BMI 633 Term Project

Improved significant SPNs selection for Polygenic Risk Score Calculation

By

James Kambere and Purushottam Panta

Introduction

Polygenic Risk Score (PRS) is a number that provide an estimation on an individual's likelihood of developing a particular heritable disease based on the number of genetic variants that they have. Polygenic Risk Scores were first used in predicting human genetic liability to a trait in 2009. Polygenic Risk Score have shown a great impact in predicting individual's risks in developing heritable diseases since its inception in 2009. PRSs are playing an important role in precision medicine. PRS values determines how risk an individual is to a particular trait. This enables medical personnel to identify and provide proper medical care to such individuals.

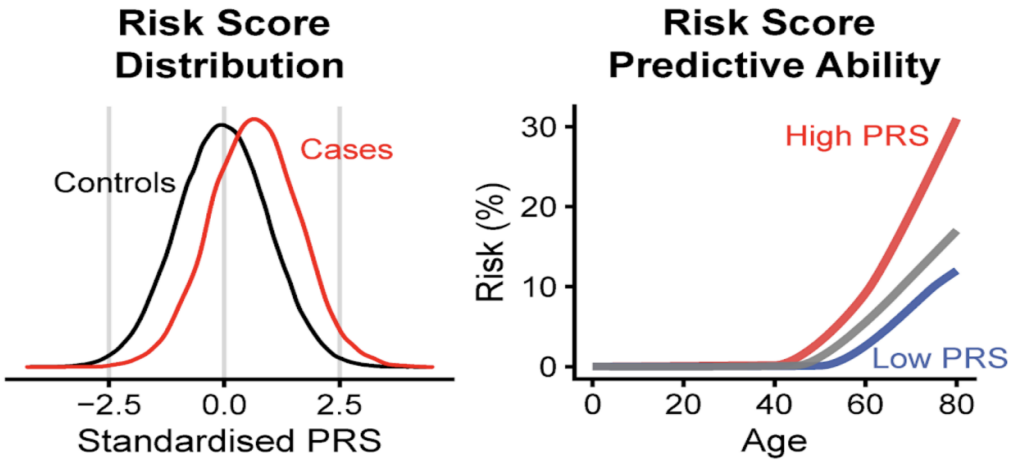


Fig1. The left panel shows how in the predictions of disease risk, the PRS on the x-axis, can separate from the controls. The y-axis describes how many in each group are assigned a certain PRS. To the right, the same population is divided into three groups according to the predicted risk, their assigned PRS. The observed risk is shown on the y-axis and the separation of the groups is in correspondence with the predicted risks. (https://www.medrxiv.org/content/10.1101/2020.04.23.20077099v2.full-text)

PRSs are calculated from Genomic Wide Association Study (GWAS) summary statistics data. The rapid increase in GWAS's sample size is leading more accurate and useful PRSs for complex heritable diseases. GWAS identifies genetic variants most of which are single nucleotide

polymorphisms (SPNs) that are significantly associated with a variety of traits. These variants usually have small effect on the trait hence they have a limited predictive power. To improve the predictive power of the variants to trait, statically correct methods of aggregating the effects of the variants were proposed. PRSs is one of the proposed variants effect aggregation method.

Polygenic Risk Scores are calculated as weighted sum of risk alleles among a set of SPNs that an individual have. The weights are the risk allele effect sizes from GWAS summary statistics data. There are several techniques and procedure that are used in calculation of PRSs. PLINK implemented PRS calculation algorithm that is widely used by many researchers. PLINK's PRS calculation algorithm improvements have been suggested and implemented by other researchers. These improvements include PRSice – 2, Ldpred – 2 and many others. The PRS, ***S*** is given by:

$$\boldsymbol{S} = \sum_{\boldsymbol{l=1}}^{\boldsymbol{n}} \boldsymbol{\beta}_{\mathrm{l}} \boldsymbol{x}_{\mathrm{l}}$$

(Eq1: PRS calculation)

Where β is the weight of the SNP *l* (either an odds ratio or a beta coefficient) and *x* is the number of risk alleles an individual have for SNPs *l.*

Problem Statement.

In order to calculate PRS, two files are required. These two files are GWAS summary statistics data and individual level genotype and phenotype data. The actual PRS calculation has four of steps. The first step is quality control (QC).

