Developing methods to improve and broaden polygenic risk score analyses

Ruan, Yunfeng

Awarding institution:
King's College London

Developing methods to improve and broaden polygenic risk score analyses

Yunfeng Ruan

A Thesis in the Field of Statistical Genetics for the Degree of PhD

Supervised by Dr Paul O'Reilly and Dr Shing Wan Choi

Institute of Psychiatry, Psychology and Neuroscience, Kings College London

Submitted in December 2018

Abstract

Powerful large-scale GWAS not only discover loci associated with various traits but also provide data for performing downstream analyses that lead to further aetiological insights. One such group of downstream analyses involve the use of polygenic risk scores. A polygenic risk score (PRS) is a weighted sum of risk alleles across the genome, which acts as a proxy of an individual's genetic propensity for a phenotype. The weight is usually the effect size estimated from a GWAS on the phenotype. PRS were first used to test whether an outcome had a polygenic basis, especially when the corresponding GWAS yielded no significant results. More recently, PRS have been used for a huge range of applications, most commonly so far to evaluate evidence for shared genetic aetiology between different phenotypes.

In this thesis, I evaluate the power of various PRS analyses by exploiting UK Biobank data, I develop two novel shrinkage methods for increasing the power of PRS analyses, which may have applications beyond GWAS and PRS studies, and finally I develop a set of methods for extending the PRS approach to individual-level gene-set analyses.

My PhD begins by developing a method that we call "Permutation Shrinkage", which shrinks GWAS effect size estimates in order to make them closer to the true effect sizes. The motivation of this method is to improve the PRS prediction model, which is based on GWAS effect size estimates. The accuracy of effect size estimates greatly affects the power of the prediction model based on them. This shrinkage method estimates 'noise' in the observed effect size estimates from a null distribution of the effect sizes generated by permuting raw phenotype data and then subtracting these estimated null effects from the observed estimates. Permutation shrinkage was tested in UK Biobank data. The corrected GWAS leads to an average 35% increase in PRS R^2 across a range of traits tested. In the next chapter, I extend the method to an order statistic method ("Order Statistics Shrinkage") applicable for use on summary statistic data, which is an important extension because most available GWAS data are on summary statistics only. I compare this new shrinkage method to several other well-established shrinkage methods, such as Ridge and LASSO regression and tailed the new method to GWAS data. Order Statistics Shrinkage had similar performance with Permutation Shrinkage in the tests.

In the final work chapter of my thesis, I extend the conventional PRS analysis method to a group of gene-set analysis methods, which we collectively call 'PRSet'. We add PRSet to the PRSice suite of software packages. PRSet calculates gene-set PRS to study aetiology on the gene-set or pathway level. Gene-set analyses can be either self-contained (testing general association) or competitive (testing enrichment compared to other gene-sets). The performance of PRSet is compared with MAGMA, a leading gene-set analysis method.

Acknowledgements

First, I am deeply grateful to my supervisor Dr Paul O'Reilly. His creativity and passion for science deeply influenced me. In addition to all the help and guide on the specific issues, he always encouraged me to have the confidence and patience to investigate and improve the work rather than get depressed by the unexpected or "bad" results. From him, I learnt about how to be a scientist who can think both critically and positively, as well as the knowledge on statistical genetics. It is always a pleasure to work with him.

Second, I cannot imagine finishing my thesis without the help of my second supervisor Dr Shing Wan Choi. His collaboration was an essential part of my PhD project. Meetings with him were usually short but very productive, from which I always benefit. Besides his quick mind, amazing coding skills and profound knowledge, he is a great example for me of what a good postdoc researcher should be.

I would like to extend my thanks to Dr Clive Hoggart for his help and advice on designing and implementing order statistics methods. He was very patient and clear when he discussed with me.

I also wish to thank Dr Jonathan Coleman for helping me with the UK Biobank data, Dr Jack Euesden for having been working on PRSice before I started the PhD project and then introducing the software to me, and Prof Cathryn Lewis, Dr Eva Kraphol, Dr Gerome Breen for discussing the project and giving useful advice. I would like to thank the PhD students who shared the office with me and made my working environment extremely enjoyable.

Finally, I thank my family for supporting me throughout my stay in the UK and my friends for chatting and practicing kendo with me and sending all the funny memes that made me laugh so hard.

Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1.Genetic variation

The human genome consists of 22 autosomal pairs of chromosomes, one from each parent, as well as the sex chromosomes – a pair of X chromosomes in females, and an X and Y chromosome in males. The Human Genome Project characterised the genome, finding that it consists of approximately 3.3 billion nucleotide bases, comprising Adenine (A), Cytosine (C), Guanine (G) and Thymine (T), containing approximately 22000 genes. If two individuals are sampled randomly from the human population then they will have the same nitrogenous base at approximately 99% of their DNA sequence. In terms of base positions that are variable in the human population, around 1 in 100 bases can be defined as being genetic variants where the minor allele has a frequency of at least 1% in the population[1].

Genetic variants come in various forms[2]. Ordered by the scale of variation, the known genetic variants are:

1. Chromosome abnormality, including variation in chromosome number or large translocations.
2. Structural variation of DNA sequence, which are not as large as to cause a considerable change to the chromosome: such as deletions, inversion polymorphisms, insertions, duplication of chromosome fragments, repeat sequences and microsatellites.
3. Variations that occur at the single nucleotide level, such as single nucleotide polymorphisms (SNPs) and small indels (***in***sertions and ***del***etions).

The 1000 Genomes Project estimated that the typical difference between 2 individuals is 20 million base pairs (or 0.6% of the total of 3.2 billion base pairs), the vast majority of which are small differences (SNPs and brief indels)[1].

1.2. History of genetic research

Identifying the genetic basis of human traits and disease is one of the main objectives of genetic research. Historically, the identification of genetic regions that influence a trait can be tracked back to Morgan's experiments[3] with fruit flies in the early 20th century. Morgan's experiments showed that genes are located linearly on the chromosomes and in linkage disequilibrium (LD), which means that the association or linkage of the genotypes of different genetic variants is not random in the population. Since then, various techniques have been developed to identify gene function, the location of genes on the chromosomes and the distance between genes (gene mapping).

In order to identify which genes or regions of the genome are associated with human traits or diseases, there are two main study designs: linkage analyses and association analyses.

Linkage analysis exploit pedigrees and tests the association between molecular markers, such as restriction fragment length polymorphisms (RFLPs), and the phenotype to narrow down the region associated with the phenotype. This method mainly applies to diseases caused by a highly penetrant rare alleles, often caused by *de novo* mutations.

The molecular markers for linkage analysis have various forms. They are not necessarily the DNA sequence with any relevant biological function. Before the Human Genome Project and the wide usage of high-throughput genotyping techniques, such as SNP microarray and next generation sequencing (NGS), the techniques to measure the genetic variations were mainly based on restriction endonuclease, electrophoresis and polymerase chain reaction (PCR)[4], which only indicate the position of genes or other markers on a genome and their relative positions.

After a marker is found to be in linkage with the trait, the location of the casual variant(s) can be approximated. The following-up study is to fine map the causal region with techniques such as yeast artificial chromosome[5] and bacterial artificial chromosome[6] until the candidate region is relatively small. In the case of Huntington's disease, the candidate region was then approximately 500 kb[7]. Then the exons encoded by the candidate region could be amplified to isolate the candidate gene(s)[8].

Linkage analyses have contributed to discovering the genetic basis of a huge fraction of Mendelian and oligogenic diseases and has shed light on how genetic variation affects human traits and diseases. However, linkage analysis has limited power: only risk alleles of strong effect that lead to very severe phenotypes, such as Huntington's disease[9, 7], can be detected. If traits are caused by multiple variants and the risk alleles are of moderate and weak effect, which is true for most common diseases, then the indirect association test via linked markers in relatively small samples of pedigrees may be insufficiently powered. Sample sizes are inevitably limited in linkage analyses because collecting family data is time-consuming and expensive. Moreover, the recombination events in relatively small samples of pedigrees are of limited mapping resolution[10].

By contrast, association analysis uses "unrelated", which are distantly related, samples and directly tests the association between the genotypes of each variant and the phenotype, usually using linear (quantitative traits) or logistic regression (binary traits). Associated analyses are typically substantially more statistically powerful than linkage analyses since they can be more easily performed on very large sample sizes, meaning that they are more able to detect causal variants that have modest effects.[11] Beside, the unrelated samples from the general population are usually more diverse and contain many more historical recombination events, so the mapping resolution of population samples is much higher[10].

The majority of diseases are now known to be polygenic traits, for which thousands of genetic variants with small effects collectively influence the phenotype. As the focus of the field shifted to common diseases and more genetic variants were identified through the association study design, association analysis started to dominate the field of Genetic Epidemiology.

At first, candidate gene association studies were widely conducted. In this design, candidate genes are selected for testing based on the previous biological and clinical knowledge of the traits and the possible causal genes or associated markers that may likely be causal for them. However, the results from early association studies on polygenic traits, typically performed on candidate genes, were usually inconsistent and the studies were underpowered[12]. Negative or inconsistent results can be caused by various factors: the sample sizes are too small, none of the candidate genes are associated with the phenotype despite the phenotype having a genetic basis, or that the phenotype has little or no genetic basis. However, the design of the candidate gene association study makes it impossible to know which of the factors is the case for a particular study. Therefore, there was a critical need for a more powerful experimental design.

Since the Human Genome Project[13] was completed in 2001, the sequence and structure of the whole human genome has been well characterised. In the following-up research, such as the ENCODE[14] project, the human genome was further sequenced and annotated. Public databases, such as the NCBI (https://www.ncbi.nlm.nih.gov/), make the information of the human genome available to the researchers globally. Mapping a gene or a section of DNA sequence to the chromosome has become increasingly easy. At the same time, the cost of high-throughput methods of genotyping genome-wide has dramatically decreased. The current mainstream human genetics assaying methods are DNA microarray (also known as "gene chip") and next-generation sequencing (NGS), which can process thousands of samples and assay millions of alleles within days.

Accumulation of the knowledge about human genome and large and high-resolute data make it possible to design genome-wide association studies (GWAS) (see section 1.3) to detect various genetics variants, even if their effect size is very small. Genetic variants, such as copy number variations (CNVs) and single nucleotide polymorphisms (SNPs), can be accurately measured and systematically studied.

Various studies have shown that common genetic variants, which are mostly detected by microarrays in GWAS, can explain a considerable proportion of polygenic traits (see section 1.3). The aetiological interpretation and prediction of traits can be improved based on these findings and the availability of large-scale biobank data (see section 1.7). In this thesis, we mainly focus on GWAS and SNPs and develop new methodology to further and better exploit GWAS data.

1.3.The GWAS design

A genome-wide association study (GWAS) is a large-scale study design to test if genetic variants across the genome, typically single nucleotide polymorphisms (SNPs), are associated with the phenotype. In GWAS, DNA samples from thousands of individuals are analysed with DNA microarrays, which predominantly assay SNPs.

In terms of which SNPs are tested, unlike candidate gene association studies, GWAS aims to cover the whole genome instead of making assumptions about which genes or regions are likely to be associated with the phenotype. The design of GWAS is not limited to or directed by previous findings.

However, the current GWAS typically genotypes between 500k and 1M SNPs across the genome, which represents only a subset of SNPs across the 3.3 billion nucleotide bases of the whole genome). Since the Linkage disequilibrium (LD) structure across the genome is relatively pervasive, each SNP is typically correlated with many neighbouring SNPs in the same genomic region. SNPs are specifically selected for inclusion on genotyping arrays that cover genetic variation in each region best, and these SNPs are known as “tag SNPs”. By genotyping tag SNPs, we obtain a high fraction of the genetic variation of an individual’s genome with relatively low cost. Moreover, a method known as “genotype imputation”[15] in the field can be applied to infer the genotypes for many additional SNPs by using sequence information, as a reference panel for imputation, from genetic variation projects such as the 1000 Genomes project[16], to increase coverage further. However, despite this relatively high SNP coverage, there are still many SNPs in the genome not included in GWAS, which means that GWAS results indicate only whether there is a variant(s) in a region that contributes to the phenotype under study, but the SNPs with the smallest *P*-values in GWAS are not necessarily the causal SNPs.

GWAS typically assumes that: 1) each causal SNP contributes independently and additively to the phenotype; 2) For each SNP, its genetic effect has an additive mode of inheritance, in that the trait effect is the sum of the effect of each of the two alleles. Under these assumptions, any interaction between SNPs is ignored and the SNPs can be analysed independently. While these assumptions may oversimplify the reality, they are still widely adopted by the field, because the calculations based on these assumptions are straight-forward and yet appear to be reasonably powerful: in simulation tests, the single-marker-based approach was as powerful as haplotype-based approach[17]. Generally, association testing between SNP genotypes and the phenotype are performed using linear regression for continuous phenotypes and logistic regression for binary phenotypes. A major advantage of performing regression compared to simpler tests, such as the chi-squared test or Armitage trend test, is that covariates can be controlled for. In GWAS a major potential confounder is that of population structure[18], whereby genotypes and phenotypes can be associated due to non-random mating (mostly by geographical location), but one way to overcome this issue is to include principal components from a principal component analysis (PCA) of the genome-wide SNP data as covariates in the regression models.

When performing GWAS we encounter the "multiple testing problem": since thousands or even millions of tests are performed, the usual statistical significance level of $P < 0.05$ is too liberal because a huge number of tests would be declared as significant just by chance if this threshold was applied. Due to LD structure, the effective number of tests performed in GWAS is much lower than the number of SNPs on SNP chips. Therefore, testing 2-3M genotyped and imputed SNPs is equivalent to testing approximately 1M independent SNPs. Based on this, the genome-wide significance threshold applied in GWAS is typically 5×10^{-8} [1], equivalent to a Bonferroni correction on 1M independent tests. In GWAS, usually only common SNPs, defined as the SNPs of minor allele frequency (MAF) > 1%, are included in analyses because testing rare SNPs (MAF < 1%) is typically underpowered, more prone to genotyping errors and false positives, and thus considered a relatively poor investment. However, as sample sizes grow, with the availability of data from large-scale studies such as the UK Biobank, it is becoming more worthwhile to test the association of rare variants for complex traits.

1.4.Findings from GWAS

GWAS has proven to be a fast and powerful way to identify the genetic basis of both oligogenic and polygenic traits. Before the Human Genome Project and the availability of large-scale GWAS, finding the causal genetic variants or gene influencing a disease, even a Mendelian one, could be extremely time-consuming and expensive since multiple rounds of tests are typically needed to narrow down the candidate region. The first genetic marker associated with

Huntington's Disease was found in 1983[9]. After 10 years of recurrently searching for more accurate markers and narrowing down the candidate region, the causal gene was finally identified[7]. Nowadays, if the GWAS is powerful enough, the candidate regions can be found in a single GWAS analyses, which can take as little as a few hours on already collected data.

GWAS results provided a comprehensive picture of how much the common SNPs contribute to the phenotype, especially when the sample size is large enough to detect the signals of SNPs with very small effect size. In 2007, the first large-scale GWAS tested 7 common diseases with 2000 cases for each disease and 3000 shared controls. 24 independent loci were found to be associated with these diseases[19]. Since then more large-scale GWAS have been conducted to identify the new risk alleles and reveal aetiological insights.

Consider schizophrenia disorder, for example: in its early stage, GWAS used sample sizes of several thousand individuals. The first GWAS on schizophrenia[20] used only 3,322 cases and 3,587 controls. However, GWAS on such a sample size is relatively underpowered for very complex polygenic traits such as psychiatric disease. In the first schizophrenia GWAS study, none of the loci passed the genome-wide significant threshold. Later, thanks to more collaborative work across institutions and big data sets such as national biobanks, GWAS for polygenic traits such as schizophrenia became more powerful. More and more associated loci were identified in GWAS. In 2014, when the sample size increased to 36,989 cases and 113,075 controls, 108 loci were identified as to be associated with schizophrenia[21]. The GWAS of schizophrenia shows that it is not only quantitative traits, such as height[22] and BMI[23], that can be explained by common variation, but that even rare binary traits and disesases[24] that result from the interaction between genetics and environment can be explained by common genetic variation.

There has been a debate in the field for many years about whether common diseases or complex traits are caused by common variants (MAF > 1%) or rare variants (MAF< 1%)[25]. However, the findings of many common variants affecting complex disease does not reject the hypothesis that rare variants have an important influence on complex diseases. In fact, the role of rare variants can be better estimated as the influence of common variants becomes better estimated.

