified (in this case only the pTs for the bar-chart levels)

Log output

Same as above except ➔

Begin plotting

Current Rscript version = 2.3.3
Plotting the quantile plot
Plotting Bar Plot
Files that are output:

Same as above except:

Run2.all_score
File with every calculated PGS + FID and IID, similar to test.all_score above, but this file only contains scores at the bar levels of 0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1, along with FID and IID.

Run2.snp
A file containing all SNPs with CHR, BP, P-value and an indicator for SNPs that are included in the ‘best’ score.

<table>
<thead>
<tr>
<th>CHR</th>
<th>SNP</th>
<th>BP</th>
<th>P</th>
<th>Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs2272908</td>
<td>1721479</td>
<td>4.6e-16</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>rs3737628</td>
<td>1722932</td>
<td>4.3e-16</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>rs9660180</td>
<td>1723031</td>
<td>1.4e-17</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>rs10907185</td>
<td>1733219</td>
<td>1.1e-10</td>
<td>1</td>
</tr>
</tbody>
</table>

Run2_QUANTILES_2021_03_12.txt
A text file containing the quantile information, coefficient, confidence limits, group and number of individuals.

Run2_QUANTILES_PLOT_2021_03_12.png
Quantile plot with the middle quantile as the reference. Y axis shows change in phenotype given score in quantile versus quantile of polygenic score.
Third Run: Use the clumping defaults to see how the score changes

```
Rscript PRSice.R \
--prsice PRSice_linux \
--target pdhp_geno \
--base Yengo_meta_analysis_summarystats_BMI.txt \
--out Run3 \
--binary-target F \
--print-snp \
--quantile 10 \
--extract test.valid \
--pheno-file pdhp_phenocov.txt \
--pheno-col BMI \
--cov-file pdhp_phenocov.txt \
--cov-col sex,age,age2,@PC[1-5]
```

**Syntax explained**

Same as above except ➔

We removed the --no-clump flag in the code. The scores that are created will be clumped in a distance of 250kb, with an R2 threshold of 0.1 and a p-value threshold of 1 (the default settings).

We also removed the --all-score and --fastscore flags

**Log output**

Same as above except ➔

Start performing clumping

Clumping Progress: 100.00%

Number of variant(s) after clumping : 96510

**Files that are output:**

Same as above
Let’s compare outputs:

From Run 1 and 2 – no clumping at a p-value threshold of 1

From Run 3 – default clumping at a p-value threshold of 1
Step 5: What do you do after you have a polygenic score?

Depends... What’s your research question?

**Research question:** What is the association between a polygenic score for BMI based off the GIANT BMI and UK BioBank meta-analysis summary statistics and measured BMI in the PDHP data set?

In your terminal (PuTTy) window, make sure you are in the data folder

```bash
$ cd /home/train###/PDHP_Lab/data
Type in a capital letter R and hit enter
$ R
```

R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

```r
## Read in our files - phenotype and PGSs
cov<-read.table("pdhp_phenocov.txt",sep="",header=T)
scoreR2<-read.table("Run2.best",sep="",header=T)
scoreR3<-read.table("Run3.best",sep="",header=T)

## Take a quick look at our files
head(cov); dim(cov)
head(scoreR2); dim(scoreR2)
head(scoreR3); dim(scoreR3)
```
## Merge the phenotype file and the Run2 polygenic score

temp1 <- merge(cov, scoreR2[, c(1, 2, 4)], by = c("FID", "IID"))
colnames(temp1)[14] <- "All_SNPs_PGS"

## Merge the combined file and the Run3 polygenic score

wpgs <- merge(temp1, scoreR3[, c(1, 2, 4)], by = c("FID", "IID"))
colnames(wpgs)[15] <- "Clumped_PGS"

## A very basic scatter matrix of some of the continuous vars

png(file = "scattermatrix_BMIheightagePGSs.png")
pairs(wpgs[, c(4, 5, 6, 14, 15)])
dev.off()
No clumping, pT = 1
## Scatterplot of PGS at pT=1 and BMI
```
png(file="scatter_R2pgs_BMI.png")
plot(x=wpgs$All_SNPs_PGS, y=wpgs$BMI, ylab="BMI", xlab="Polygenic score at pT=1, no clumping", main="Run 2 polygenic score by BMI", col=as.factor(wpgs$sex))
legend("topleft", legend=c("Male", "Female"), col=c(2,1), pch=c(1,1))
dev.off()
```

