Instructions for polygenic risk scoring (PRS) using PLINK

Last changes: 27.06.2018

Swapnil Awasthi, Ph.D. student, swapnil.awasthi@charite.de
Prof. Stephan Ripke, M.D., Ph.D., Group Leader, stephan.ripke@charite.de
GResu (GWAS Research Unit, BIH, Berlin)

I. Target genotype data.
II. LD clumped discovery data set.
III. Q-Range file.

1. Create a new directory

```
mkdir test_prs
cd test_prs
```

2. Create a symbolic link of the genotype data (target data set)to be scored in this directory. (make sure it is post-imputation best guess genotypes with at least several million SNPs)

```
ln -s
/user/plink/genotype/mix_XXXX_eur_sr-qc.hg19.ch.fl.bgn.bim
./
ln -s
/user/plink/genotype/mix_XXXX_eur_sr-qc.hg19.ch.fl.bgn.bed
./
ln -s
/user/plink/genotype/mix_XXXX_eur_sr-qc.hg19.ch.fl.bgn.fam
./
```

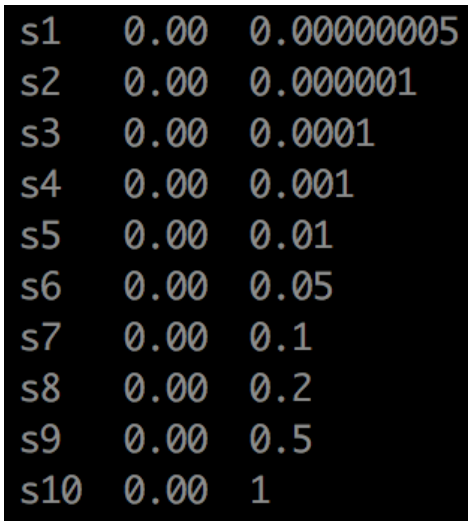
3. Copy the LD-clumped discovery data set. This will be used as weights in PRS.

```
cp /user/plink/clump/PGC2012/pgc.scz.clump.2012-04.txt.gz
./
```

Convert odds ratios to log odds ratios. Using odds ratios instead of log odds will yield incorrect results.

```
gunzip -c pgc.scz.clump.2012-04.txt.gz | awk
'$9=log($9){print}' > pgc.scz.clump.2012-04.txt.gz.xhmc
```

4. Create a q range file: A text file in which each row corresponds to a **different score**, containing a label, then a lower and upper bound for the values as given in the discovery (weights) file. For example one can give a lower and upper bound of **p-value** then only the SNPs that lie between this range will be used. Example of q range file below



```
s1   0.00  0.00000005
s2   0.00  0.000001
s3   0.00  0.0001
s4   0.00  0.001
s5   0.00  0.01
s6   0.00  0.05
s7   0.00  0.1
s8   0.00  0.2
s9   0.00  0.5
s10  0.00  1
```

5. Create PRS for the target dataset based on the discovery set using following plink command

```
plink \
--allow-no-sex \
--bfile mix_XXXX_eur_sr-qc.hg19.ch.fl.bgn \
--q-score-range range pgc.scz.clump.2012-04.txt.gz.xhmc 2
11 header \
--score pgc.scz.clump.2012-04.txt.gz.xhmc 2 4 9 header sum
\
--out OUTNAME.XXXX.bgn.score
```

--q-score-range: take the argument as range file (from Step 4), the LD-clumped file(Step 3) and column number of variant ids (Eg. 2) and the column number of p-values (eg.11)
--score: take the argument as LD-clumped file(Step 3), column number of variant ids (Eg.

2), column number of allele codes (Eg.4) and scores(Eg.Log of odds ratio,9)

6. Look at the different (based on -q-score-range) results files (explanation for S1-S10 see screenshot under 4.)
   OUTNAME.XXXX.bgn.score.S1.profile
   OUTNAME.XXXX.bgn.score.S2.profile
   OUTNAME.XXXX.bgn.score.S3.profile
   .
   .
   .
   OUTNAME.XXXX.bgn.score.S10.profile

7. For Correlation Analyses you usually need to integrate at least some PCA covariates. For case/control datasets this is usually done via “Nagelkerke”. Separate scripts available by request