1.5.Aetiological insights into complex traits based on GWAS findings

GWAS findings provide not only information about significant associated loci but also aetiology insights. Here, some of the main areas of progress relating to aetiological understanding that has been made thanks to the findings of GWAS are reviewed.

1.5.1. Heritability estimation

Heritability is defined as "the proportion of the phenotypic variation that can be explained by the genetic variation". In the broad-sense heritability, the variations of phenotype can be explained by genetic factor, environmental factor and interaction between the two.

$$\text{Var(Phenotype)=Var(Genetic)+Var(Environment)+2 Cov(Genetics, Environment)}$$

Ideally, the covariance between genetic factors and environmental factors should be controlled to be zero. Here, *Var(Genetic)* includes all the genetic variations, such as additive, dominant, epistatic. The broad-sense heritability is:

$$H^2 = \frac{Var(genetic)}{Var(Phenotype)}$$

The narrow-sense heritability is defined as:

$$h_2 = \frac{Var(additive\ genetics)}{Var(Phenotype)}$$

Only additive genetic variants are included. As explained in the previous section, GWAS make the additive assumption. Therefore, narrow-sense heritability is essentially equal to the maximum of how much the genetic variations detected by GWAS (or similar designs) can explain the phenotype variation.

Note that the estimation of heritability is not simply adding up all the variations that is directly observed in the GWAS. It is to estimate the upper-limit of the contribution of genetic variations to the phenotype according to the observed data. There are two main types of heritability estimation method.

1) GCTA-GREML method

GCTA-GREML[27] uses raw genotype GWAS data to estimate narrow-sense heritability. The basic idea of this method is to compare the genetic similarity and the phenotypic similarity between two unrelated individuals.

The genetic similarity of the samples is denoted in a genetics relationship matrix (GRM) containing the genetic distance between individuals. The genetic relationship between individual i and j based on SNPs i=1,2,…N, with genotypes $x_i \in \{0,1,2\}$, is：

$$A_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

in which p_i is the minor allele frequency (MAF) of SNP i.

The main underlying concept of this method is the mixed linear model (MLM): the phenotype is a function of a group of variables of fixed effect such as sex, age, eigenvectors from principle component analysis (PCA), and SNPs with random effects:

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{Wu} + \varepsilon$$

in which y is the phenotype vector; β is the vector of variables with fixed effects; W is the standardized genotype matrix, u ~ N (0, $\boldsymbol{I}\sigma_u^2$).

Wu can be written as $\boldsymbol{g}$, a vector of total additive genetic effects of the individuals. $\boldsymbol{g} \sim N$ (0, $\mathbf{A}\sigma_g^2$). **A** is the genetic relationship matrix (GRM).

The variance components of this model can be written as:

$$\mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_\varepsilon^2$$

Based on the GRM, the variance explained by all the SNPs σ_g^2 is estimated by restricted maximum likelihood (REML) approach. In this way, the narrow-sense heritability can be estimated.

2) LDSC method

LD Score Regression (LDSC)[28] uses summary statistics data to estimate narrow-sense heritability. The basic idea of this approach is to assume that the observed effect sizes are the combination of real polygenic signal and confounding biases and that the more causal variants a SNP is correlated with, then the larger its observed effect size will be. Thus, LDSC assumes that traits with high heritability will have a strong association between the SNP-phenotype association test statistic (eg. chi-squared statistic) and local LD, while traits with low heritability will have a weak association between the test statistic of SNPs and local LD.

First, for any genetic variant *j*=1, 2, …M, its LD score is defined as:

$$l_j := \sum_{k=1}^{M} r^2{}_{kj}$$

The χ^2 statistics of genetic variant *j,* given its LD score l_j, is:

$$E\left(\chi^2 | l_j\right) = \frac{Nh^2 l_j}{M} + Na + 1$$

where *N i*s the sample size; *M* is the number of SNPs; *a* measures the confounding biases. In the regression model of χ^2 statistics from GWAS results on the LD score, the slope can be rescaled as an estimate of heritability, explained by the *M* SNPs used to build this LD score model.

1.5.2. Polygenic nature of complex traits

GWAS findings have demonstrated that most complex traits are polygenic[26]. Attention is mainly focused on the loci that pass genome-wide significance when the aim is to identify the causal variants, but studies show that not only genome-wide significant SNPs contribute to the phenotypes. Less statistically significant SNPs turn out to explain the phenotypes and the correlations between different traits. For example, in Purcell *et al* (2009) the authors used polygenic risk scores to argue that schizophrenia had a polygenic basis, which was shared by bipolar disorder[20]. In this study, the polygenic risk score based on the both genome-wide significant and less significant SNPs, that is, PRS using different SNP *P*-value thresholds, significantly predicted the phenotype.

As the sample size of GWAS increased, more SNPs were found to be genome-wide significantly associated with complex traits. For instance, after Purcell *et al*[20], more GWAS on schizophrenia were conducted using Caucasian samples and the number of genome-wide significant loci increased to 5[27], 22[28], and 108[21] as the sample size increased. This indicates that the complex traits are collectively influenced by SNPs of small effect size and more SNPs of even smaller effect size may be found in the future as sample sizes increase.

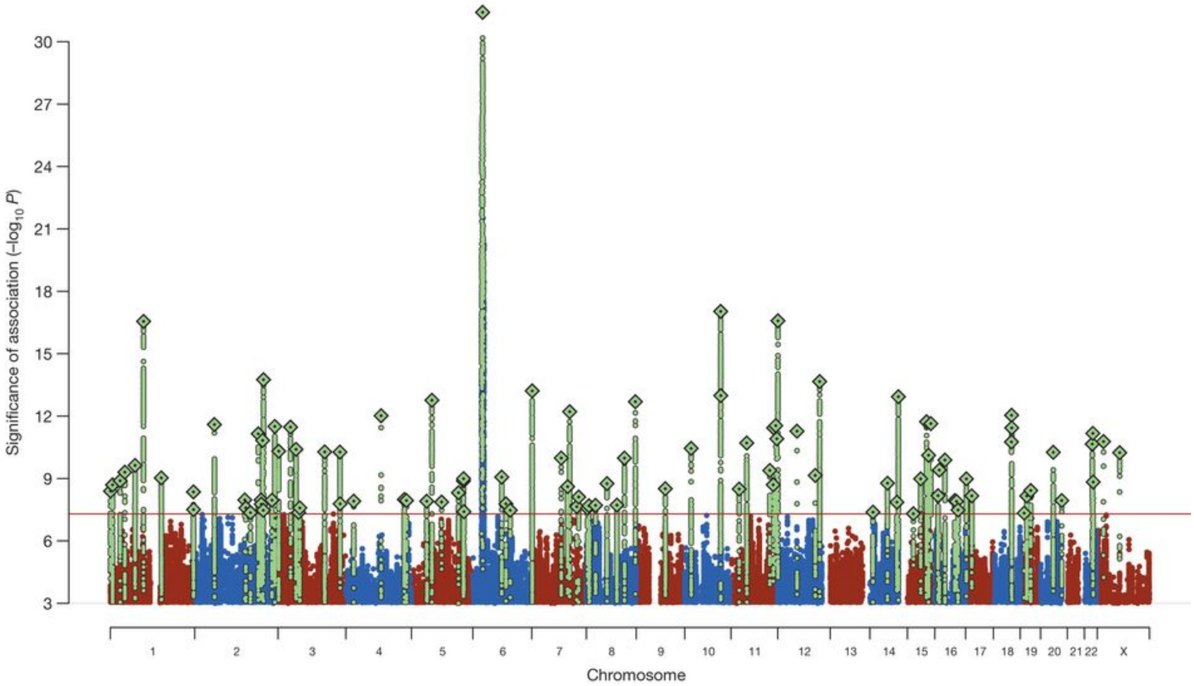


Figure 1 Manhattan plot showing the 108 genome-wide significant loci found in the GWAS study by the PGC[21]. The red horizontal line shows the genome-wide significance threshold; the diamonds show the significant index SNPs, while the SNPs in LD with the index SNPs are in green.

The polygenicity means that each causal genetic variant has a small or modest effect on the phenotype. Therefore, it is difficult to study polygenic phenotypes based on each single locus. Polygenic Risk Scores (PRS) are a useful tool to represent the genetic burden across the genome and have become widely-used in research on polygenic traits[29]. The implementation and application of PRS will be further discussed in section 1.6.

1.5.3. Genetic correlation between phenotypes

GWAS can be used to estimates the genetic correlation between different phenotypes. There are two popular methods for testing the correlation. First, a polygenic risk score (PRS) based on one trait can be used to predict another trait. More detailed introduction of PRS will be given in 1.6. In short, a PRS is a weighed sum of risk alleles relating to an individual. The effect size estimated in a previous GWAS can be used as the weight. PRS can be used as a proxy of the genetic burden for a phenotype. For example, if the PRS that uses the schizophrenia GWAS as the weight represents a person's genetic burden for schizophrenia. If the schizophrenia PRS can predict other phenotypes, e.g. the risk of bipolar disorder, it indicates that schizophrenia is genetically correlated with other phenotypes. Using this PRS method, the International Schizophrenia Consortium showed that schizophrenia shared a common genetic basic consisting of common polygenic variations with bipolar disorder, but did not appear to share a common genetic basis with non-psychiatric disease (coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type I diabetes and type II diabetes) [20]. Although none of the SNPs reached genome-wide significance in this early-stage study, the association between schizophrenia PRS and the bipolar disorder (BD) phenotype was significant and the smallest P-value in this study is 1×10^{-12}.

Another method is cross-trait LD Score regression[30]. In addition to calculate heritability, LDSC can be used to calculate genetic correlation by adjusting the model. In a large scale study, Bulik-Sullivan *et al* tested the correlation between 24 traits with this methods[31]

The model that calculates the heritability (see section 1.5.1) can be adapted to cross-trait data. Assume that the two GWAS each have sample sizes of N_1 and N_2 and have N_s overlapping samples. The genetic correlation between the two traits is ϱ_g and the phenotypic correlation between the two samples is ϱ. Since the LD structure is the same in the two samples, the LD score remains the same. The Z score of the variants j in the two GWAS is z_{1j} and z_{2j}. The original model can be written as

$$\mathrm{E}\left(z_{1j}z_{2j}\middle|l_j\right) = \frac{\sqrt{N_1N_2}\varrho_g l_j}{M} + \frac{\varrho N_s}{\sqrt{N_1N_2}}$$

The slope of the regression model of $z_{1j}z_{2j}$ regressed against LD score can be used to estimate the genetic correlation.

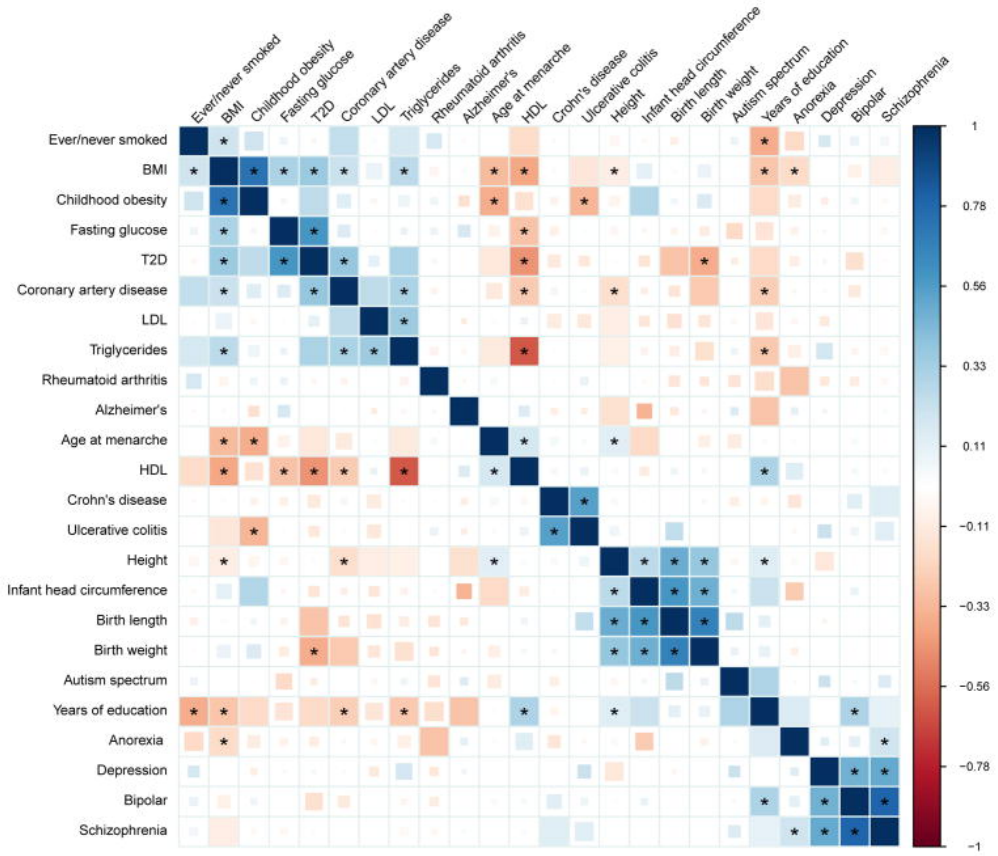


Figure 2 The genetic correlations estimated by LD score regression[31]. The size of the coloured squares indicates the P-value of the genetic correlation. The larger the square, the more significant the P-value. The result of the correlation difference from zero, at a false discovery rate of 1%, is shown as full-size squares. The asterisks mark the results that are significant after Bonferroni correction for the 300 tests. The colour and shade indicate the value of genetic correlation, blue for positive and red for negative. The darker the shade, the higher the absolute value.

As shown in Figure 2, statistically significant genetic correlations were found among various traits. Notably, no SNPs reached genome-wide significance for anorexia nervosa and only three for educational attainment in this study, but significant genetic correlations were observed between these two traits and other traits.

These studies showed that even when only a few or no SNPs were found to be genome-wide significant, the genetic correlations could still be detected with a statistically significant result. These findings show that the variants of small or modest effect sizes contribute to the correlations between different polygenic traits.

1.6.Polygenic risk score

1.6.1. Basic principles

Polygenic risk score (PRS) or polygenic score is the weighted sum of alleles that may contribute to phenotype.

$$\mathrm{PRS} = \sum \beta_x \, x$$

The weight is usually the effect estimates from previous research such as GWAS. The effect estimates or the GWAS data that produce these estimates is usually called base data. The individual raw genotype and phenotype data based on which the prediction is made is usually called target data. Admittedly, the SNPs included in the model are not necessarily the causal SNPs but the PRS model can be used to represent the genetic burden of the individual.

Despite the simple and straight-forward rational of PRS, it is necessary to pay attention to some details when calculating PRS:

First, the base data and the target data should be independent but homogeneous samples drawn from the same populations. The overlapping between the base and target data leads to overfitting[32]. If the base and target data are from different populations, the model can be under-powered or biased because of the differences in allele frequency, population structure, environmental factors etc[29]. Although there have been successful multi-ethnic PRS analysis[33], PRS based on cross-population or heterogenic samples required great caution to correct for the possible confounding factors.

Second, the SNPs in LD will cause redundant signals in PRS. One of the commonly used methods to avoid the redundant signal is to clump SNPs before calculating the PRS.

Third, missing SNPs need to be properly corrected for. In PRS calculation, missing SNPs are usually set to be zero. This will introduce a bias correlated with the missingness. To correct for this bias, in the QC step before calculating the PRS, samples with too many missing SNP should be excluded; in the PRS calculation, the raw $\sum \beta_x \, x$ should be divided by the number of SNPs (excluding all the missing SNPs) used in the PRS calculation.

$$\mathrm{PRS}' = \frac{1}{N} \sum \beta_x \, x$$

PRS' is the average genetic burden per non-missing SNP account for. Therefore, it is not biased because of the different missingness of the sample.

PRS should be controlled for covariates such as sex, age, top principle components, etc. when testing the association between the phenotype and PRS.