Default clumping, pT = 1
## Scatterplot of clumped PGS and BMI
```
png(file="scatter_R3pgs_BMI.png")
plot(x=wpgs$Clumped_PGS, y=wpgs$BMI, ylab="BMI", xlab="Clumped polygenic score at pT=1", main="Run 3 polygenic score by BMI", col=as.factor(wpgs$sex))
legend("topleft", legend=c("Male", "Female"), col=c(2,1), pch=c(1,1))
dev.off()
```

Incremental R2:
```
> summary(mod1)
Call: lm(formula = BMI ~ age + age2 + as.factor(sex) + PC1 + PC2 +
PC3 + PC4 + PC5 + AT1_SNPs_PGS, data = wpgs)
Residuals:
       Min        1Q  Median        3Q       Max
-11.972 -3.1340   0.387   3.009   13.611
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.777e+01  2.847e+01  1.318 0.21995
age            2.773e-01  8.583e-01  0.323 0.74970
age2           2.567e-03  6.551e-03 -0.392 0.69832
as.factor(sex)1 -3.049e+00  1.500e+00 -2.033 0.04585 *
PC1            -4.793e+01  6.934e+01 -0.722 0.47240
PC2             3.748e+01  6.943e+01  0.538 0.59225
PC3            -6.311e+01  9.796e+01 -0.647 0.51993
PC4             4.194e+01  8.510e+01  0.492 0.62687
PC5             9.874e+01  6.599e+01  1.511 0.13313
AT1_SNPs_PGS   3.102e+05  1.168e+05  2.655 0.008981 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 6.056 on 70 degrees of freedom
Multiple R-squared: 0.2478, Adjusted R-squared: 0.1511
F-statistic: 2.563 on 9 and 70 DF, p-value: 0.01311
```
```
> summary(mod2)
Call: lm(formula = BMI ~ age + age2 + as.factor(sex) + PC1 + PC2 +
PC3 + PC4 + PC5 + Clumped_PGS, data = wpgs)
Residuals:
       Min        1Q  Median        3Q       Max
-12.899 -3.1211  0.3838  2.9707  15.2345
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.293e+01  2.794e+01  1.179 0.24249
age            7.586e-02  8.458e-01  0.551 0.58506
age2           2.782e-03  6.430e-03 -0.431 0.67581
as.factor(sex)1 -3.122e+00  1.505e+00 -2.078 0.04120 *
PC1            -3.122e+01  6.551e+00 -0.475 0.63644
PC2             3.758e+01  6.709e+00  0.560 0.57973
PC3             9.364e+00  8.349e+00  1.143 0.25714
PC4             1.522e+01  8.324e+00  1.831 0.07153 .
PC5             1.522e+01  8.324e+00  1.831 0.07153 .
Clumped_PGS    1.066e+05  3.659e+04  2.918 0.004475 **
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 6.047 on 70 degrees of freedom
Multiple R-squared: 0.167, Adjusted R-squared: 0.122
F-statistic: 2.784 on 9 and 70 DF, p-value: 0.008036
```
```
> Incremental R2:
Residual standard error: 5.713 on 78 degrees of freedom
Multiple R-squared: 0.08991, Adjusted R-squared: 0.0876
F-statistic: 8.58 on 1 and 78 DF, p-value: 0.004443
```
Conclusions

We found a strong, positive association between a BMI polygenic score at a p-value threshold of 1 with no clumping and BMI ($B=310200 \text{ CI } (77184, \ 543199), \ p=0.01$), adjusting for age, age$^2$, sex, and five genetic principal components. The percent of variation in BMI explained by this “all SNPs” polygenic score is around 6.9%.

OR, if you chose the clumped score:

We found a strong, positive association between a BMI polygenic score at a p-value threshold of 1 clumping at 250kb, a linkage r$^2$ of 0.1 and BMI ($B=106800 \text{ CI } (33799, \ 179760), \ p=0.005$), adjusting for age, age$^2$, sex, and five genetic principal components. The percent of variation in BMI explained by this “clumped” polygenic score is around 9.9%.

Some last minute notes

Many of ways to construct a PGS

- biological vs prediction
- don’t train in your own data

How well do ancestry, phenotype, life course match (most GWAS are on European adults)

Social outcomes GWAS are likely to have environmental correlates due to ascertainment/training bias

Vast majority of GWAS results are based on those of euro ancestry

[Image of a graph showing the ancestry of GWAS participants over time compared to the global population, with a note by Martin et al. 2019 Nature Genetics]
Some reviews and commentaries – polygenic scores and precision medicine

• A. Torkamani, et al. The personal and clinical utility of polygenic risk scores. Nature Reviews Genetics May 2018
• L. Hercher. Genome Culture: A Personal Risk Score May Be the Next Big Thing in Genetic Medicine, Genome Magazine, April 2018
• K. Beaney, et al. How close are we to implementing a genetic risk score for coronary heart disease? Expert review of molecular diagnostics 2017 Oct 17(10) 905-915
### PRSice options

