# **runPRS script documentation:**

**Description:**

**By Adrian Campos and Scott Gordon v3.0 4/10/21.**

**This script performs automatically a PRS pipeline on the genotype data from the genepi group. It has two options for dealing with the problem of LD in PRS: i) Clumping + Thresholding or ii) Estimating multivariate effect sizes using SBayesR. It should be readily modifiable to be run on other genotype datasets and clusters. It was developed on python 3.6 and requires the following libraries (all available on hpc):**

* **pandas**
* **re**
* **argparse**
* **subprocess**
* **os**
* **glob**
* **time**
* **numpy**
* **warnings**
* **string**
* **sys**
* **datetime**
* **socket**

**The following diagram shows the overall PRS process and naming conventions:**

GWAS sumstats

Genepi dataset markers e.g. R10

QC

GWAS sumstatsQCed

LD ref panel (1000G)

Independent SNPs (clumping)

Multivariate effect sizes (SBayesR)

LD matrix (UKB)

Match to

Match to

Genepi dataset individual genetic profiles (scores)

**Quickstart guide:**

**Call to run a default PRS including both SBayesR and clumping + thresholding:**

**$ runPRSv2.py inputSumstats PRSname**

**Call to run only clumping + thresholding:**

**$ runPRSv2.py inputSumstats PRSname -runSBayesR False**

**Call to run only SBayesR:**

**$ runPRSv2.py inputSumstats PRSname -runclumping False**

**Please see the guide below for input sumstats format, and the required columns to run SBayesR or C+T**

**It performs the following**

1. **Opens the summary statistics (they have to be ‘QCed’) and in a specific format (see below)**
   1. **Removes indels**
   2. **MAF RSQ as needed**

**The script also does the following:**

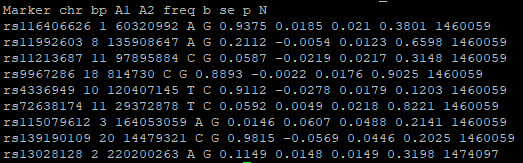
* + 1. **Interprets summary statistics headers**
       1. **(try to use REF and ALT instead of A1 and A2)**
       2. **Also accepts column names as parameters**
    2. **Transforms OR to log(OR)**
    3. **Removes X chr**
    4. **Removes indels**
    5. **Filters ambiguous strand SNPs**

1. **Runs SBayesR analysis to reestimate multivariate SNP effect sizes (Optional step).**
   1. **SBAYESR requires the GWAS to have:**
      1. **An N column**
      2. **RS number as Marker (regardless of whether youre matching by ch:bp:a1:a2)**
      3. **Effect allele frequency**
      4. **Effect allele beta and standard error**
2. **Performs QC on metadata (Genepi dataset).**
   1. **Removes indels**
   2. **Filters imputation quality Rsq>=0.6 DEFAULT**
   3. **Filters MAF 0.01<=x<=0.99**
   4. **Filters ambiguous strand alleles by removing them**
3. **Matches GWAsumstats and SBayesR results to the QCed metadata** 
   1. **Matching can be done by:**
      1. **rs number**
      2. **chr:bp:a1:a2 (also checks ch:bp:a2:a1 just in case) DEFAULT**
      3. **marker name in the Genepi release (not recommended)**
4. **Runs polygenic scoring based on SBayesR effect sizes (Optional step).**
5. **Creates a list of all usable SNPs for the clumping by merging the GWA summary stats and the genepi dataset genotype data, removing duplicates and incoherent SNPs (same rs number but different chr:bp or ref alt (beyond strand ambiguity) alleles)**
6. **Checks if there are duplicates after matching and raises a warning if there are (if there are duplicates rows (exact same) then one is ignored, but sometimes there are sumstats with duplicate SNPs and different pvalues or effect sizes.**
7. **Performs the clumping:**
   1. **Creates a local working directory ‘clumping’**
   2. **Creates the list of available SNPs on that working dir**
   3. **Creates the list of SNPs for extract (allSnpRsNumberUniqueNoIndels)**
   4. **Creates all the clumping scripts ending with ‘.PBS’**
   5. **Submits all the clumping scripts to the cluster**
   6. **Waits for completion**
   7. **Checks for errors and reports them**
8. **Makes the pvalues and betas working directories (to store said files)**
9. **Makes the clumped betas and pvalue files**
10. **Calculates the clumping + thresholding PRS (using plink):**
    1. **Creates two nested working directories: PRS\_calc and PRS\_calc/PRS\_out**
    2. **Creates the submission scripts (one for each block)**
    3. **Submits the scrips to the cluster**
    4. **Waits for job completion**
    5. **Checks for errors and reports them**
11. **Compiles all PLINK profiles matching the cutoffs used**
12. **Outputs the file to the working directory**