1.6.2. Methods of optimising PRS

The raw GWAS result contains SNPs that are correlated or in LD. Calculating PRS using all the SNPs will lead to redundant signals. Usually, SNPs are clumped before the calculation of PRS. Clumping is a function that implemented by PLINK[34]. Clumping takes all the SNPs that passes a *P*-value threshold and have not been clumped (*index SNPs*) and put all the other SNPs within a physical distance measured in kilobase that are correlated with the *index SNP* because of LD into a clump. The correlation is measured in the R^2. Clumping is a greedy algorithm so that each SNP is only included in at most one clumps. Thus, clumped SNPs can be viewed to be independent or only modestly correlated and are the most significant SNPs in each clump.

PLINK provide 4 parameters for clumping: p_1, the significant threshold SNPs to be considered as *index SNPs*; p_2, the significant threshold for non-index SNPs to be listed in the output file; kb: the distance within which the clumping is performed; r^2: the correlation threshold of for the non-index SNP to be considered with the index SNP. In PRS calculation, the aim of clumping is to extract all the independent SNPs so p_1 is usually set to be 1 and kb and r^2 can be adjusted according to each specific analysis.

In GWAS, the effect size estimates contain true signal as well as error quantity caused by chance ('noise'). When using the GWAS as the base data for PRS model, reducing the 'noise' in the model will increase the prediction power.

The most common way to optimising PRS is P-value thresholding. It is intuitive to assume that the smaller P-value, the more likely the effect size estimate is significant and close to the truth. The SNPs that pass the P-value threshold are viewed to have the valid effect size estimates and included in the PRS and the other SNPs are excluded.

Since different traits and data may have different genetic architecture, P-value thresholding method iterates through different P-value thresholds in order to get the best signal-noise ratio. The best results of all the iterations is viewed to have the highest the signal-noise ratio and used as the final PRS result. P-value thresholding on the clumped SNPs (Clumping + P-value thresholding, or "C+P") is currently the most used approach for calculating PRS[29].

There are also other ways of improving the PRS prediction power. Two alternative options of the clumping and thresholding approach to optimize PRS, LDpred[35] and lassosum[36], are briefly introduced here.

LDpred[35] uses a Bayesian approach that uses the genetic architecture and LD structure information as the prior instead of clumping when controlling for the LD structure. The posterior mean for effect size is approximated as the following formula with some assumptions:

$$E(\beta^l \mid \tilde{\beta}^l, D) \approx \left(\frac{M}{Nh_g^2} I + D_l\right)^{-1} \tilde{\beta}^l.$$

where N is the number of individuals; M is the number of genetic variants; D_l denotes the regional LD matrix under the assumption that only makers within a small region are linked; and h_g denotes the heritability estimated with LD score regression[36]. In addition, it iterates a series of possibility of variants being causal, p:

$$\beta_i \sim \begin{cases} N\left(0, \frac{h^2}{Mp}\right) \text{with probability } p \\ 0 \text{ with probability } (1-p) \end{cases}$$

The default series of p in the methods is 1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 3E-4, 1E-4, 3E-5, 1E-5. The SNP weights generated by the above model is then used to build the PRS model.

Lassosum[36] optimises the prediction power with a modified LASSO model that shrink the effect size estimates in the base data. The original LASSO[37] method is a penalized regression that depends on the raw individual-level data:

$$\begin{aligned} f(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}) + 2\lambda||\boldsymbol{\beta}||_1^1 \\ &= \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + 2\lambda||\boldsymbol{\beta}||_1^1 \end{aligned}$$

where β is the estimated effect size; $\|\beta\|_1^1 = \sum_i |\beta_i|$; ***X*** is the individual-level genotype matrix, ***y*** is the phenotype vector.

The raw individual level data $\boldsymbol{X^T X}$ and $\boldsymbol{X^T y}$ can be replaced by summary statistics. $\boldsymbol{r} = \boldsymbol{X^T y}$ is the SNP-wise correlation between the SNPs and the phenotype; $\boldsymbol{R} = \boldsymbol{X^T X}$ is the LD matrix. Both ***r*** and ***R*** can be estimated from other summary statistics databases. Therefore, the regression can be written as

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{R\beta} - 2\boldsymbol{\beta}^T\boldsymbol{r} + 2\lambda||\boldsymbol{\beta}||_1^1$$

Since ***R*** is estimated from another set of genotype data $\boldsymbol{X_r}$, the model is modified as:

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}_r^T\boldsymbol{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + 2\lambda||\boldsymbol{\beta}||_1^1$$

The model is not a LASSO problem. In order to make the modified model equivalent to a LASSO problem, the formula is regularized:

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{R}_s\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{r} + 2\lambda||\boldsymbol{\beta}||_1^1$$

Where $\boldsymbol{R}_s = (1 - s)\boldsymbol{X}_r^T X_r + s\boldsymbol{I}$ (0<s<1).

A set of BLUP methods[38–41] have been developed to improve the SNP effect size estimates. GWAS estimates the effect size of each SNP separately regardless of the LD structure. Using all the SNP effect size estimates can be problematic because of high collinearity of the SNPs. Clumping can get rid of the collinearity but may exclude informative SNPs. To address this problem, Genomic BLUP (GBLUP) uses a linear mixed model (LMM) with best linear unbiased predictor (BLUP) properties to improve the ordinary least squares OLS predictors of original GWAS estimates.

In a general linear mixed model:

$$\mathbf{y} = \mathbf{Wb} + \epsilon$$

In which $\mathbf{y}$ denotes the phenotype; $\mathbf{b}$ denotes the SNP effects; ϵ denotes the residues; $\mathbf{W}$ denotes the standardized genotype, where the ij^{th} element (the risk allele number of i^{th} individual's j^{th} SNP) is standardized as:

$$w_{ij} = \frac{(x_{ij} - 2p_j)}{\sqrt{2p_j(1 - p_j)}}$$

The distributional properties are denoted as var($\mathbf{b}$) = $\mathbf{B}$, var($\epsilon\epsilon$) = $\mathbf{R}$ and var($\mathbf{y}$) = $\mathbf{WBW}' + \mathbf{R}$.

Usually, GWAS calculate the effect of one SNP at a time using OLS regression as:

$$\hat{\mathbf{b}}_{\mathrm{OLS}} = \mathrm{diag}[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\mathbf{y}$$

Where $\mathrm{diag}([\mathbf{W}'\mathbf{W}])$ has diagonal elements = $w_j{}'w_j$ and off-diagonal elements = 0.

In a general form, BLUP solution for $\mathbf{b}$ based on individual-level data is

$$\hat{\mathbf{b}}_{\mathrm{BLUP}} = [\mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \mathbf{B}^{-1}]^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{y}$$

It accounts for the correlation between variants and also is unbiased in terms of $\mathrm{E}\big[\mathbf{b}|\hat{\mathbf{b}}\big] = \hat{\mathbf{b}}$**.**

If **R** is diagonal, the above equation can be reduced to:

$$\hat{\mathbf{b}}_{\mathrm{BLUP}} = [\mathbf{W}'\mathbf{W} + \mathbf{B}^{-1}\mathbf{R}]^{-1}\mathbf{W}'\mathbf{y}$$

Assuming **b** follows the distribution $\mathbf{b}{\sim}N(0,\ \mathbf{I}_M\sigma_b^2)$, then $\mathbf{B} = \mathbf{I}_M\sigma_b^2$; assuming the residue $\epsilon{\sim}N(0, \mathbf{I}_M\sigma_\epsilon^2)$. The above equation can be written as:

$$\hat{\mathbf{b}}_{\mathrm{BLUP}} = [\mathbf{W}'\mathbf{W} + \mathbf{I}_M\lambda]^{-1}\mathbf{W}'\mathbf{y}$$

In which $\lambda = \frac{\sigma_\epsilon^2}{\sigma_b^2}$.

The above is the solution of genomic BLUP.

When the individual-level raw GWAS data is not available, BLUP estimates can be obtained from summary statistics GWAS result by replacing the individual level data $\mathbf{W}'\mathbf{W}$ and $\mathbf{W}'\mathbf{y}$ with the expectation that can be obtained with public available data.

$$\mathbb{E}[\mathbf{W}'\mathbf{W}] = N\mathbf{L}$$

$$\mathbb{E}[\mathbf{W}'\mathbf{y}] = N\hat{\mathbf{b}}_{\mathrm{OLS}}$$

In which L is an M × M scaled SNP LD correlation matrix estimated from a reference SNP data set and $\hat{\mathbf{b}}_{\text{OLS}}$ are obtained from GWAS summary statistics.

Assuming phenotype variance =1 and the proportion of phenotypic variance contributed by SNPs $h_{SNP}^2 = M\sigma_b^2$, then $\lambda = \frac{M\sigma_\epsilon^2}{h_{SNP}^2} = \frac{M(1-h_{SNP}^2)}{h_{SNP}^2}$.

Thus, GBLUP can be transformed to summary statistic approximate BLUP (SBLUP):

$$\hat{\mathbf{b}}_{\text{SBLUP}} = [N\mathbf{L} + \mathbf{I}_M\lambda]^{-1}N\hat{\mathbf{b}}_{\text{OLS}}$$

$$\hat{\mathbf{b}}_{\text{SBLUP}} = [\mathbf{L} + \mathbf{I}_M\lambda/N]^{-1}\hat{\mathbf{b}}_{\text{OLS}}$$

This method[40] is similar with LDpred.

In addition to optimizing SNP effect size estimates of a single trait, the genetic predictors of multiple traits can be jointly analysed with in a mixed-effects model to improve the prediction power if the traits are genetically correlated. The above BLUP model can be adjusted for multiple traits. For *k* traits that are measured on different individuals, with N_k observations for trait *k,* the phenotype and genotype data can be written as:

$$\mathbf{y}' = \mathbf{Wb} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}' = [\boldsymbol{y}'_{\mathbf{1}}, \boldsymbol{y}'_{\mathbf{2}} \dots \boldsymbol{y}'_{\boldsymbol{k}}]$, $\mathbf{W} = \begin{bmatrix} \mathbf{W_1} & \mathbf{0} & \cdots \mathbf{0} \\ \mathbf{0} & \mathbf{W_2} & \cdots \mathbf{0} \\ \vdots & \vdots & \ddots \vdots \\ \mathbf{0} & \mathbf{0} & \cdots \ \mathbf{W}_{\boldsymbol{k}} \end{bmatrix}$.

The distribution of the parameters is $\text{var}(\epsilon) = \mathbf{R} = \text{diag}(\mathbf{R}_k) = \text{diag}[\mathbf{I}_{N_k}\sigma_{\epsilon_k}^2]$, $\text{var}(b) = B = \sum_{\boldsymbol{b}} \otimes \mathbf{I}_M$. $\sum_{\boldsymbol{b}}$ is a k × k matrix, with diagonal elements $\sigma_{b_k}^2$ and off-diagnoal elements the covariances of SNP effects between traits. Substituting these expressions into $\hat{\mathbf{b}}_{\text{BLUP}}$ will generate multi-trait BLUP solution for effect size estimates:

$$\hat{\mathbf{b}}_{\mathrm{MT-BLUP}} = \left[\mathbf{W}'\mathbf{W} + \Sigma_{\epsilon}\Sigma_{\boldsymbol{b}}^{-\mathbf{1}} \otimes \mathbf{I}_{M}\right]^{-1}\mathbf{W}'\mathbf{y}$$

If only summary statistics is available, $\mathbf{W}'\mathbf{W}$ and $\mathbf{W}'\mathbf{y}$ can be replaced by their expectation as in the single-trait BLUP:

$$\mathbb{E}[\mathbf{W}_{\boldsymbol{k}}{}'\mathbf{W}_{\boldsymbol{k}}] = N_{k}\mathbf{L}$$

$$\mathbb{E}[\mathbf{W}'\mathbf{y}] = N_{k}\hat{\mathbf{b}}_{\mathrm{OLS}_{k}}$$

Thus, the multi-trait summary statistics BLUP is:

$$\hat{\mathbf{b}}_{\mathrm{MT-BLUP}} = \left[\mathbf{I}_{k}\otimes\mathbf{L} + \Sigma_{\epsilon}\Sigma_{\boldsymbol{b}}^{-\mathbf{1}}\mathbf{N}^{-1} \otimes \mathbf{I}_{M}\right]^{-1}\hat{\mathbf{b}}_{\mathrm{OLS}}$$

However, the inversion of non-diagonal matrix $\left[\mathbf{I}_{k}\otimes\mathbf{L} + \Sigma_{\epsilon}\Sigma_{\boldsymbol{b}}^{-\mathbf{1}}\mathbf{N}^{-1} \otimes \mathbf{I}_{M}\right]$ is highly computational expensive. Thus, the method assumes that SBLUP solutions have BLUP properties: $\mathrm{cov}\left(\mathbf{b}_{k}, \hat{\mathbf{b}}_{\mathrm{SBLUP}_{k}}\right) = \mathrm{var}\left(\hat{\mathbf{b}}_{\mathrm{SBLUP}_{k}}\right) = \mathrm{var}\left(\hat{\mathbf{b}}_{\mathrm{BLUP}_{k}}\right)$ and thus independent LD reference sample can be used to generate approximate solution. The method is implemented in SMTpred (https://github.com/uqrmaie1/smtpred)[41].

Please note that all these above methods may have the overfitting problem. It is recommended that after the correction, the optimized prediction model is validated with an out-of-sample data (see section 2.3.7).

In Chapter 2 and Chapter 3, I developed new shrinkage methods working on the GWAS effect size estimates.

1.6.3. Applications

PRS represent the genetic burden that makes an individual vulnerable get an outcome. Regressing the outcome on PRS or stratifying the individuals according to their PRS can reveal the aetiology and facilitates clinical therapy.

The earliest application of PRS is to test whether the polygenic genetic basic exists when no or few SNPs are genome-wide significant. For example, in an early-stage GWAS of schizophrenia and bipolar disorder in 2009[20], the association between schizophrenia PRS and the phenotype showed that thousands of common SNPs of small effect sizes contribute to the schizophrenia phenotype.

In more recent studies where the sample size become larger and more genome-wide significant SNPs were found, PRS was used to study the aetiology of the traits and investigate the interaction between the genetic factor and the environmental factors.

For example, stratification of samples according to their PRS showed that the accumulation of genetic burden led to more severe/extreme outcome or higher chance to have the outcome. For example, in the work of Selzam *et al* on the PRS of total year of enducation[42], individuals with higher PRS had better educational achievement as shown in Figure 3. In the work of Wray *et al* on major depression[43], the similar pattern was found that individual with higher depression PRS had higher risk of depression.

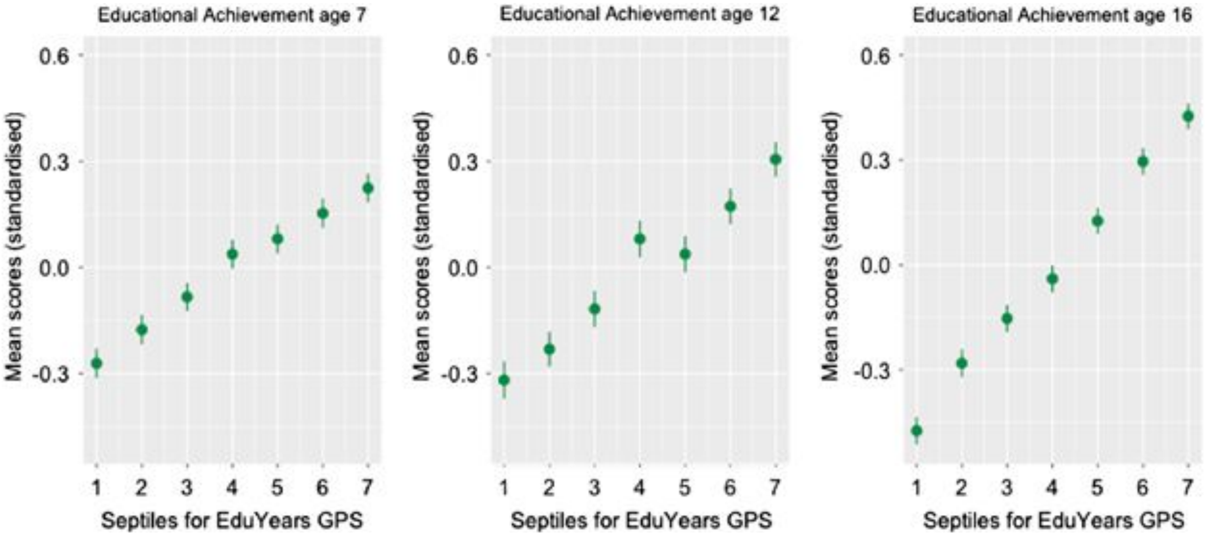


Figure 3 Individuals with higher Education year PRS have better educational achievement. Graph adapted from research by Selzam et al 2017[42]. The x-axis shows the groups of individuals divided according to their education polygenic risk score; y-axis shows their educational achievement at different ages. EduYears GPS: Genome-wide polygenic score for total year of education. The mean and standard error were standardized.