<table>
<thead>
<tr>
<th>Options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base files</strong></td>
<td></td>
</tr>
<tr>
<td>--base</td>
<td>Base association file</td>
</tr>
<tr>
<td>--beta</td>
<td>Whether the test statistic is in the form of BETA or OR. If set, test statistic is assume to be in the form of BETA.</td>
</tr>
<tr>
<td>--A1</td>
<td>Column header containing allele 1 (effect allele) Default: A1</td>
</tr>
<tr>
<td>--A2</td>
<td>Column header containing allele 2 (reference allele) Default: A2</td>
</tr>
<tr>
<td>--bp</td>
<td>Column header containing the SNP coordinate Default: BP</td>
</tr>
<tr>
<td>--chr</td>
<td>Column header containing the chromosome Default: CHR</td>
</tr>
<tr>
<td>--index</td>
<td>If set, assume the INDEX instead of NAME for the corresponding columns are provided. Index should be 0-based (start counting from 0)</td>
</tr>
<tr>
<td>--info-base</td>
<td>Base INFO score filtering. Format should be &lt;Column name&gt;,&lt;Threshold&gt;. SNPs with info score less than &lt;Threshold&gt; will be ignored Column name default: INFO Threshold default: 0.9</td>
</tr>
<tr>
<td>--maf-base</td>
<td>Base MAF filtering. Format should be &lt;Column name&gt;,&lt;Threshold&gt;. SNPs with maf less than &lt;Threshold&gt; will be ignored</td>
</tr>
<tr>
<td>--pvalue</td>
<td>Column header containing the p-value Default: P</td>
</tr>
<tr>
<td>--se</td>
<td>Column header containing the standard error Default: SE</td>
</tr>
<tr>
<td>--snp</td>
<td>Column header containing the SNP ID Default: SNP</td>
</tr>
<tr>
<td>--stat</td>
<td>Column header containing the summary statistic If --beta is set, default as BETA. Otherwise, will search for OR or BETA from the header of the base file</td>
</tr>
<tr>
<td><strong>Clumping</strong></td>
<td></td>
</tr>
<tr>
<td>--clump-kb</td>
<td>The distance for clumping in kb Default: 250</td>
</tr>
<tr>
<td>--clump-r2</td>
<td>The R2 threshold for clumping Default: 0.100000</td>
</tr>
<tr>
<td>--clump-p</td>
<td>The p-value threshold use for clumping. Default: 1.000000</td>
</tr>
<tr>
<td>--ld</td>
<td>LD reference file. Use for LD calculation. If not provided, will use the post-filtered target genotype for LD calculation. Support multiple chromosome input Please see --target for more information</td>
</tr>
<tr>
<td>--ld-keep</td>
<td>File containing the sample(s) to be extracted from the LD reference file. First column should be FID and the second column should be IID. If --ignore-fid is set, first column should be IID Mutually exclusive from --ld-remove No effect if --ld was not provided</td>
</tr>
<tr>
<td>--ld-remove</td>
<td>File containing the sample(s) to be removed from the LD reference file. First column should be FID and the second column should be IID. If --ignore-fid is set, first column should be IID Mutually exclusive from --ld-keep</td>
</tr>
<tr>
<td>--ld-type</td>
<td>File type of the LD file. Support bed (binary plink) and bgen format. Default: bed</td>
</tr>
<tr>
<td>--no-clump</td>
<td>Stop PRSice from performing clumping</td>
</tr>
<tr>
<td>--proxy</td>
<td>Proxy threshold for index SNP to be considered as part of the region represented by the clumped SNP(s). e.g. --proxy 0.8 means the index SNP will represent region of any clumped SNP(s) that has a R2&gt;=0.8 even if the index SNP does not physically locate within the region</td>
</tr>
<tr>
<td><strong>Covariate options</strong></td>
<td></td>
</tr>
<tr>
<td>--cov-file</td>
<td>Covariate file. First column should be FID and the second column should be IID. If --ignore-fid is set, first column should be IID</td>
</tr>
<tr>
<td>--cov-col</td>
<td>Header of covariates. If not provided, will use all variables in the covariate file. By adding @ in front of the string, any numbers within [ and ] will be parsed.</td>
</tr>
</tbody>
</table>
E.g. @PC[1-3] will be read as PC1,PC2,PC3. Discontinuous input are also supported: @cov[1.3-5] will be parsed as cov1,cov3,cov4,cov5

---

**Dosage options**

--hard-thres

Hard threshold for dosage data. Any call less than this will be treated as missing. Note that if dosage data is used as a LD reference, it will always be hard coded to calculate the LD. Default: 0.900000

--hard

Use hard coding instead of dosage for PRS construction. Default is to use dosage instead of hard coding.