**The code is available at the copy provided (add it to your bin). Feel free to copy it, edit it or just link it to your path to be able to execute it from wherever (you have to edit your $HOME/.bashrc file and add the line export PATH="$PATH:/mnt/backedup/home/adrianC/bin/runPRS", soon available in genepi)**

**To run this script you require a working directory (accessible by the nodes {not in labdata}) that contains the following files:**

1. **yfp (your favourite phenotype) sumstats (preferred format below):**



**This is a tab or space delimited file containins at least the following columns (extras do not matter, and it is capslock insensible so SNP==snp):**

* **Marker name - SNPrs [snp,snpid,rsnumber,rs,snprs] – rsnumber**
* **CHR [chr,chrom,chromosme]- chromosome**
* **BP [bp,pos]– basepair**
* **Reference\_allele [ALLELE0, A2,a2,ref,reference] – reference allele (non-effect allele CAREFUL WITH THIS ONE)**
* **effect\_allele [ALLELE1, A1,a1,minor\_allele,alt,ALT] – alt, effect or minor allele (CAREFUL WITH THIS ONE)**
* **beta [BETA,effsize,beta,or]– effect size or OR**
* **pval [p,pvalue,pvalues]– pvalues**

**For SBayesR, the Marker HAS to be rsnumber, the frequency should be the effect allele frequency and the beta standard error (se) is also required!**

**NOTE: Column order doesn’t matter, the headers should be identical or one of the [synonyms]. Extra columns will be ignored**

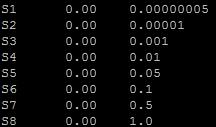
**If you do not have rsnumber you can match your data to our rsnumbers by using a script like this:**

**awk 'NR==FNR{id[$1]=$2;next}($1 in id) {print id[$1]"\t"$0}' allchr.markerlist.IDs GWASUMSTATS > GWASUMSTATS\_rs**

**Can be found in the release metadata**

1. **pvalue.ranges file:**

**Name lower upper**



**The pvalue.ranges file must lie in your working directory (where your input sumstats are). If not provided it will be automatically generated to match the one showed above.**

**This script requires the python module to be loaded (module load python) and at least 15Gb of memory so it should be ran either interactively or as a Job using qsub or a PBS script.**

**How to run interactively:**

**Start interactive session:**

**$ qsub -I -l mem=32gb,walltime=10:00:00**

**Load python module**

**$ module load python**

**Move to the working directory (either as you would normally do or):**

**$ cd $PBS\_O\_WORKDIR**

**Submit the command**

**$ runPRSv2.py sumstats JOBNAME**

**How to run it with a PBS script (more reproducible):**

**Create the script file (where the input files are):**

**$ touch PRS\_SUBMIT.PBS**

**Edit the script file:**

**$ nano PRS\_SUBMIT.PBS**

**Add the following lines:**

**#!/bin/bash**

**#PBS -N JOBNAME**

**#PBS -l mem=20GB,walltime=12:00:00**

**module load python**

**cd $PBS\_O\_WORKDIR**

**runPRSv2.py sumstats JOBNAME**

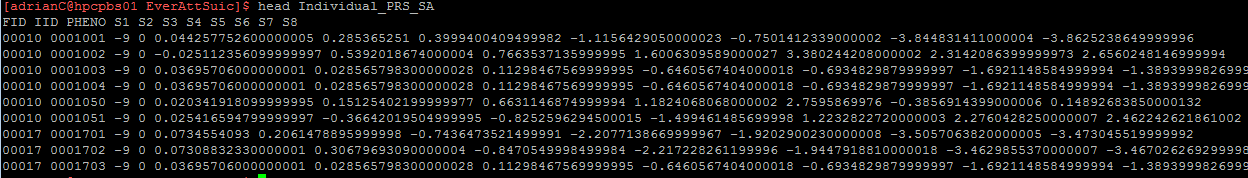
**To submit it run the following:**

**$ qsub PRS\_SUBMIT.PBS**

**Important Note: This script requires the submission to be called from inside (within) the working directory (where your input files are located), as it will use the environmental variable $PBS\_O\_WORKDIR when it submits ‘children’ jobs.**