Besides, PRS was also evidence to show that genetic burden wound influence which disease subtype an individual might develop. For example, schizophrenia PRS in patients with schizoaffective BD was significantly higher than patients with non-schizoaffective BD[44], bipolar disorder PRS in patients with psychotic BD was significantly higher than patients with non-psychotic BD[45]. The work of Wray *et al*[43] also showed that individuals with higher PRS tended to develop severer subtype of depression.

In addition to the relation between genetic basis and outcome, PRS can be used to study the interaction between genetic and environmental factors. Meyers *et al* shows that cigarette use PRS predicted the number of cigarettes smoked per day and interacted with traumatic events and neighbourhood social cohesion in Africa America samples[46]. A series of studies investigate the interaction between schizophrenia PRS and childhood adversity[47–49] in European Caucasian samples and gave different results, which indicated that the interaction testing might be statistically vulnerable[49].

Cross-trait PRS analysis can reveal whether the two traits are genetically correlated. In the research on schizophrenia and bipolar disorder mentioned above, PRS based on schizophrenia and bipolar disorder GWAS could predict the schizophrenia and bipolar disorder but could not predict non-psychiatric traits such as coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type I diabetes and type II diabetes[20]. This showed that schizophrenia and bipolar disorder may share genetic basis.

Besides, several studies attempted to facilitate clinical decision with cross-trait PRS analysis. Higher schizophrenia PRS tended to respond less to antipsychotic drug treatment[50]. So far, no neither major depression PRS nor schizophrenia PRS predicted the response to antidepressants[51].

1.7.Biobank

PRS requires powerful large-scale data to have good prediction performance[52]. It is usually difficult for individual research group to collect thousands or millions of samples. Therfore, the availability of biobank data greatly boosts the research on PRS and other statistical genetic methods. A biobank is a repository that collects and stores biological samples and data for research purpose. It provides large-scale samples and data to multiple research programmes.

UK Biobank (https://www.ukbiobank.ac.uk/) is a large-scale biorepository aiming to facilitate researches on the cause of the diseases and improve the prevention, diagnosis and treatment. It collected 500,000 people aged between 40-69 across the country. The participants donate their blood, urine and saliva samples for genotyping and provide information of various phenotypes. UK Biobank enables large-scale researches by providing genotype and various phenotype data of millions of anonymous samples.

1.8.Shrinkage methods

As described in section 1.6.2, in order to improve the prediction of the polygenic risk score model, the effect size estimates from the base GWAS data can be shrunk. Shrinkage methods have been long used to improve prediction accuracy, especially for high-dimensional data. They usually constrain the value of the regression coefficients and/or reduce the number of predictors in the model. The rationale behind shrinkage methods is that the raw coefficients contain stochastic variation or "chance association" with the outcome in the training data. Removing the predictors that contains too much noise or constraining their coefficients can improve the power of the prediction model.

Many parameter-based shrinkage methods were designed to correct the result of ordinary least squares (OLS) regression, which is the most commonly approach used for multiple linear regression:

In OLS, a dependent variable y is the linear combination of k independent variables x_i:

$$y_j = \sum_{i=1}^{k} \beta_i x_{ij} + \varepsilon_j$$

The model can be written in matrix form as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

OLS aims to generate the estimate $\boldsymbol{\beta}$, i.e. $\widehat{\boldsymbol{\beta}}$ to minimize the residues of the model:

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\ ||\boldsymbol{y} - \boldsymbol{X\beta}||^2$$

The OLS estimate is:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

However, OLS estimates have large variance[37] and only work well when multicollinearity between predictors is low or absent[53]. Shrinkage methods were developed to overcome the pitfalls of OLS. Among the most popular methods for shrinkage are the James-Stein estimator[54], LASSO[37], and Ridge regression[53].

The James-Stein estimator is an empirical Bayes method. The prior distribution, a single parameter μ, is estimated from observation *x*

$$\mu \sim N(M, A) \ \text{ and } x|\ \mu \sim N(\mu,\ 1)$$

Here, μ has posterior distribution:

$$\mu|x{\sim}N(M+B(x\text{-}M),\ B)$$

in which $B=A/(A+1)$.

If *M* and *A* are unknown, they are to be estimated from x= (x$_1$, x$_2$, …, x$_N$):

$$\widehat{M} = \overline{\boldsymbol{x}}$$

$$\widehat{B} = 1 - \frac{N-3}{S} \text{ in which } S=\sum_{i=1}^{N}(x_i - \bar{x})^2$$

The James-Stein estimator is:

$$\hat{\mu}^{JS} = \widehat{M} + \hat{B}(x - \widehat{M})$$

The ridge regression and LASSO methods set an upper-limit for the sum of the squared value or absolute value of all the estimates, respectively. LASSO shrinkage sets an upper-limit for the sum of the absolute value of the parameters as follows:

$$\min_{\beta \in \mathbb{R}^P} \{\frac{1}{N} \|y - X\beta\|_2^2\} \ subject \ \ to \ \|\beta\|_1 \leq t$$

Ridge shrinkage sets the upper-limit for the sum of the square of the parameters.

$$\min_{\beta \in \mathbb{R}^P} \{\frac{1}{N} \|y - X\beta\|_2^2\} \ subject \ \ to \ \|\beta\|_2^2 \leq t$$

On the other hand, many SNPs are highly multicollinear due to the LD structure and are also noise-prone due to the small effect size of SNPs for polygenic traits and the large number of SNPs tested in GWAS. It is therefore clear that shrinkage methods have great potential to increase the prediction power of models based on GWAS results.

Elastic net regularisation[55] uses a linear combination of the constraint of LASSO and ridge to overcome the shortcomings of LASSO, especially in the situation where the number of covariates is greater than the sample size. The basic form of elastic net is:

$\min_{\beta \in \mathbb{R}^P} \{\frac{1}{N} \|y - X\beta\|_2^2\} \ subject \ \ to \ \frac{\lambda_1}{\lambda_1 + \lambda_2} \|\beta\|_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} \|\beta\|_2^2 \leq t$, in which λ_1 and λ_2 is the tuning parameter decided by cross validation.

Although there has been much research on shrinkage methods, applying these methods to the current GWAS can be problematic. Due to the large scale of current GWAS, directly applying shrinkage methods to raw genotype data can be extremely computational expensive. For example, the complexity of an ordinary LASSO regression for a sample containing p SNPs in *n* individuals (usually $p \gg n$), the complexity is *O(np min{n, p})*[56]. Besides, LASSO and ridge regression were first designed for multiple linear regression, while in GWAS, the SNPs are typically tested separately.

Therefore, shrinkage methods that are tailored to GWAS data and less computational expensive are needed.

In Chapter 2 and Chapter 3, I will introduce new shrinkage methods based on estimation of effect sizes under the null hypothesis.

1.9.Gene set analysis

1.9.1. The basic concept

Polygenic risk scores have many applications (see section 1.6.3), but most of them generate genome-wide conclusions. More specific biological mechanisms are yet to be investigated by PRS analyses. Gene set analysis (GSA) have been widely used to interpret the underlying biological mechanism with ‘omics data. In omics data, the signals from a single gene or gene product may be weak or it may be difficult to interpret the underlying mechanism. However, it is easier to interpret the enrichment of signals in a gene set whose biological meaning is clear and well-studied.

The term “gene set” is a broad concept. A gene set can be a group of genes that take part in a certain biological process (e.g. the genes that take part in a metabolism pathway), or a group of genes that have the same functional annotation (e.g. Gene Ontology[57]), or even a group of genes that meet the same customized criteria in a particular research project (e.g. a group of genes that are expressed in a certain tissue). By defining a gene set from previous studies and analysing new data with gene set knowledge, we can combine our previous knowledge with the new findings.

1.9.2. Widely-used GSA database

The information used to define gene sets usually comes from gene set or pathway data base such as Kyoto Encyclopedia of Genes and Genome (KEGG)[58], Gene Ontology (GO)[57] and MSigDB[59]. These databases collect information and knowledge of genes and molecules involved in biological pathways. For example, Figure 4 is an example of a gene set or pathway diagram provided by KEGG (https://www.genome.jp/kegg/). KEGG provides the information relating to which genes comprise the gene set/pathway and the interactions between the genes and between genes and other biomolecules.

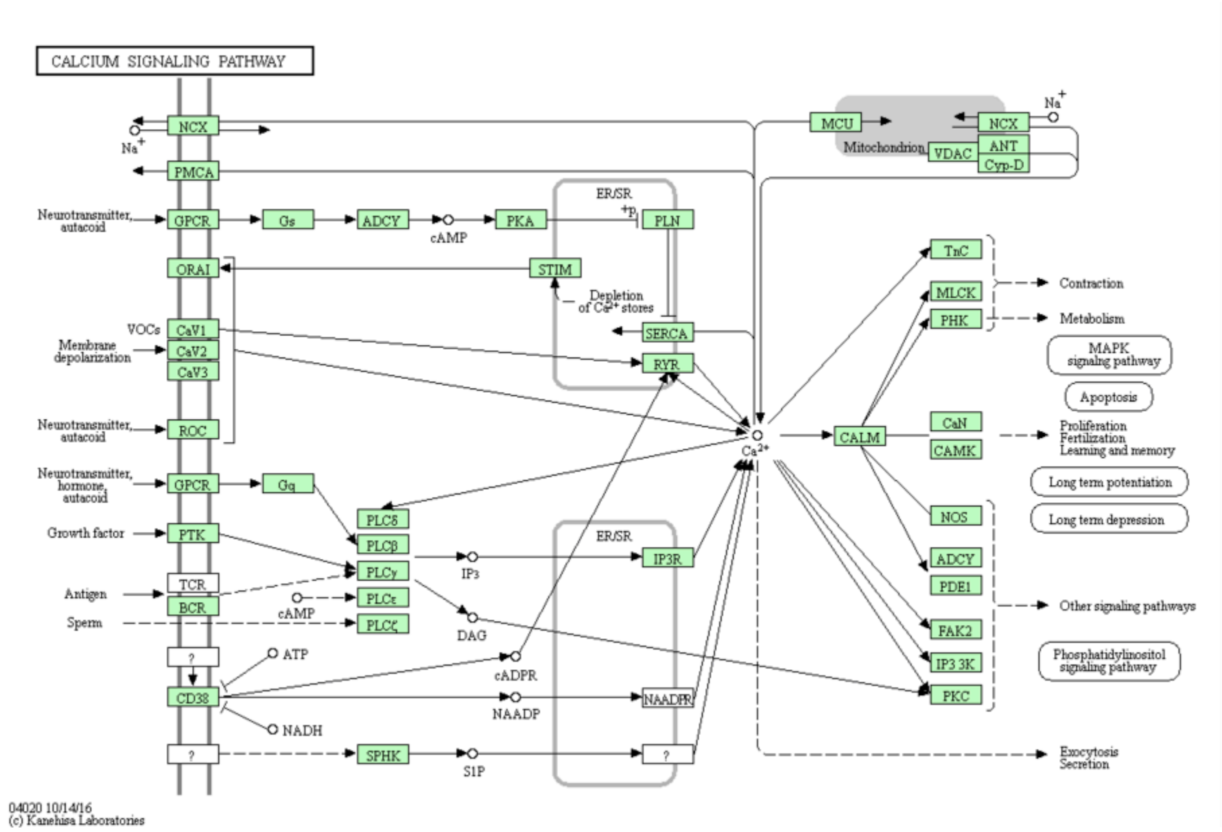


Figure 4 The diagram of calcium signalling pathway provided in KEGG database. This is an example of interacting genes (shown in green boxes) forming a pathway that performs a biological function. Information about involved metabolites and cellular structures is also included in the database.

1.9.3. Classification of GSA

Classification based on raw data

Gene set analysis can be performed on different types of data. Long before the availability of large-scale GWAS data like UK biobank, which are powerful enough to detect the small signals of polygenic traits, gene set analyses were mostly based on high-throughput measurement of RNAs and proteins, i.e. transcriptome and proteome data. In these datasets, the measures of gene products are the raw output data. Then the significance of each gene is directly calculated from the gene products. Multiple statistics methods, such as GSEA[59] were developed to calculate the gene set significance from the gene significance (e.g. the expression difference between testing sample and control sample).

Gene set analysis was introduced into GWAS data in order to 1) enrich the weak SNP association signals of GWAS to study whether there is a genetic association when only underpowered samples are available [60], and 2) study the aetiology and genetic basis of the traits. However, GWAS results are the association of single SNPs, instead of the expression or the function of the genes, such as the transcripts. Following the same idea of analysing transcriptome and proteome data, the approach of converting gene *P*-values into gene set *P*-values is still used in analysing genome data. However, the gene *P*-value has to be converted from the *P*-value of SNPs or other statistics that can summarize the genotype-phenotype association. Then the gene set association is calculated from gene *P*-value with GSEA.

As the sample size have increased, genome data such as GWAS or next generation sequencing have become powerful enough to perform gene set analyses. Yet, more recent methods may still take the same "two-tier" structure as summarized in a review on GSA[61]. The authors summarize this structure of GSA as "two-tier" as follows: tier one is to derive gene associations from SNPs within genes, and tier two is to obtain gene set associations from the gene associations.

In this thesis, the main focus is on GWAS data, rather than transcriptome or proteome data, as input.

Classification based on hypothesis

Depending on the null hypothesis, GSA can be divided into two groups: self-contained and competitive. Self-contained methods assume that the gene set (or the genes in that gene set) are not associated with the phenotype at all. Competitive methods assume that the genes inside the gene set are not more associated with the phenotype than the genes outside the gene set, or to put it in another way, that the gene set under study is not more associated with the phenotype than a random gene sets of the similar properties, such as gene set size.

Classification based on the statistical method implemented

In addition, de Leeuw *et al*[62] classified the GSA methods based on how they calculated the signal in the gene or gene set as 'mean-based', 'count-based' or 'rank-based'. Mean-based methods calculate the average significance of the variants in the gene set; Count-based methods calculate the number of significant variants in the gene set; and rank-based methods calculate the ranking of the all variants and test whether the variants in the gene sets are enriched in the top of the ranking list. Mean-based and count-based methods can be either self-contained or competitive.

Table 1 GSA classification based on the statistical method implemented by the methods adapted from the work of de Leeuw et al[62]

Method	**Type**	**Description**
Mean-based		
Fisher's method	Self-contained	Tests mean of –log or transformed *P*-values in the set against the null mean
Fisher's method	Competitive	Tests mean of –log or transformed *P*-values in the set against mean outside of the set
Single sample Z-test	Self-contained	Tests mean of probit transformed *P*-values in the set against the null mean
Two-sample t-test	Competitive	Tests mean of probit transformed *P*-values in the set against mean outside of the set
Linear regression	Competitive	Tests whether being in the set or not is a predictor of having higher probit transformed *P*-values
Count-based		

Binomial test	Self-contained	Tests whether proportion of *P*-values in the set below the threshold is greater than the null proportion
Hypergeometric test	Competitive	Tests whether proportion of *P*-values below the threshold in the set is greater than the proportion outside the set
Logistic regression	Competitive	Tests whether being in the set or not is predictor of having *P*-values below the threshold
Rank-based		
Two-sample Kolmogorov-Smirnov test	Competitive	Tests whether genes in the set are overrepresented at the top of the list of all genes ranked by *P*-value
Rank+mean-based		
GSEA	Self-contained or competitive	Modified Kolmogorov-Smirnov test, weight ranks by –log or transformed *P*-values

1.9.4. Possible problems of GSA

GSA can be highly challenging because the association we directly observe can be due to other aspects that are irrelevant to the biological function of the gene set. Here are some common problems for GWAS-based GSA:

1) The signal may also come from the overlapping genes or SNPs that also belong to other causal gene sets. The signal observed is not caused by the function of the observed gene set but other gene sets.
2) The annotation of gene set is wrong or outdated. It will cause the failure of detecting the signals from the gene sets[63].
3) Signal of gene set driven by only one gene or a small number of genetic variants, such that the gene set itself is not important for the outcome
4) The GSA test is statistically biased. For example, self-contained method is biased in favour of big gene sets because the larger gene sets may get more associated genes/SNPs only by chance because of their large size, or LD structure, rather than their function.