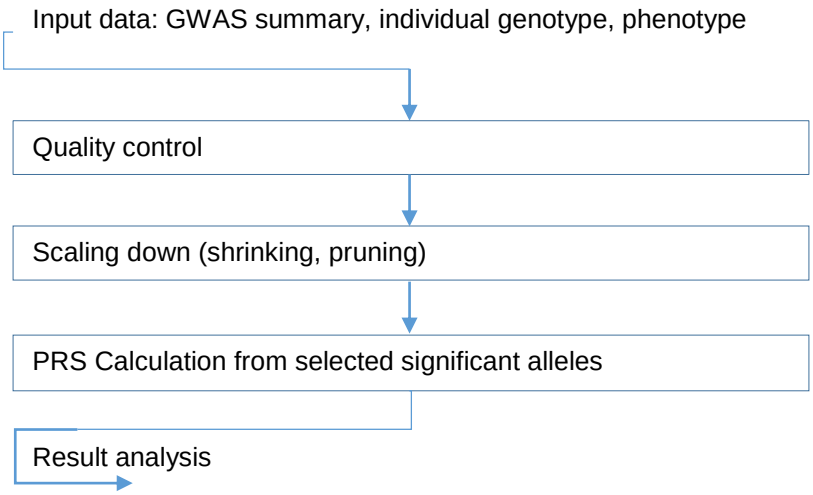


Fig2: PRS Calculation and analysis

Both GWAS summary statistics data and the target individual's data undergoes QC steps that aims at removing variable and values that may led to generation of false positive results. There are several QC methods implemented in various bioinformatics software. PLINK has a robust and easy to use QC algorithm. If one is using a one SNP at a time based GWAS summary statistic data, Linkage Disequilibrium (LD) has to be accounted for (Crumping). The most commonly used crumping mechanism uses correlation coefficient as measure. All SNPs with a correlation coefficient of greater than a given value (mostly 0.2) are retained. The second step in PRS calculation involves scaring down (shrinking or pruning) the effect size of the alleles. This step involves selection of significant SNPs from a given GWAS summary statistics data. This step is followed by the actual calculation of PRS from the selected significant SNPs. The last step in PRS calculation is the analysis of the obtained results.

Pruning is the most important step in calculation of PRS. Several pruning techniques have been developed and used in different PRS calculation software. Standard PRS calculation tools such as PLINK uses a combination of Linkage Disequilibrium (LD) crumping and p – value threshold in selecting significant SNPs. This p – value threshold is based on Bonferronis adjustment. Bonferroni adjustment suggest that a p – value, say $\boldsymbol{p}$ is significant if:

$$\boldsymbol{p \leq \frac{\alpha}{n}}$$

(Eq2: Bonferroni adjustment of significant p-value)

Where $\boldsymbol{n}$ is the number of SNPs in GWAS summary statistics data and $\boldsymbol{\alpha}$ is the given level of significant which is mostly 0.05. In this case, every SPNs with LD correlation coefficient, $\boldsymbol{r^2 < 0.2}$ and p – value that meet the bonferroni adjustment are selected as significant SNPs.

The problem with the above stated pruning method is that there is no statistics that quantifies the significant extent of LD. Using a static p – value threshold as proposed in bonferroni adjustment leads to increase in type II error as the experimental wise error rate remains unchanged despite the numerous comparisons made. Selecting significant SPNs this way may end up leaving important alleles that might be significantly associated with the trait. It is so difficult to avoid bias in determining the static p – value threshold under this threshold criteria.

Proposed Solution

In this project, we are proposing to solve the above stated problems by introducing a new simple method (improved bonforreni adjustment) for determining significant SPNs from a given GWAS statistic file. Since studies have shown that inclusion of more SNPs in calculation of PRS improves the predictive power of the score, **our method will be a success if it will get more alleles** included into the calculation of PRS. In this case we will be sure that no alleles that might be significant to the trait have been left out. The **improved bonforreni adjustment** state that, for any order collection of p – values, a p value at position ***i*** is significant if:

$$p_i \leq \frac{i\alpha}{n}$$

(Eq3: Improved Bonferroni adjustment of significant p-value)

Where ***i*** is the position of a p – value in the ordered list of p – values and $i > 0$, ***n*** is the number of SNPs and α is a given level of significance, mostly 0.05. In this case we are a voiding the use of a static p – value threshold as each and every p – value within the ordered list will be compared to a different threshold. This method reduces bias by subjecting each p – value to its corresponding comparison value. It is statistically proven that it reduces type II error in multiple comparisons of p – values. When applied to independent tests, this method has been proven to have a reduced type I error probability of equal to alpha.