**The –log flag is deprecated**

**A quick overview of the many outputs of this script:**

* **clumping directory – contains: the snp list used for –extract, the available snps used for –clump, the clumping scripts submitted to the cluster, the clumping logs of plink, the clumped files and the hpc output and error files.**
* **Betas directory – contains the clumped matched betas from the summary statistics used.**
* **Pvals directory – contains the clumped matched pvalues from the sumstats.**
* **PRS\_calc directory – contains: PRS\_out directory, the PBS scripts to calculate polygenic risk scores, the hpc output and error files.**
* **PRS\_calc/PRS\_out directory – contains the plink logs of the PRS calculation, the PRS profiles (one per block per chromosome per pvalue cutoff)**
* **SBayesR – SbayesR working directory and output files, as well as log files**
* **SBayesRbetas – Matched SbayesR betas top the genepi dataset for allelic scoring**
* **ClumpThresh\_JobName.profile [also Std\_ClumpThresh\_JobName.profile] – Final compiled output with the following columns:**
  + **FID**
  + **IID**
  + **CNT – marker count to calculate PRS**
  + **PRS1 – with the name you placed on your pvalue.ranges (e.g. S1)**
  + **PRS2**
  + **…**
  + **PRSN – the last cutoff you used (ordered based on the name)**
* **SBayR\_JobName.profile [also Std\_SBayR\_JobName.profile] – SbayesR scoring results.**
* **If desired a final file with all the merged output (c+t and sbayesr can be made)**

**Script OPTIONAL flags or parameters:**

* **-o (--output) = Output file, Individual\_PRS\_$jobname by default**
* **-std = Output ALSO standardized PRS True by default**
* **-dataset =Which dataset from the release (release + imputation) to use, default is Release10\_HRCr1.1**
* **-eafield =Column name of the effect allele**
* **-oafield= Column name of the other allele**
* **-pvaluefield= Column name of pvalue field**
* **-SNPfield =Column name of Marker name field (rs for SbayesR)**
* **-chrfield= Column name of CHR**
* **-BPfield= Column name of base position**
* **-betafield= Column name of effect size or beta**
* **-ORfield =Column name of OR (will be transformed to beta)**
* **-FREQfield= Column name of effect allele frequency field (Required for SBayesR not for CT)**
* **-SEfield= Column name of standard error field (Required for SBayesR not for CT)**
* **-Nfield= Column name of per SNP GWAS N (Required for SBayesR not for CT)**
* **-assocstrand= Strand alignment of summary statistics + - or unkown**
* **-analysischromosomes= (list of chromosomes to analyse autosomes by default)**
* **-minRsq= QC rsquared value default 0.6**
* **-minMAF= QC MAF value default 0.01**
* **-maxMAF= default 0.5**
* **-matchby= [rsID,name,positionandalleles,position] – How to match GWAS to Genepi dataset**
* **-rangesfile = Pvalue ranges file name (not required)**
* **-runmetadata = [True|False] whether to run metadata matching step default=True**
* **-runclumping = [True|False] whether to run clumping step default=True**
* **-runSBayesR= [True|False] whether to run SBayesR method default=True**
* **-runPRS = [True|False] whether to run PRS scoring (independent of CT or SBayesR will try all) default=True**
* **-compileblocks = [True|False] Whether to compile the putput. default=True**
* **-mergeCTandSB = [True|False] Whether to merge CT and SB outputs into a single profile file. default=True**

**Summary:**

* **This calculates, with either clumping or SBayesR, a PRS on our genotype data based on summary statistics from other sample.**
* **It is a python script that should not be run on login nodes.**
* **To run it you need input file and choose a job name:**
  + **GWAsumstats - with the format specified above**
  + **Job name – it is required**
* **An example of how the script is to be ran (from within the working directory):**
  + **$ runPRS infile jobname**
  + **Be sure to run it from an interactive session or via qsub**
* **Several outputs created for debugging/reproducibility, but the important ones end in .profile**

**The next steps involve assessing the variance explained by this calculated PRS in the Twin data. Because they are related individuals an lmm analysis using GCTA should be performed. (see PRSlmm Documentation)**