While the first three problems listed above are important and should be considered carefully in future research on this topic, this thesis only focuses on the 4th problem.

1.9.5. Widely used GSA Methods

In this section, several of the most popular gene set analysis methods are reviewed. According to a previous comparison study[61], MAGMA is the current leading method. In chapter 4, the methods developed in my PhD project are compared with MAGMA.

GSEA

GSEA[59] was designed to analyse expression data. It uses the expression data and ranks the genes' expression change in the test samples compared with the control samples. It estimates whether the genes in a gene set are randomly distributed across the ranking list or enriched in the top or the bottom of the list by calculating an Enrichment Score (ES) according to the ranking list. The statistical significance of the ES is tested by permuting the phenotype data multiple times and obtaining an empirical *P*-value of the ES.

FORGE

FORGE[64] combines the SNP *P*-values with a correction for the LD structure and the correlation between the SNPs to obtain a combined *P*-value for the gene set or a gene. It can either directly combine the *P*-values of all the SNPs mapped to a gene set or calculate the *P*-value in a two-tier structure, that is, where the *P*-values of genes are first calculated and then combined into a gene set *P*-value with correction for LD structure and the correlation between genes.

MAGMA

MAGMA[65] has a two-tier structure. The first tier is to calculate the gene association from genotype data. A model predicts the phenotype with the principle components of SNPs in a gene j, controlling for covariates such as sex, age, genome-wide PCA of this individual i, etc.

$$Outcome_i \sim PCA_{ij} + Cov_i$$

Then the P-value of this model is converted to a one-tailed Z_j score, which is presented as the gene association in the subsequent analysis.

The second tier calculates the gene set analysis by compare the gene association of genes inside and outside the gene set in a regression model. To test the association of gene set s, a model is built to predict Z_j with an indicator of the gene set membership, S_{sj}, controlling for covariates such as the gene size, minor allele counts and gene density (the ratio of clumped PCs to the total number of SNPs in the gene). $S_{sj} = 1$ if the gene j is in the gene set s; $S_{sj} = 0$ if the gene j is not in the gene set. ε is the error term, which takes the correlation between genes into consideration:

$$Z_j \sim \beta_{sj} \cdot S_{sj} + Cov_j + \varepsilon$$

One of the advantages of the "two-tier" methods is that it is easy to incorporate the results from different types of research. However, not all the GSA methods necessarily follow the "two-tier" structure. For example, one option of FORGE[66] is to directly calculate the combined P-values of all the SNPs in the gene sets.

To summarise, when developing a GSA method, one needs to answer two questions: first, how to summarize the genetic signals from the gene set and generate a statistic for this gene set; second, how to evaluate the statistical significance based on this statistic. In Chapter 4, a GSA method based on PRS is introduced and tested. We test several options to estimate the statistical significance in both self-contained and competitive tests.

Stratified LD score regression

Stratified LD score regression[67] expand the application of typical LD score regression[30] (see 1.5.1) to analyse partitioned heritability of different functional elements accounting for LD. It makes the same assumption as LDSC that χ^2 association statistic for a given SNP includes the effects of all the SNPs in LD with this SNP and uses the same mathematical model except that only the SNPs belonging to a certain functional category are included in the calculation of LD score.

In LDSC, the mathematical model is:

$$E(\chi^2|l_j) = N\frac{h^2}{M}l_j + Na + 1$$

Where N is sample size, $\frac{h^2}{M}$ is the heritability explained per SNP, l_j is the LD score of SNP j; a is a term that measures the contribution of confounding biases.

In stratified LD score regression, the mathematical model is:

$$E(\chi^2) = N\sum_{C} \tau_C \ell(j, C) + Na + 1$$

where C indicates category, $\ell(j, C)$ is the LD score of SNP j that calculated within category C, τ_C is heritability explained per SNPs of category C.

Thus, the significance of enrichment of a category is to test whether the heritability explained per SNP of a category is larger than the baseline level or that of SNPs outside the category. The statistics $\frac{h^2(C)}{|C|} - \frac{h^2 - h^2(C)}{M - |C|}$ follows normal distribution with the expectation of 0 and standard error that can be estimated using a block jackknife over SNPs with 200 equally sized blocks of adjacent SNPs. Thus, z score and *P*-value can be estimated.

1.10. Summary

GWAS is a useful tool to identify genetic variants associated with complex disease. PRS is an individual-level proxy of the polygenic burden of risk of a disease or propensity to a trait, based on GWAS results. PRS have many successful applications within biomedicine. In this PhD project, I aim to increase the power of PRS by improving the accuracy of GWAS effect size estimates and broaden the application of PRS by developing a group of PRS-based GSA methods.

In Chapter 2, I developed a shrinkage method called "Permutation Shrinkage" to improve GWAS effect size estimates. Permutation Shrinkage first estimates the null distribution of genetic effect sizes by running GWAS on permuted null phenotype data. Then the estimated null effects are subtracted from the observed data to generate corrected estimates. This method was tested with quantitative traits in the UK Biobank and it can increase the PRS R^2 by approximately 35%.

In Chapter 3, I developed a similar shrinkage method called "Order Statistics Shrinkage" that uses a similar framework to Permutation Shrinkage: estimating the null distribution and subtracting the estimated null effects from the observed effects. Here, the null distribution is estimated from order statistics instead of permuting the individual-level data. Order Statistics Shrinkage can be applied to summary statistics data. We tested the method with summary statistics data from the UK Biobank and Order Statistics Shrinkage had similar performance as Permutation Shrinkage.

In Chapter 4, I developed PRSet, a group of PRS-based GSA methods. I used gene set PRS to represent the genetic burden in the gene set. Self-contained and competitive analyses were implemented and tested with UK Biobank data and were compared with MAGMA, the current leading GSA method.

Chapter 2. A novel shrinkage method – Permutation Shrinkage – and its application to GWAS data to increase predictive power of PRS

2.1.Introduction

GWAS was originally designed to discover the genetic variants significantly associated with the phenotype under study but in recent years has been used for more applications as its power has increased[26]. In particular, GWAS results are now commonly used to calculate polygenic risk scores (PRS) in samples with individual-level genotype-phenotype data, which have a wide range of applications[29]. However, GWAS can produce inflated SNP effect size estimates[68] and -$\log_{10}$(P-values) and the inflation makes the GWAS results less reproducible[69]. Thus, prediction based on the GWAS data, such as from polygenic risk scores, could have reduced power and accuracy. This can hinder the translation of GWAS results into further research and application.

Inflation can have a particularly large impact on effect size estimates in the GWAS setting due to a combination of: 1) the millions of tests performed across genome-wide SNPs in GWAS (multiple comparisons problem), and 2) the very small effect sizes of common genetic variants (low signal-noise ratio) (see section 2.2). Therefore, the top results in GWAS are likely to suffer from "Winner's Curse", whereby the top results among a large number of tests on predictors of similar effect size are ranked top partly because of their severe inflation[70]. "Winner's Curse" is more likely to happen if the true effect size of the top causal variables is not distinguishingly higher than the rest, which can be true for common variants in relation to many polygenic traits[71].

GWAS inflation can lead to producing an overfit prediction model, whereby the top predictors have inflated estimates of the regression coefficients and null predictors are included due to their chance association with the outcome in the study data. This may affect the predictive power of polygenic risk scores because they are the sum of risk alleles weighted by the corresponding effect size estimates. In the standard PRS approach, known as the Clumping and Thresholding (C+T) method[29], PRS are calculated across a range of P-value threshold and the PRS most associated with the outcome in the target data is selected. Thus, while PRS can suffer from inflation due to this *P*-value threshold optimization, it may also produce misleadingly PRS R^2 due to the inflated GWAS effect size estimates that it is based on. While the GWAS results are likely to suffer from both inflation and deflation due to the effect of 'noise', the C+T method ensures that the top SNPs, and thus most inflated, from the GWAS are included in the PRS, while the SNPs with the weakest associations may be excluded from the optimised PRS.

Given the generality of the inflation problem in association testing and prediction modelling, numerous statistical techniques have been developed to deal with the inflation[37,53,55,72,73]. The most common approach to account for the inflation of $-\log_{10}$ (*P*-value) caused by multiple comparisons problems, is to apply a stricter *P*-value threshold, for example, based on a Bonferroni correction[72, 73] of the effective number of independent tests or a False Discovery Rate (FDR)[73]. For SNP association results in GWAS, SNPs are considered to be genome-wide significant if $P < 5 \times 10^{-8}$, which was a significance threshold derived based on a consensus of several approaches[19].

To correct the inflation of effect size estimates, most existing shrinkage methods optimize the value of the regression coefficients by adding constraint or penalty terms and/or reduce the number of predictors used in the models. Thus, some shrinkage methods are referred to as Penalised Regression. Some of the most commonly applied shrinkage methods are the Least Absolute Smooth Shrinkage Operator (LASSO)[37], Ridge Regressioan[53] and Elastic Net[55] . Software such as lassosum[36] and LDpred[35] have been developed to integrate some forms of these methods in application to GWAS data.

The aim of this chapter is to produce a simple, interpretable and computationally efficient alternative shrinkage method, "Permutation Shrinkage". The principle of this method is to estimate the error quantity of the effect size estimates and subtract the inflated part away from the raw estimates. The method was tested with UK Biobank data and with simulation studies. Improved power of PRS-based predictions according to improved GWAS effect size estimates would be evident from increased out-of-sample PRS R^2.

2.2.Winner’s (and Loser’s) Curse

When multiple tests are performed in order to investigate a broad hypothesis, such as performing millions of association tests between genetic variants and a phenotype across the genome based on the hypothesis that some fraction of genetic variants affect the phenotype, the effect of stochastic variation plays a key role. If the true effects are relatively small and largely homogenous, then the top results of millions are likely to be largely influenced by chance variation, and thus have inflated effect size estimates – a phenomenon known as “Winner’s Curse”. Less commonly discussed is the opposite effect, which could be thought of as “Loser’s Curse”, in which the true effect size is underestimated due to having a large contribution of chance association in the opposite direction to that of the effect. The contribution of stochastic variation, and thus Winner’s and Loser’s Curse, is greater the smaller the sample size and the larger the number of tests performed. These effects are illustrated in Figure 5.

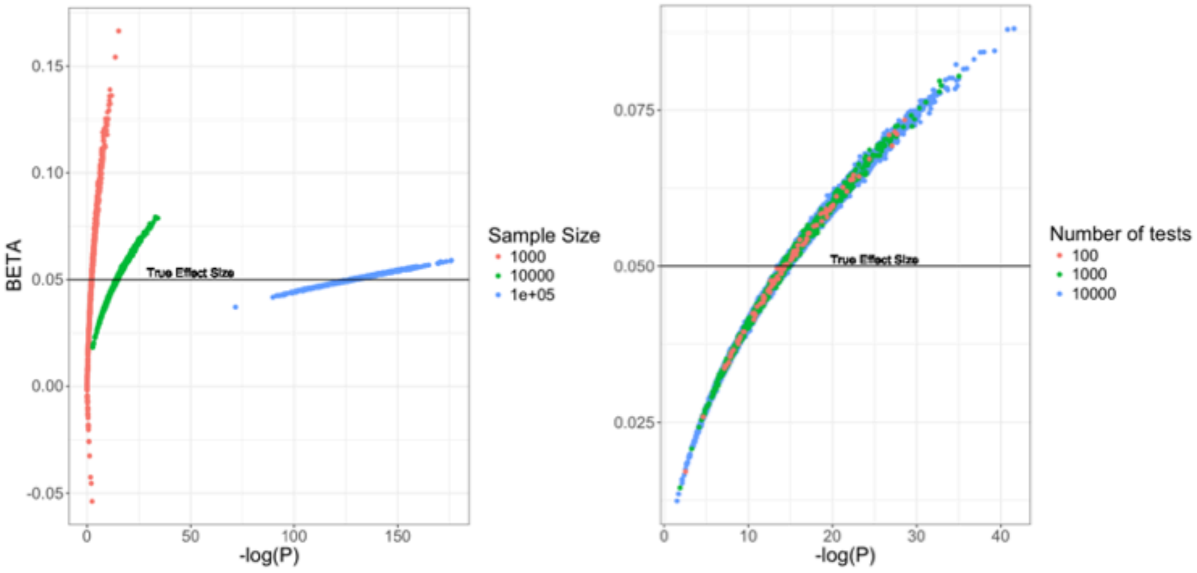


*Figure 5 Winner's and Loser's Curse in simulated multiple test scenario. Each data point in the plot shows the statistics of one effect size estimated from the following simulation: the random variable y was simulated so that y=beta*x+e, where beta is true effect size; e is the error quantity; and x and e follow normal distribution. The x-axis shows the -log$_{10}$(P-value) of the effect size estimates and the y-axis shows the corresponding estimated values of effect size. The black line shows the true effect size. The top results have the biggest effect size estimates and -log$_{10}$(P-value) because they are most inflated rather than their true effect size is higher than the rest results. The left plot shows that the smaller sample size, the more inflated the effect size estimates; the right plot shows that the most tests performed the more inflated the top results are*

2.3.Methods and Data

2.3.1. Overview of the prototype method

The main principle underlying Permutation Shrinkage is that an estimated regression coefficient, $\widehat{\beta},$ is the combination of the true effect size, $\beta,$ and an error quantity, or 'noise', *e*.

$$\hat{\beta} = \beta + e$$

The formula can be written as:

$$\beta = \hat{\beta} - e$$

Assuming that the error is independent of the true effect size and follows a distribution that relates to the sample size and the number of tests performed overall (as shown in Figure 5), then the distribution under the null will correspond to the distribution of the error.

Since the distribution of the error term is unknown due to the complicated nature of GWAS data, in particular the complex correlation structure among nearby genetic variants across the genome, here we take a permutation approach to estimate the distribution of the error under the null. The estimated error quantity $\hat{e}$ can be generated by taking the average of multiple permuted null GWAS effect size distribution. Thus, the corrected effect size estimate $\hat{\beta}'$will be:

$$\hat{\beta}' = \hat{\beta} - \hat{e}$$

Please be note that this method takes two assumptions: First, the true effect size of all or most the variables is the same; The error quantity is independent with the true effect sizes and follows a certain distribution. The following section will explain how this method is implemented to GWAS data.

2.3.2. Adjusting the prototype method for GWAS data

This method assumed that all or most of the variables have the same effect size expectation and the deviation of the observed estimates from the expectation is caused by the error quantity. Another assumption is that the difference between the estimated error quantity distribution and the distribution of the deviation of the observed values from the expectation is caused by the true signals instead of systematic bias.

The prototype method needs to be adjusted accordingly when the data to be corrected have certain features that may violate the assumptions mentioned above. The distribution of GWAS effect sizes have two features: first, the majority of the alleles are estimated to be null[71] so it is instinctive to assume that the expectation of SNP effect size is zero; second, the distribution of causal SNPs may be skewed due to selection. If the risk alleles are randomly assigned, it is instinctive to assume that the effect sizes of causal SNPs follow normal distribution. However, because of selection, the sign of the effect size may be skewed while this will not change the assumption of the zero expectation of the effect size. If a symmetric distribution of expected error quantity is used to correct a dissymmetric distribution of the observed deviation, the corrected result will be biased.

The prototype method is adjusted as the following to correct the GWAS data: the distribution of the absolute deviation of the observed estimate to the expectation, i.e. 0 is estimated by assuming the error quantity following a symmetric distribution and take the absolute value of the distribution. Only the absolute value is corrected, and the sign of the estimate remains the same after shrinkage.

2.3.3. Implementation of Permutation Shrinkage on GWAS data

First, GWAS is performed on the original observed phenotype and genotype data. Then the phenotype data is permuted to get a null trait and the GWAS is performed on the permuted null data. The permutation process is repeated for 100 times.

Minor allele frequency has a great impact on the standard error of effect size and possibly the effect size[74 75]. Since the method assumes that the true effect size of the most variables is the same and the error quantity follows the same distribution, the SNPs are divided into 1% MAF bins so that the SNPs in each MAF bin are more homogenous than SNPs across the whole genome. The SNPs are corrected within each MAF bin in the following step.