Method

In this project, we have implemented the improved bonferroni adjustment algorithm in R. To read the large zipped GWAS summary statistics dataset, we installed and used an r package called ***R.utils*****.** Some QC steps are implemented to check the quality of the GWAS summary statistics data and the target individual's phenotype data. Firstly we removed SNPs with a Minor Allele Frequency (MAF) with a value of less than 0.01 and Imputation Information Score of less than 0.8 in R. Duplicate SNPs were then removed from the resulting dataset using an r function called ***duplicated ().*** To read .bed, .fam and .bim target individual phenotype files, we installed and used r package called ***BEDMatrix***. BEDMatrix created a matrix of individuals ID, SNPs, Effect Alleles and the number of risk alleles for each SNPs. We installed and used ***LDlinkR*****,** an r package for SNPs crumping. We implemented a crumping (LD) algorithm and crumped the GWAS summary statics dataset with it. For testing purpose and limitation in computation power, we only limited our input to 5000 SNPs. Both the old (PLINK model) and the improved p – value threshold techniques were implement. The data set obtained after crumping was then supplied to

these two techniques independently. We compared the performance of the algorithm on two levels. The first comparison was made on the number of SNPs selected as significant on both old and improved algorithms. The second comparison was made on the PRS score obtained. We merged the pruned GWAS summary statistics data and the BEDmatrix file of the phenotype information obtained in our earlier steps into a single file using the SNPs in R. The weight scores were calculated as the sum of the effect alleles an individual had multiplied by the beta value as weight. Due to big volume of the GWAS summary statistics data, our local computers took more minutes to process the data. We selected R because it is so fast in data manipulation in comparison with python. We did not perform analysis of the scores obtained due to time factor.

Data

In this project, we used simulated GWAS summary statistics data set obtained from the link: https://drive.google.com/file/d/1RWjk49QNZj9zvJHc9X_wyZ51fdy6xQjv/view. This is a file of simulated variants associated with height and body mass index. This simulated data is widely used in many PRS calculation tutorial sites. Our GWAS summary data file had 529504 SNPs in total.

```
> dat
        CHR       BP         SNP A1 A2      N         SE           P        OR      INFO       MAF
     1:   1   756604   rs3131962  A  G 388028 0.00301666 4.83171e-01 0.9978869 0.8905579 0.3693896
     2:   1   768448  rs12562034  A  G 388028 0.00329472 8.34808e-01 1.0006873 0.8958935 0.3368458
     3:   1   779322   rs4040617  G  A 388028 0.00303344 4.28970e-01 0.9976036 0.8975083 0.3773680
     4:   1   801536  rs79373928  G  T 388028 0.00841324 8.08999e-01 1.0020357 0.9089629 0.4832122
     5:   1   808631  rs11240779  G  A 388028 0.00242821 5.90265e-01 1.0013083 0.8932125 0.4504096
    ---
529500:  22 51174939  rs73174435  T  C 388028 0.00453846 7.33506e-03 1.0122431 0.8878835 0.4870748
529501:  22 51175626   rs3810648  G  A 388028 0.00425146 7.83096e-05 1.0169332 0.8908870 0.2903021
529502:  22 51183255   rs5771002  A  G 388028 0.00215615 7.99115e-04 1.0072560 0.8905665 0.2377800
529503:  22 51185848   rs3865764  G  A 388028 0.00463123 5.54611e-01 0.9972673 0.9053373 0.4034021
529504:  22 51193629 rs142680588  G  A 388028 0.00383730 1.03876e-01 0.9937786 0.9028023 0.4191540
```

In this file the effect allele are the ones under A1. The p – values are under the column P and the SNPs are under SNP. After undergoing some QC procedures, the file remained with 529493 SNPs. Out of these SNPs, only **5000** well subjected to crumping procedure. **2004** SNPs were retained. The significant SNPs were selected from this list of 2004 SNPs.

```
 [50] "rs2710887"   "rs9803031"   "rs113261977" "rs3813193"   "rs11260596"  "rs4075116"   "rs113592356"
 [57] "rs3934834"   "rs9442394"   "rs9442372"   "rs115057577" "rs115723010" "rs61766340"  "rs115662838"
 [64] "rs6687776"   "rs9651273"   "rs11579015"  "rs147606383" "rs12080505"  "rs61766343"  "rs61766344"
 [71] "rs57414802"  "rs9442373"   "rs55945496"  "rs2298216"   "rs4072537"   "rs11260598"  "rs113355263"
 [78] "rs6604971"   "rs116661896" "rs74045142"  "rs77791262"  "rs7538773"   "rs9660710"   "rs77277598"
 [85] "rs72894004"  "rs11260539"  "rs11260542"  "rs11260544"  "rs79295821"  "rs6671609"   "rs111751804"
 [92] "rs13374146"  "rs72631898"  "rs1320571"   "rs77950429"  "rs4634847"   "rs12120877"  "rs115516254"
 [99] "rs12060422"  "rs2274791"   "rs58004717"  "rs114946555" "rs3753350"   "rs35038461"  "rs78479912"
[106] "rs11721"     "rs4970422"   "rs114757189" "rs77472198"  "rs11260563"  "rs6603782"   "rs7536568"
[113] "rs715643"    "rs115765087" "rs78555129"  "rs12093154"  "rs72631899"  "rs75668831"  "rs114569858"
```

