**Improved significant SPNs selection for Polygenic Risk Score Calculation from GWAS summary statistics data**

**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 genetic variants that individual has. 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. Since PRSs are calculated from summary statistics data of genome – wide association study (GWAS), the increase in GWAS’s sample size is leading more accurate PRSs of 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 allele’s effect sizes estimated by GWAS. 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. The PRS, S is given by:

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.**

PRS calculation has a couple of steps, the first step involves quality control (QC) check on the GWAS summary statistics data and the target individual’s data. There are several QC methods implemented in various software. PLINK has a robust and easy to use QC algorithm. The second step 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. Selection of significant alleles is done to avoid over fitting the PRS calculation predictive model. This step is followed by the actual calculation of PRS from the selected significant alleles. 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 several PRS calculation software. Standard PRS calculation tools such as PLINK uses Linkage Disequilibrium (LD) and p – value threshold in selecting significant alleles. This p – value threshold is based on Bonferronis adjustment. Bonferroni adjustment suggest that a p – value is significant if:

Where ***n*** is the number of SNPs and alpha is the given level of significant which is mostly 0.05. In this case, every SPNs with LD and p – value that meets the bonferroni adjustment are selected as significant alleles.

The problem with the above stated method is that there is no statistic 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. This way of selecting significant SPNs may end up leaving important alleles that are heavily associated with the trait. It is so difficult to avoid bias in determining the static p – value threshold.

In this project, we are proposing a new simple method of determining significant SPNs from a given GWAS statistic file. Since studies have shown that inclusion of more alleles 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. The improved bonforreni adjustment state that, for any order collection of p – values, if:

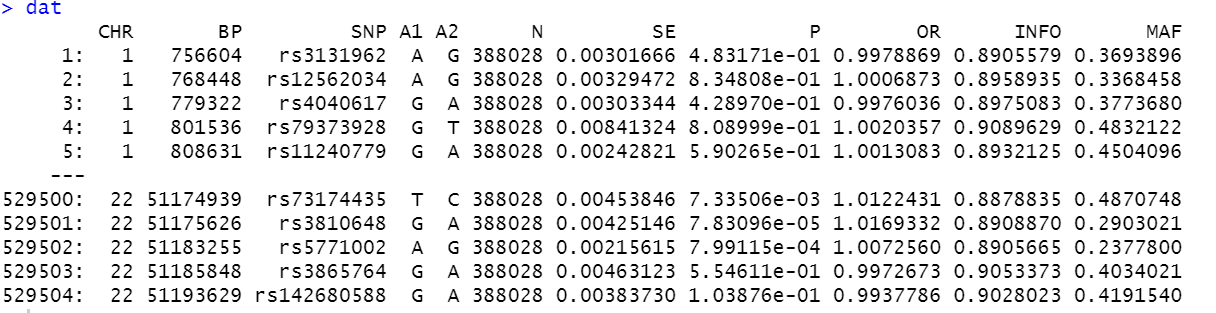
Where ***l***is the position of a p – value in the ordered list and  **> 0**, ***n*** is the number of SNPs and alpha is a given level of significance, mostly 0.05. In this case we are a voiding the use of a static p – value threshold as each and every p – value will be compared to a different threshold. This method reduces bias by subjecting each p – value to its corresponding comparison value. This method 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 type I error probability of equal to alpha.

**Procedure**

In this project, the improved bonferroni adjustment algorithm is implement in R. Some QC steps are also implement to check the quality of the GWAS summary statistics data used in the calculation of the scores. We compared the performance of the algorithm on two levels. The first comparison is made on the number of SNPs selected as significant on both algorithms. The second comparison is made on the PRS score obtained. Due to time limitation, we did not proceed to do results analysis. This process involves performing regression analysis to test the significance of the score to the trait. Due to large of the 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.

**Data**

For demonstration sake, we calculated our PRS from GWAS summary statistics data. We did not have a target file data in this project. We simulated an individual’s phenotype data and combine it with the GWAS summary statistics data using R function. No R biostatistics packages are used in our code. Our GWAS summary data file has 529504 SNPs in total.



In this file the effect allele are the ones under A1. We removed SNPs with a MAF (minor allele frequency) of greater than less than 0.01 and INFO (imputation information score) of less than 0.8 as part of QC since there are more likely to generate false positive results. After removing duplicates, the file had 529493 SNPs. The significant SNPs were selected from this list of 529493 SNPs.

**Results Obtained**

After performing QC on the GWAS summary statistics data, we first run the pruning algorithm based on improved bonferroni adjustment and LD plus p – value threshold by using a static bonferroni adjustment value. It was found out that pruning through improved bonferroni adjustment algorithm produced more significant SNPs than using LD plus static p – value threshold. 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. It was found out improved bonferroni adjustment produced more significant SNPs than the later method. In terms of PRS scores, it is obvious that improved bonferroni adjustment produced a big risk score than the later method.

**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). This process will involve adjusting our code in a way that it should be able to accept the two files give. QC methods for phenotype data file will be also included in the code. We also are 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 increasing being used in predicting human reliability to complex traits. They are playing an 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. Most the proposed techniques usually use external information panels. Our proposed method entirely depends on widely available data, the GWAS summary statistics data.

Comparing with the standard method, PLINK, our method performs better. With proper variable adjustments and inclusion other relevant statistic methods, our approach has a great chance of improving.

**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>