Since the sign of effect size is decided by the choice of reference allele, the correction is applied to the absolute value of the effect size. For each permutation, the absolute effect sizes are ranked ascendingly. The ranked null absolute effect size from each single permutation is an estimate of the null effect size distribution but it would contain noises (i.e. inflation and deflation). A smoother estimate of the null absolute effect size distribution is generated so that the kth value of the ranked null distribution $null|\beta_k|$ is the average of all the kth value of the ranked null distribution estimated from each single permutation.

In each MAF bins, the absolute observed effect sizes are ranked to get the observed distribution. For the kth SNP in the distribution:

$$Corrected\ |\beta_k| = \begin{cases} Observed|\beta_k| - null|\beta_k|, & if\ observed|\beta_k| - null|\beta_k| > 0 \\ 0 & , if\ observed|\beta_k| - null|\beta_k| \leq 0 \end{cases}$$

The corrected absolute effect size estimates may not be monotonic with the original estimates. To overcome this problem, the absolute corrected estimates in each MAF bin are ranked and then the ranked estimates are re-assigned to the SNPs. For example, if the corrected absolute estimates for SNP1, SNP2, SNP3, SNP4 is 0.1, 0.2, 0.4, 0.3. After this step, the absolute estimate of SNP3 will be 0.3 and that of SNP4 will be 0.4.

The sign of corrected effect size remained the same as the original.

2.3.4. UK Biobank data for testing the method

Genotype data of 2nd release were quality-controlled with the following parameters: For SNPs: SNPs that had MAF < 0.01, sample missingness>0.01 or HWE p < 10-8 were excluded; For samples: individuals of SNP missingness > 0.02 were excluded, only Caucasians were included; One of each pair of relatives was removed, individuals in KING relatedness criterion < 0.088 (r ~= 0.25) and UKB Recommended exclusions and UKBileve exclusions were removed; Confirmed sex was as reported; After correction the remaining sample size was about 385k.

The sex, age, top 15 PCs were regressed out from the traits. The standard residues were used as the outcome in the following analysis.

2.3.5. Simulated data for testing the method

The simulation was based on QC'ed UK Biobank chromosome 1 genotype data. Phenotype data was simulated by an in-house software SGCP written by my co-operator Shing Wan Choi. In this simulation, y_j, the outcome of individual j is

$$y_j = sum(x_{ij}u_i) + e_j$$

in which the u_i is the effect size of the SNP_i, x_{ij} is the number of reference alleles in SNP_i of individual j.

e_j is the non-genetics effect which follows the distribution:

$$\mathrm{e_j} \sim N(0, va(sum(w_{ij}*u_i))(1 / h^2 - 1))$$

We calculated the chromosome 1 heritability of the traits listed in the 2.3.4. The majority of heritability of chromosome 1 range from 0.01 to 0.05. Based on the real data heritability, we simulated phenotype of heritability of 0.001, 0.005, 0.01, 0.015, 0.02, 0.03, 0.075, 0.1, 0.04, 0.05, 0.06 and ratio of causal SNPs 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2. Please note that the simulated data are only based on chromosome 1. Therefore, the heritability was in a different scale. Assuming the casual variants are evenly distributed along the genome, chromosome 1 will contribute about 8% of the genome-wide heritability.

Another simulation was performed by GCTA[76]. The GCTA method uses standardised genotype data w_{ij} instead of the raw data x_{ij}:

$$w_{ij} = (x_{ij} - 2p_i) / \mathrm{sqrt}[2p_i(1 - p_i)]$$

in which the p_i is the minor allele frequency of $\mathrm{SNP_i}$. Except this, the other parameters were the same.

Please note that the simulated data are only based on chromosome 1. Therefore, the heritability was in a different scale. Assuming the casual variants are evenly distributed along the genome, chromosome 1 will contribute approximately 8% of the genome-wide heritability.

2.3.6. Heritability estimation

We estimated the heritability of each trait with LD Score regression[30], using the default parameters and based on the base data. The sample size used to estimate the heritability of each trait was approximately 138k.

2.3.7. Construction and validation of polygenic risk score prediction model

The prediction power of the corrected effect size estimates was tested using polygenic risk score (PRS). PRS is the sum of risk alleles weighted by their effect size estimates. The PRS prediction model contains a base dataset and a target dataset. The base data is summary statistics of GWAS to provide the effect size estimates; the target is the individual level raw genotype from which the PRS is calculated and phenotype. In the target data, the phenotype is regressed on each individual's PRS. In this study, PRSice[77] were used for PRS calculation and regression. PRSice maximises the prediction power by optimising the *P*-value threshold for SNPs to be included in the model.

The UK Biobank data were divided into three equal subsets: discovery data, testing data and validation data. GWAS was performed on the discovery data and the effect size estimates were corrected with Permutation Shrinkage. The original and corrected effect size estimates were used as the base data for PRS model. The testing data were used as the target data. PRSice optimized the *P*-value threshold for building the PRS model and the *P*-value threshold might be overfit toward the testing data. A validation PRS model were built using the out-of-sample validation data as the target data and the same *P*-value threshold optimized from the testing data PRS model. The relative increase of the validation data PRS R^2 after correction is used to evaluate the increase of PRS prediction due to the correction by Permutation Shrinkage.

Table 2 UK Biobank traits used to test the methods. The heritability was estimated using the base data.

TRAITS	CATEGORY	UK BIOBANK FIELD CODE	ESTIMATED HERITABILITY
HEIGHT	Body size measures	f.50	0.6005±0.0049
WHOLE BODY FAT FREE MASS	Impedance measures	f.23101	0.3894±0.0043
BASAL METABOLIC RATE	Impedance measures	f.23105	0.3746±0.0056
WHOLE BODY IMPEDANCE	Impedance measures	f.23106	0.3551±0.0071
FORCED VITAL CAPACITY (FVC), BEST MEASURE	Spirometry	f.20151	0.3467±0.0032
BMI	Body size measures	f.23104	0.3228±0.0076
FORCED EXPIRATORY VOLUME IN 1-SECOND (FEV1), BEST MEASURE	Spirometry	f.20150	0.3119±0.0049
BODY FAT PERCENTAGE	Impedance measures	f.23099	0.3079±0.0066
SIT HEIGHT/STANDING HEIGHT RATIO	Body size measures	f.20015 / f.50	0.3045±0.0057
FLUID INTELLIGENCE SCORE	Cognitive Function	f.20016	0.3013±0.0193
WAIST/ HIP RATIO	Body size measures	f.48 / f.49	0.2193±0.0094
SYSTOLIC BLOOD PRESSURE, AUTOMATED READING	Blood pressure	f.4080	0.1916±0.0014
PULSE RATE, AUTOMATED READING	Blood pressure	f.102	0.1909±0.0045
DIASTOLIC BLOOD PRESSURE, AUTOMATED READING	Blood pressure	f.4079	0.1849±0.0024
AVERAGE HAND GRIP	Hand grip strength	average of f.46 and f.47	0.1835±0.0065
AGE AT FIRST LIVE BIRTH & AGE OF PRIMIPAROUS WOMEN AT BIRTH OF CHILD	Female-specific factors	f.2754 & f.3872	0.1681±0.0062
BIRTH WEIGHT	Early Life Factors	f.20022	0.1565±0.0027
AGE FIRST HAD SEXUAL INTERCOURSE	Sexual factors	f.2139	0.1564±0.0042
TIME SPENT WATCHING TELEVISION (TV)	Physical activity	f.1070	0.1496±0.0109
BIRTH WEIGHT OF FIRST CHILD	Female-specific factors	f.2744	0.1449±0.0142
NEUROTICISM SCORE	Mental health	f.20127	0.1443±0.0072
TIME SPEND OUTDOORS IN SUMMER	Sun exposure	f.1050	0.0979±0.0044
COMPARATIVE BODY SIZE AT AGE 10	Early Life Factors	f.1687	0.0972±0.0030
SLEEP DURATION	Diet	f.1160	0.0852±0.0050
MEAN TIME TO CORRECTLY IDENTIFY MATCHES	Cognitive Function	f.20023	0.0794±0.0029
WATER INTAKE	Diet	f.1528	0.0746±0.0035
FRESH FRUIT INTAKE	Diet	f.1309	0.0679±0.0053
TEA INTAKE	Diet	f.1488	0.0655±0.0053
TIME SPENT OUTDOORS IN WINTER	Sun exposure	f.1060	0.0643±0.0047
COMPARATIVE HEIGHT SIZE AT AGE 10	Early Life Factors	f.1697	0.0566±0.0036
BREAD INTAKE	Diet	f.1438	0.0502±0.0037
COOKED VEGETABLE INTAKE	Diet	f.1289	0.0345±0.0046
SIBLING NUMBERS	Family history	f.1873 + f.1883	0.0312±0.0041
SALAD / RAW VEGETABLE INTAKE	Diet	f.1299	0.0311±0.0064
REACTION TIME	Cognitive Function	f.404.0.0	0.0287±0.0076

TOWNSEND DEPRIVATION INDEX AT RECRUITMENT	Sociodemographics	f.189	0.0111±0.0130
AVERAGE LOGMAR	Visual acuity	average of f.5201 and f.5208	-0.0109±0.0097

2.4.Results

2.4.1. Performance of Permutation Shrinkage under simple simulation scenarios

The method was tested with 10,000 individuals simulated by R. Each of them has 100 independent SNPs that act additively and produce a quantitative trait of heritability = 0.1.

Four different scenarios were tested:

Scenario 1: all the SNPs have the same minor allele frequency (MAF) of 0.25 and the same effect size;

Scenario 2: 60 SNPs have the MAF of 0.25; 40 SNPs have MAF of 0.05; all the SNPs have the same effect size;

Scenario 3: all the SNPs have the same MAF of 0.25; 20 SNPs have effect size of 1.0, 20 SNPs effect size of 0.5, 60 SNPs effect size of 0;

Scenario 4: all the SNPs have the same MAF of 0.25; 10 SNPs have effect size of 1.0, 10 SNPs effect size of 0.5, 60 SNPs effect size of 0

The following figures show how much the Permutation Shrinkage may improve the effect size estimates under different genetic architecture. The corrected estimates are shrunk towards the true value.

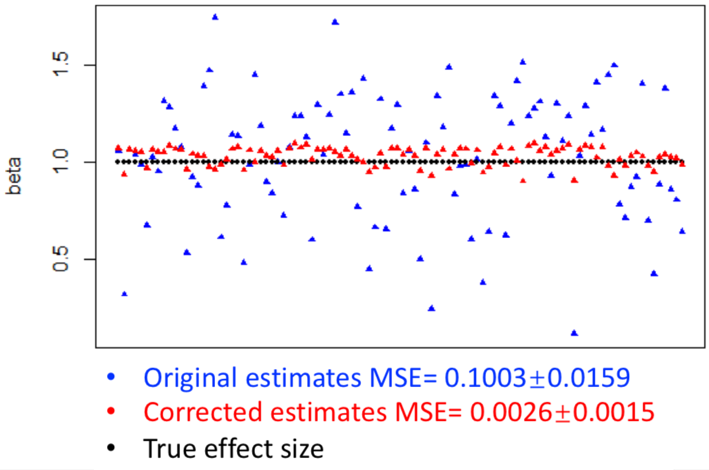


Figure 6 Corrected and original Effect size estimates in Scenario 1.

Note: from *Figure 6* to *Figure 9*, the corrected and original effect size estimates and the true values are shown in different colours. The mean squared error (MSE) was calculated for both corrected and original effect size estimates. The simulation was repeated for 10 times to calculate the mean and standard error of MSE. The figures only show the result of one round of simulation; the mean and standard error of MSE under the figures are based on 10 repetitions. All scenarios are plotted in the same way.

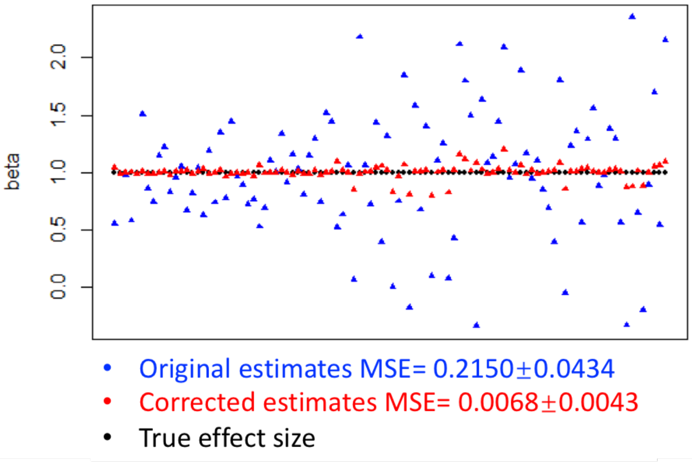


Figure 7 Corrected and original Effect size estimates in Scenario 2

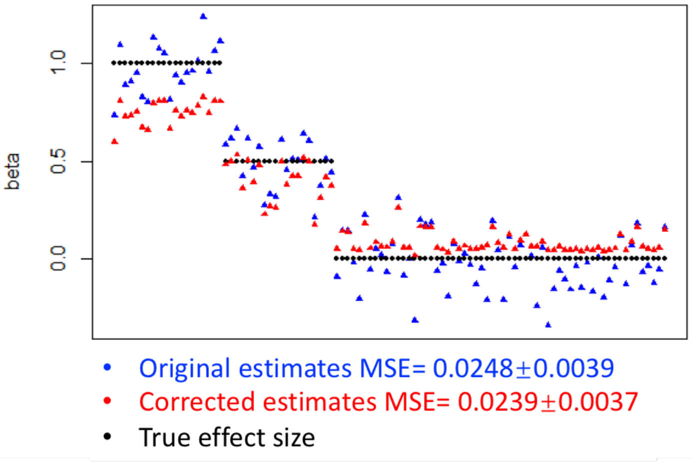


Figure 8 Corrected and original Effect size estimates in Scenario 3

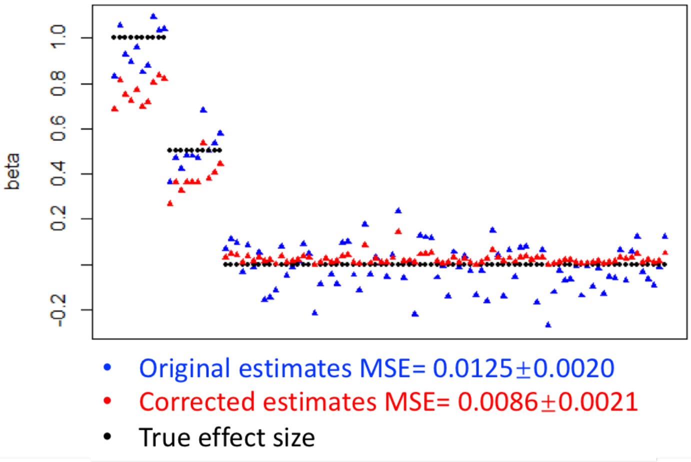


Figure 9 Corrected and original Effect size estimates in Scenario 4

The method worked best when the effect sizes of all the SNPs were the same. If the effect sizes were different, the method worked better if the proportion of the SNPs with effect size different from the majority was smaller, which accorded with the assumption of the method.

The feature of polygenic traits also makes it possible to correct effect size estimates with this method. The previous GWAS results that although for most polygenic traits, many SNPs may contribute to the traits, but the effect sizes of the causal SNPs are very close to zero; only a very small proportion of the casual SNPs have large effect sizes[71].

2.4.2. Performance of Permutation Shrinkage correcting SNP effect size in each MAF bin

For heritable traits, we took pulse rate, automated reading data (heritability estimated = 0.1909±0.0045) as an example to compare the corrected beta and original effect sizes.

Figure 10 shows that after correction, the effect sizes of SNPs with *P*-value $>10^{-10}$ were shrunk towards zero. The effect sizes of SNPs with *P*-value $< 10^{-10}$ were less shrunk. As an overall result, less significant SNPs had smaller effect size than the more significant SNPs after correction while in the original data effect size of both non-significant and significant SNPs could be large.

Figure 11 shows the comparison of corrected effect sizes (y-axis) and the original ones (x-axis) in different MAF bins. The relation of original and corrected effect size was not linear. The effect sizes of most SNPs were shrunk towards zero and the shrinkage of the majority part was monotonic. The effect size of the SNPs with top absolute estimates were shrunk with much less extend. Non-monotonic shrinkage mostly happened in the two ends of the distribution. However, this non-monotonic problem could be negligible if the sample size is large (data not shown).