List of crumped SNPs

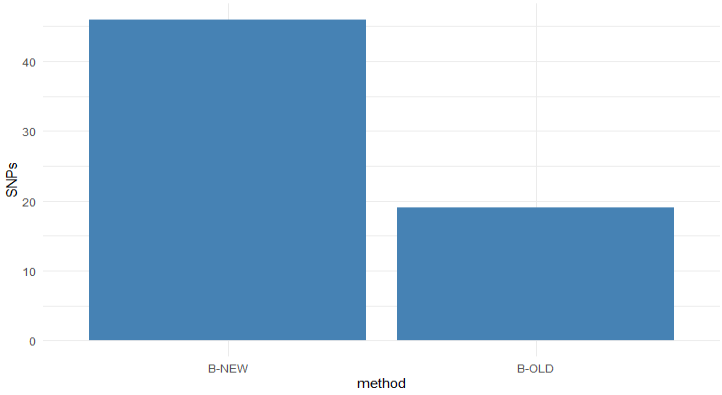
Similarly we obtained our simulated target individual phenotype data from the same website (https://drive.google.com/file/d/1uhJR_3sn7RA8U5iYQbcmTp6vFdQiF4F2/view). This a file contained 1000 genomes. Some QC techniques were applied to the file. SNPs with N/A values were removed. We tested our PRS calculation by selecting only two individuals from the BEDmatrix file.

```
> new_file_withoutna_2
              SNP A1 PERO01 PERO02
    1:  rs3131962  A      1      0
    2:  rs4040617  G      1      0
    3: rs79373928  G      0      0
    4: rs11240779  G      1      1
    5: rs57181708  G      0      0
   ---
51995: rs78959884  T      0      0
51996: rs79721654  A      0      1
51997:  rs4355165  T      0      1
51998: rs11685742  T      0      1
51999: rs17391809  G      0      0
>
```

List of individual effect allele counts

Results Obtained

After performing QC on the GWAS summary statistics data and the target individual phenotype dataset, LD crumping and p – value threshold, static p – value threshold method produced only 19 significant SNPs. Improved bonifferoni adjustment produced 331 significant SNPs. From these obtained results we can broadly claim that our pruning method through improved bonferroni adjustment algorithm **produced more significant** SNPs than static p – value threshold algorithm. For proving sake, we also compared the number of SPNs obtained through improved bonferroni adjustment and those obtained by the use of only static bonferroni adjustment value without crumping. Similarly improved boniferroni algorithm produced more significant SNPs that the static p – value threshold. The PRS scores or our method are also a bit higher than those obtained using the static p – value threshold for the two individuals we selected.



A bar plot showing the number of SNPs obtained by both static p – value threshold (B-OLD) and improved bonferroni (B-NEW) method.

The table below shows the PRS calculated for person 1 and 2 using the SNPs obtained through the old method.

```
>     prs_old_way
 new_PER001  new_PER002
 0.13793156 -0.06384602
```

The table below show the PRS calculated for person 1 and 2 using SNPs obtained through the improved bonferroni adjustment.

```
>     prs_new_way
new_PER001 new_PER002
 0.1957933 -0.2886614
```

Future work

We would like to apply the procedure to real world data on super computer and analysis the performance. In this case, we will have a real world summary statistics data and real phenotype data (target individuals). We are also planning to take this observation further by performing analysis of the results obtained. We will implement regression analysis of the results in R. Plots will be used in visualizing the significance of the results.

Conclusions

Polygenic Risk Score are being used in predicting human reliability to complex traits. They are playing important role in medical researches aimed at finding best ways of treating heritable diseases. Several studies aiming at improving the significant of the PRS have been done and others are still going on. Most of these proposed techniques usually use external information panel adjustment as way of improving the significant of the PRS. Our proposed method entirely

depends on widely available data, the GWAS summary statistics data and phenotype file. Our improved p – value threshold technique has the ability to select more significant SNPs from a crump GWAS summary statistics data. With standard and proper adjustments, this method can work efficiently with large datasets.

Reference

1. Choi, S.W., Mak, T.SH. & O'Reilly, P.F. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc 15, 2759–2772 (2020). https://doi.org/10.1038/s41596-020-0353-1
2. Biometrika, Volume 73, Issue 3, December 1986, Pages 751–754, https://doi.org/10.1093/biomet/73.3.751 (Improved Bonferroni adjustment)
3. Truong, B., Zhou, X., Shin, J. et al. Efficient polygenic risk scores for biobank scale data by exploiting phenotypes from inferred relatives. Nat Commun 11, 3074 (2020). https://doi.org/10.1038/s41467-020-16829-x
4. Zhao, Z., Yi, Y., Song, J. et al. PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics. Genome Biol 22, 257 (2021). https://doi.org/10.1186/s13059-021-02479-9
5. So, HC. Sham, P. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Sci Rep 7, 41262 (2017). https://doi.org/10.1038/srep41262