---

**PRSice options**

--bar-levels

Level of barchart to be plotted. When--fastscore is set, PRSice will only calculate the PRS for threshold within the bar level. Levels should be comma separated without space.

--fastscore

Only calculate threshold stated in--bar-levels

--full

Include the full model in the analysis

--interval | -i

The step size of the threshold. Default: 0.000050

--lower | -l

The starting p-value threshold. Default: 0.000100

--model

Genetic model use for regression. The genetic encoding is based on the base data where the encoding represent number of the effective allele. Available models include:
- add - Additive model, code as 0/1/2 (default)
- dom - Dominant model, code as 0/1/1
- rec - Recessive model, code as 0/0/1
- het - Heterozygous only model, code as 0/1/0

--quantile

Along with a number, will create a quantile plot to see effect of increasing PRS

--no-regress

Do not perform the regression analysis and simply output all PRS

--score

Method to handle missing genotypes. By default, final scores are averages of valid per-allele scores with missing genotypes contribute an amount proportional to imputed allele frequency. To throw out missing observations instead (decreasing the denominator in the final average when this happens), use the 'no_mean_imputation' modifier. If --missing SET_ZERO is set, the SNP for the missing samples will be excluded. Alternatively, if --missing CENTER is set, all PRS calculated will be minused by the MAF of the SNP (therefore, missing samples will have PRS of 0).
### Target Files

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>--binary-target</td>
<td>Indicate whether the target phenotype is binary or not. Either T or F should be provided where T represent a binary phenotype. For multiple phenotypes, the input should be separated by comma without space. Default: T if --beta and F if --beta is not</td>
</tr>
<tr>
<td>--info</td>
<td>Filter SNPs based on info score. Only used for imputed target</td>
</tr>
<tr>
<td>--keep</td>
<td>File containing the sample(s) to be extracted from the target file. First column should be FID and the second column should be IID. If--ignore-fid is set, first column should be IID Mutually exclusive from--remove</td>
</tr>
<tr>
<td>--remove</td>
<td>File containing the sample(s) to be removed from the target file. First column should be FID and the second column should be IID. If--ignore-fid is set, first column should be IID Mutually exclusive from--keep</td>
</tr>
<tr>
<td>--pheno-file</td>
<td>-f</td>
</tr>
<tr>
<td>--pheno-col</td>
<td>Headers of phenotypes to be included from the phenotype file</td>
</tr>
<tr>
<td>--prevalence</td>
<td>-k</td>
</tr>
<tr>
<td>--nonfounders</td>
<td>Keep the nonfounders in the analysis Note: They will still be excluded from LD calculation</td>
</tr>
<tr>
<td>--target</td>
<td>-t</td>
</tr>
<tr>
<td>--type</td>
<td>File type of the target file. Support bed (binary plink) and bgen format. Default: bed</td>
</tr>
</tbody>
</table>

### Miscellaneous options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>--all-score</td>
<td>Output PRS for ALL threshold. WARNING: This will generate a huge file</td>
</tr>
<tr>
<td>--exclude</td>
<td>File contains SNPs to be excluded from analysis</td>
</tr>
<tr>
<td>--extract</td>
<td>File contains SNPs to be included in the analysis</td>
</tr>
<tr>
<td>--ignore-fid</td>
<td>Ignore FID for all input. When this is set, first column of all file will be assume to be IID instead of FID</td>
</tr>
<tr>
<td>--logit-perm</td>
<td>When performing permutation, still use logistic regression instead of linear regression. This will substantially slow down PRSice</td>
</tr>
<tr>
<td>--keep-ambig</td>
<td>Keep ambiguous SNPs. Only use this option if you are certain that the base and target has the same A1 and A2 alleles</td>
</tr>
<tr>
<td>--out</td>
<td>-o</td>
</tr>
<tr>
<td>--perm</td>
<td>Number of permutation to perform. This will generate the empirical p-value. Recommend to use value larger than 10,000</td>
</tr>
<tr>
<td>--seed</td>
<td>-s</td>
</tr>
<tr>
<td>--print-snp</td>
<td>Print all SNPs used to construct the best PRS</td>
</tr>
<tr>
<td>--thread</td>
<td>-n</td>
</tr>
<tr>
<td>--help</td>
<td>-h</td>
</tr>
</tbody>
</table>