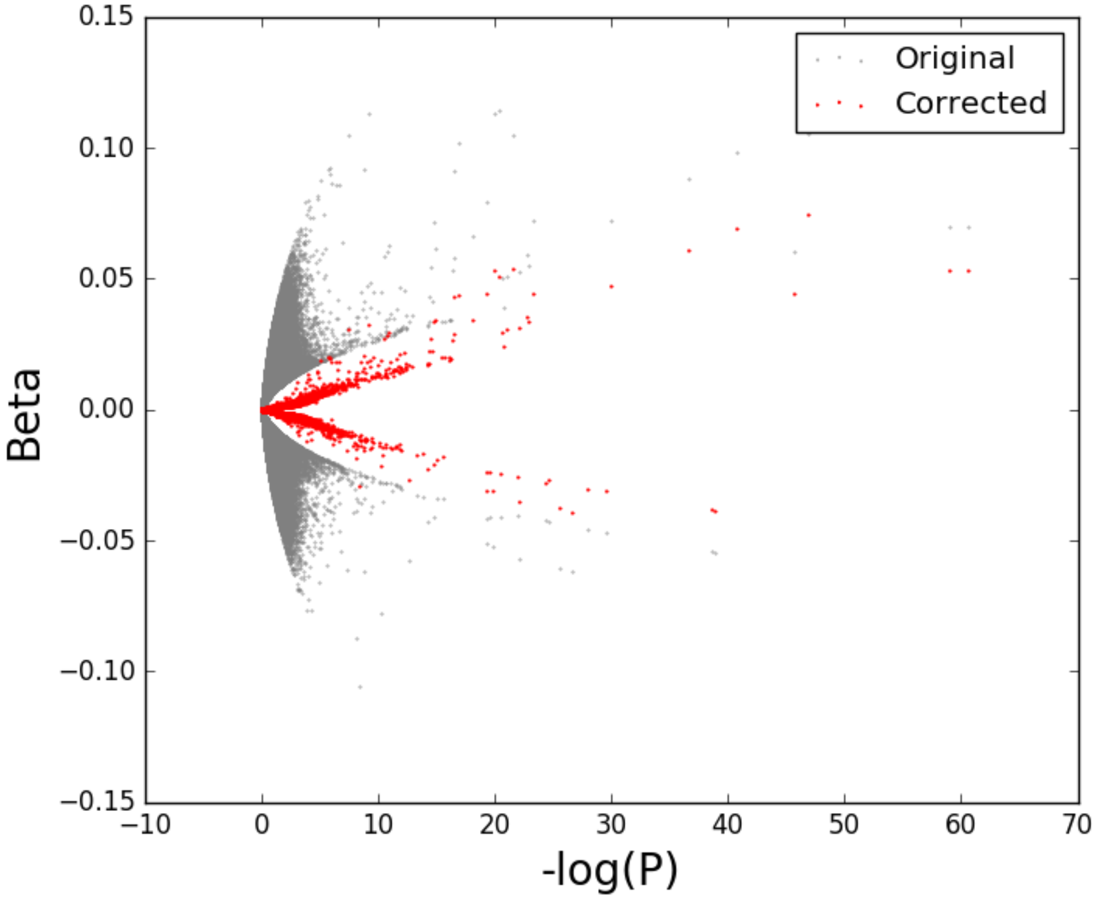
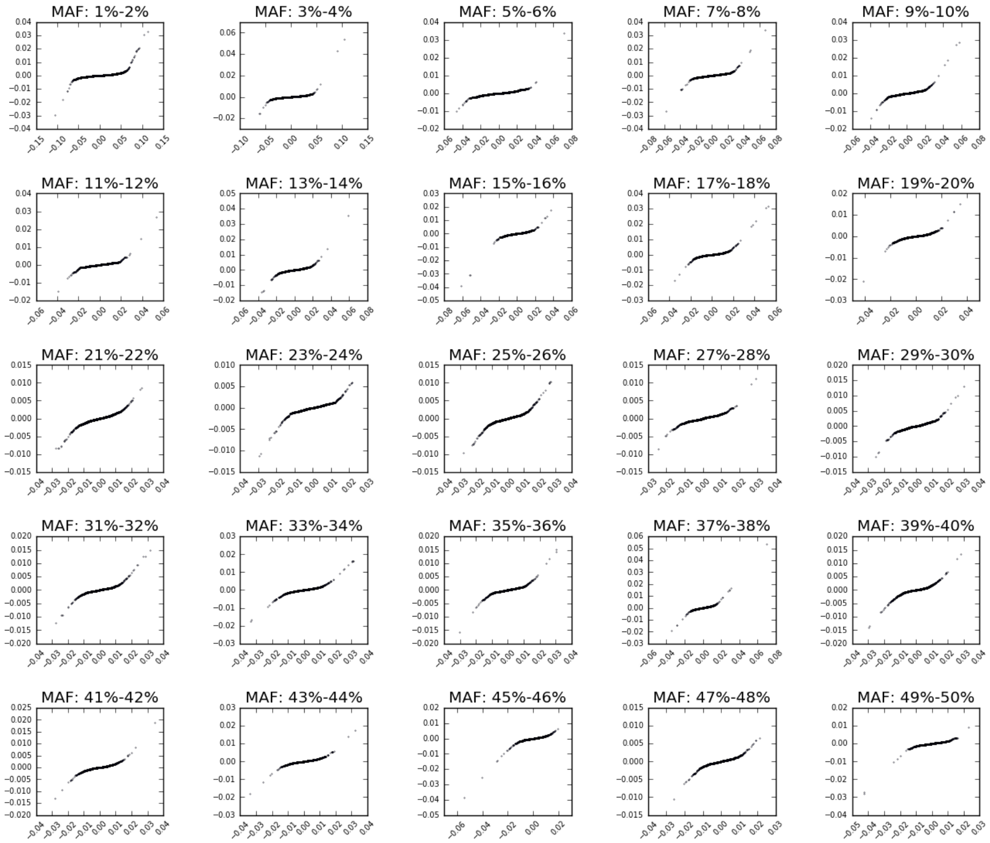


Figure 10 Comparison of corrected and original SNP effect size estimates of UK Biobank pulse rate GWAS data. Each data point represents the statistics of one SNP in the GWAS. The grey colour shows the original results and the red colour shows the corrected results. The x-axis shows the original -log$_{10}$(P-value) and the y-axis shows corresponding effect size estimates value. The plot shows that more significant SNPs were less shrunk and less significant SNPs were shrunk much closer towards zero, especially for those whose P-value doesn't pass the genome-wide significance threshold 5×10^{-8}


MAF: 1%-2%
MAF: 3%-4%
MAF: 5%-6%
MAF: 7%-8%
MAF: 9%-10%
MAF: 11%-12%
MAF: 13%-14%
MAF: 15%-16%
MAF: 17%-18%
MAF: 19%-20%
MAF: 21%-22%
MAF: 23%-24%
MAF: 25%-26%
MAF: 27%-28%
MAF: 29%-30%
MAF: 31%-32%
MAF: 33%-34%
MAF: 35%-36%
MAF: 37%-38%
MAF: 39%-40%
MAF: 41%-42%
MAF: 43%-44%
MAF: 45%-46%
MAF: 47%-48%
MAF: 49%-50%

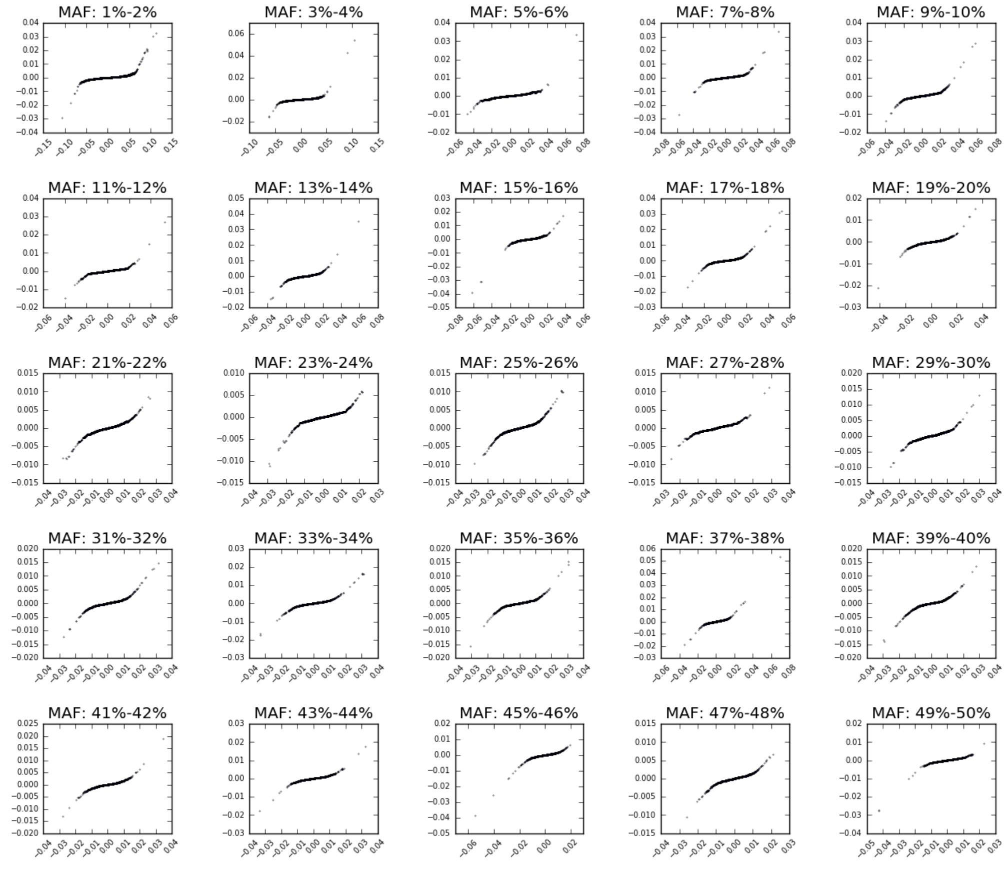

MAF: 1%-2%
MAF: 3%-4%
MAF: 5%-6%
MAF: 7%-8%
MAF: 9%-10%
MAF: 11%-12%
MAF: 13%-14%
MAF: 15%-16%
MAF: 17%-18%
MAF: 19%-20%
MAF: 21%-22%
MAF: 23%-24%
MAF: 25%-26%
MAF: 27%-28%
MAF: 29%-30%
MAF: 31%-32%
MAF: 33%-34%
MAF: 35%-36%
MAF: 37%-38%
MAF: 39%-40%
MAF: 41%-42%
MAF: 43%-44%
MAF: 45%-46%
MAF: 47%-48%
MAF: 49%-50%

Figure 11 Corrected and original effect size estimates in MAF bins. Every other 1% MAF bin is shown here. X-axis shows the original effect size estimates; y-axis the corrected ones. Up: the corrected effect sizes were further adjusted for non-monotonic problem; down: the corrected effect sizes were not further adjusted. The difference between the 2 plots is negligible.

For less heritable traits (genome-wide h^2 estimates < 0.05), most effect size estimates were shrunk to zero. Figure 12 is an example of low heritable traits (reaction time, heritability estimated = 0.0287±0.0076). When *P*-value > 10^{-8}, the effect size estimates were shrunk towards zero. In comparison, for more heritable traits SNPs of *P*-value > 10^{-8} were also shrunk towards zero, but the corrected effect size of more significant SNPs still remained non-zero as shown in Figure 11. Therefore, when the heritability is very low, the original PRS R^2 is approximately zero due to the small power and after correction, PRS R^2 remains close to zero because PRS was shrunk to zero. Therefore, we only showed the results of traits with estimated genome-wide heritability higher than 0.05 in the following analysis.

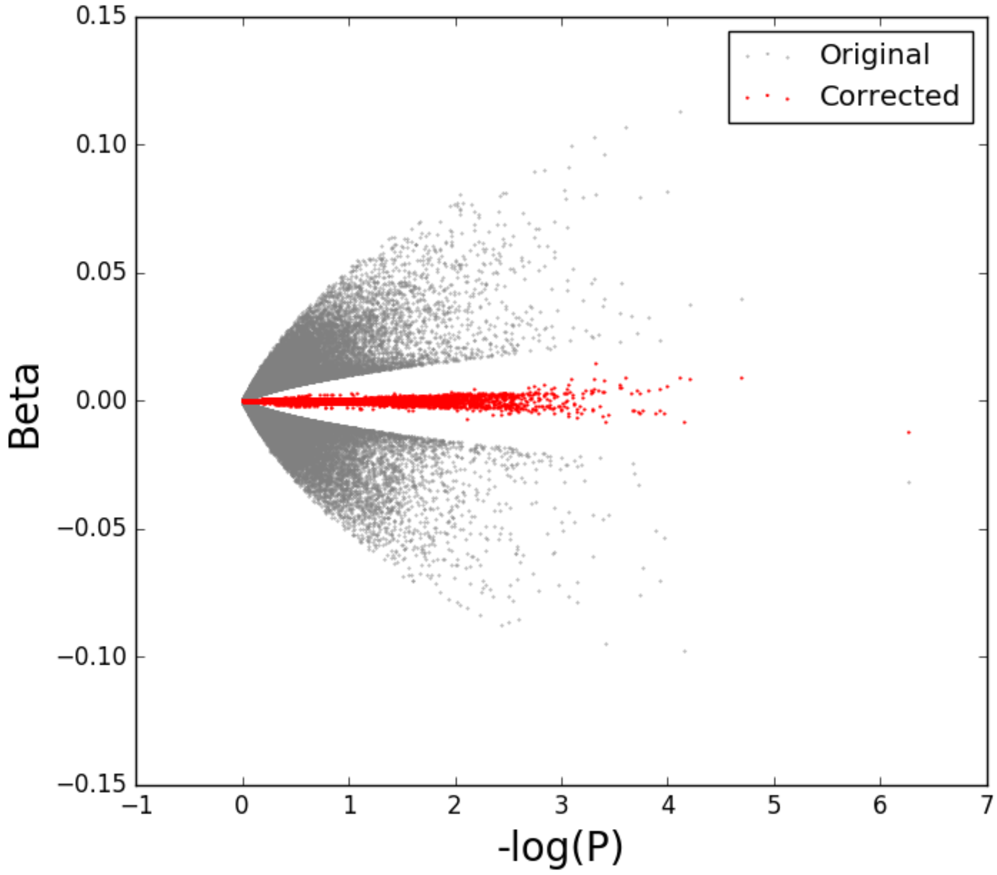


Figure 12 Corrected and original SNP effect size estimates compared against P-value of a low-heritability trait, UK Biobank reaction time. Each data point represents the statistics of one SNP. The grey colour shows the original results and the red colour shows the corrected results. The x-axis shows the original -log$_{10}$*(P-value) and the y-axis shows the corresponding estimated values of effect size. The plot shows that when all the SNPs are not genome-wide significant (all the P-values are larger than the typical threshold* 5×10^{-8}*) the corrected values were shrunk towards zero.*

2.4.3. Increased effect size correlation between testing and reference GWAS caused by Permutation Shrinkage

A wide range of traits of different heritabilities and from different categories were used to test the method. In order to test whether the correction makes the effect size estimates closer to the true value, the UK biobank data were divided by the ratio 1:2 and GWAS was performed on the 2 parts separately. The GWAS results of the larger part were the reference since the it should be closer to the true values than the smaller testing dataset. The GWAS results of the smaller testing dataset were corrected by Permutation Shrinkage. Correlation coefficients of the two parts were compared before and after the correction. Permutation Shrinkage significantly increased the correlation between the testing and reference GWAS results especially for traits with heritability > 0.05, (Figure 13, Figure 14).

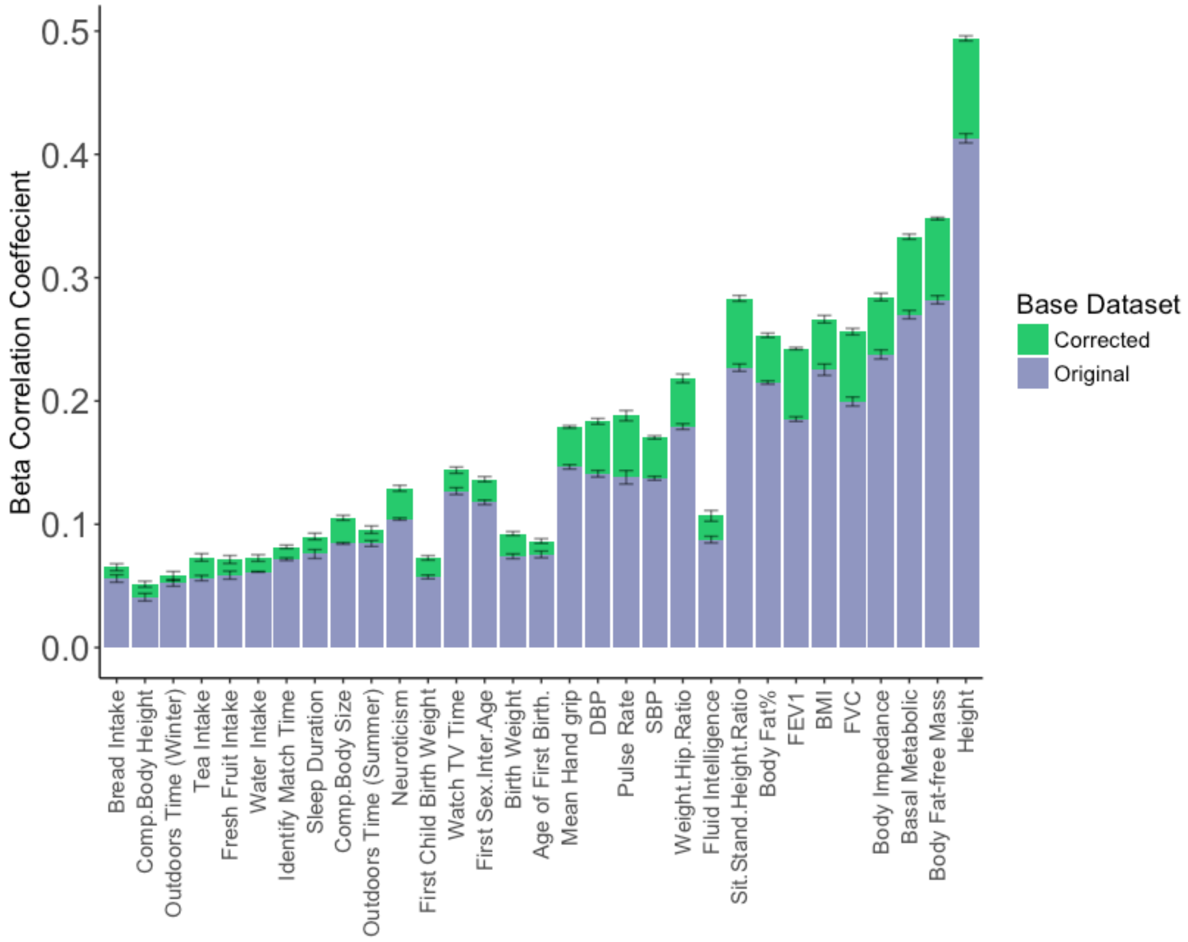


Figure 13 Correlation coefficient between effect sizes estimated from the testing data and those from larger reference data increased after correction. X-axis shows the phenotypes, ranked ascending from left to right according to their estimated heritability (for their estimated heritability, please refer to Table 2 UK Biobank traits used to test the methods); the blue bars shows the correlation coefficients of the original effect size estimated from the testing data with those from the larger reference data; the green bars shows the increased correlation coefficients between the corrected effect size estimates and the estimates calculated from the reference data.

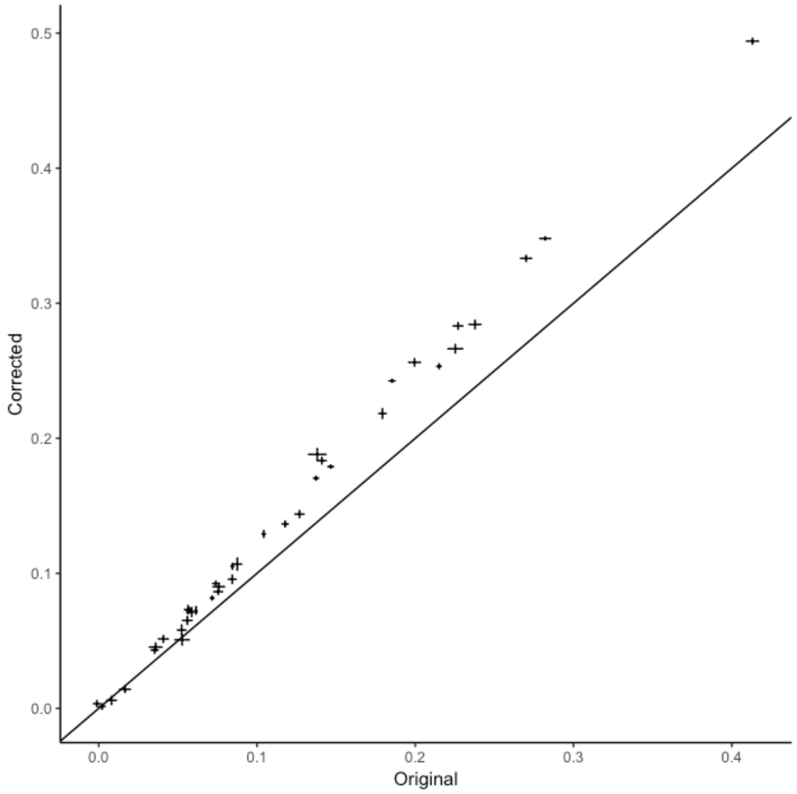


Figure 14 Increase of corrected effect size estimates correlation coefficients to the original ones. The x-axis shows the correlation coefficients of original effect size estimates of testing dataset and those of reference dataset; The y-axis shows correlation coefficients of corrected effect size estimates of testing dataset and those of reference dataset. The data points located above the reference line y=x indicates the increase of correlation coefficients after the correction.

2.4.4. Testing performance of Permutation Shrinkage on increasing PRS prediction with real data

In order to test whether the method can improve the PRS prediction power, the out-of-sample PRS R^2 were compared before and after the correction. When estimated heritability < 0.05, the relative PRS R^2 increase was fluctuated around zero. For more heritable traits with estimated heritability >0.05, our method significantly increased the out-of-sample PRS R^2. The relative PRS R^2 increase was flatten out as the h^2 estimated increase and the average of increase is around 35% (Figure 15, Figure 16).

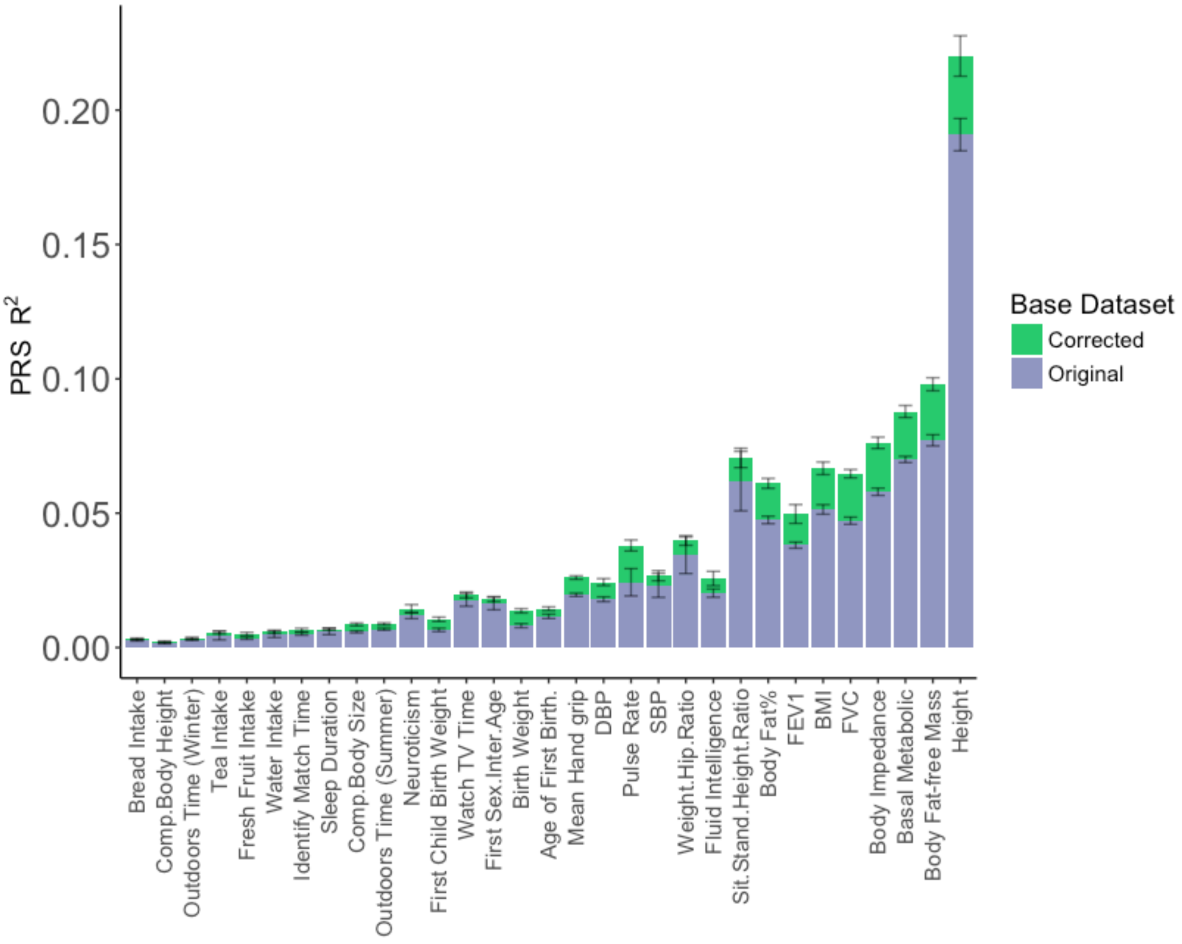


Figure 15 PRS R^2 calculated in the independent validation dataset increased after correction. X-axis shows the phenotypes, ranked ascending from left to right according to their estimated heritability (for their estimated heritability, please refer to Table 2 UK Biobank traits used to test the methods); The blue bars show the PRS R^2 based on the original effect size estimates of the discovery data; the green bars show the increased PRS R^2 after correction of the effect size estimates of the discovery data. The PRS model optimized the P-value threshold in testing dataset and calculated the PRS R^2 with the previously optimized P-value threshold in an independent validation dataset. The process of dividing UK Biobank randomly to get the discovery, testing and validation data and performing the analysis were repeated for 5 time to get the mean and sd of PRS R^2.

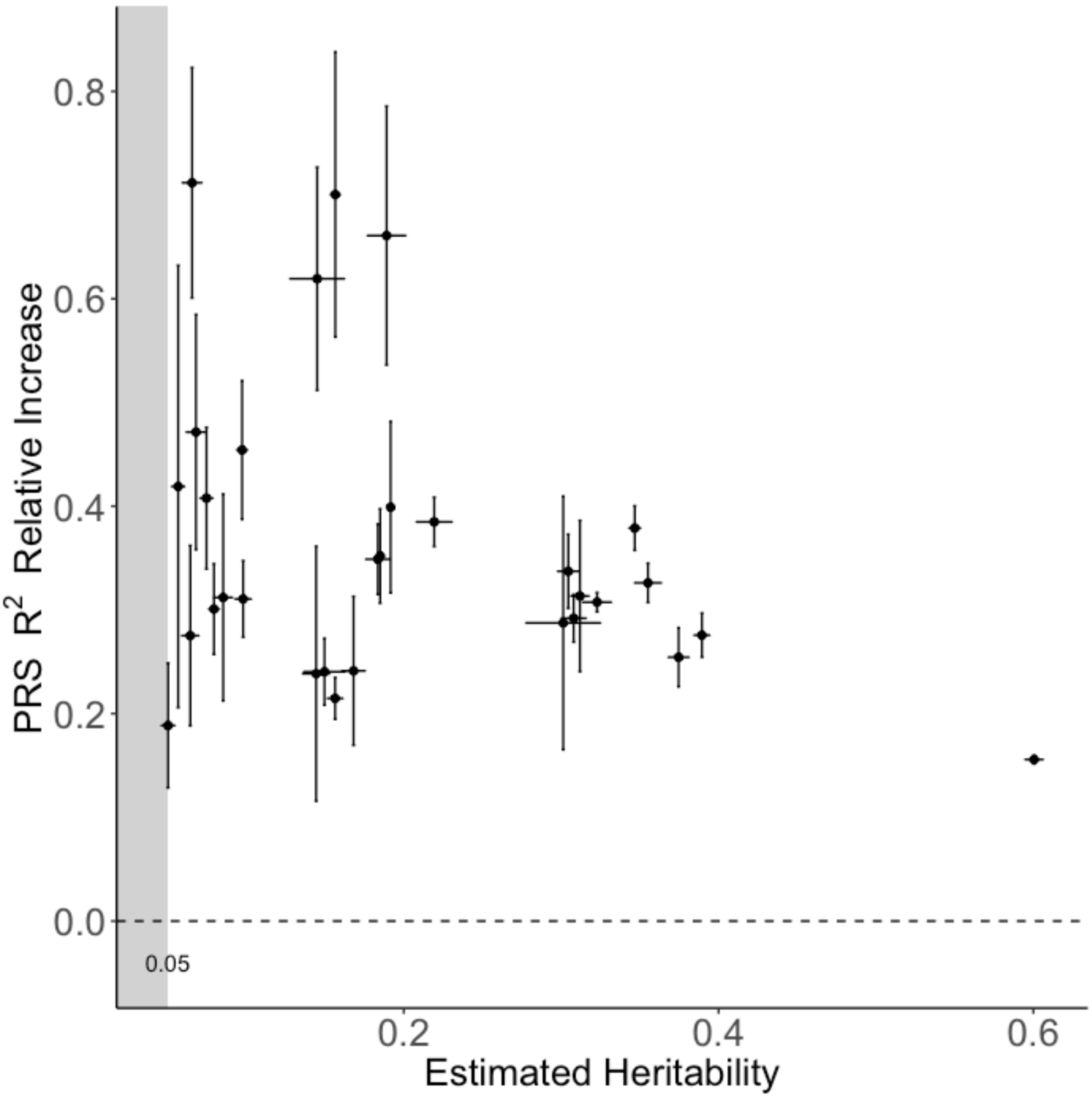


Figure 16 Relative increase of PRS R^2 of traits of estimated heritability >0.05. The data shown here is based on 5 repetitions. The PRS results is the same as those shown in the previous plots. Figure 15 shows the PRS R^2, here shows the relative increase of PRS R^2 of the same traits. The PRS R^2 of the traits of estimated h^2<0.5 fluctuates around 0. Therefore, the relative PRS R^2increase is omitted here.

2.4.5. Testing performance of Permutation Shrinkage on increasing PRS prediction with simulated data

In last two sections, we showed that Permutation Shrinkage can improve the PRS prediction of real traits by improving the effect size estimates. In order to further test our method under different possible genetic architectures, Permutation Shrinkage was tested with simulated phenotype data of different percentage of causal SNPs and different heritability based on UK Biobank genotype data as described in Section 2.3.5. A higher percentage of causal SNPs would generate a more polygenic trait. Two simulation methods were used in this test: simulated with standardized genotype by GCTA and simulated with unstandardized genotype by the in-house software SGCP. Standardized genotype unified the effect of MAF and unstandardized genotype use the genotype data as they are. By standardising the genotype, SNPs of low MAF would have more effect on the phenotype.

However, both methods generated unexpected results. The authors of LD Score Regression (LDSC) reported that the estimate of heritability is unbiased regardless of the percentage of causal SNPs[30]. However, for both simulation methods, the estimated heritability is lower than the simulated heritability. For simulated traits based on unstandardized genotype, the more oligogenic trait, the more underestimated the heritability is (Figure 17). For simulated traits based on the estimated heritability of the trait simulated based on standardized genotype, the heritability estimates were much noisier and have the opposite tendency regarding the percentage of causal SNPs: the more polygenic trait, the more underestimated the heritability is. This tendency was the same whether the distribution of causal SNP effect size was normal distribution or chi squared distribution (Figure 18).

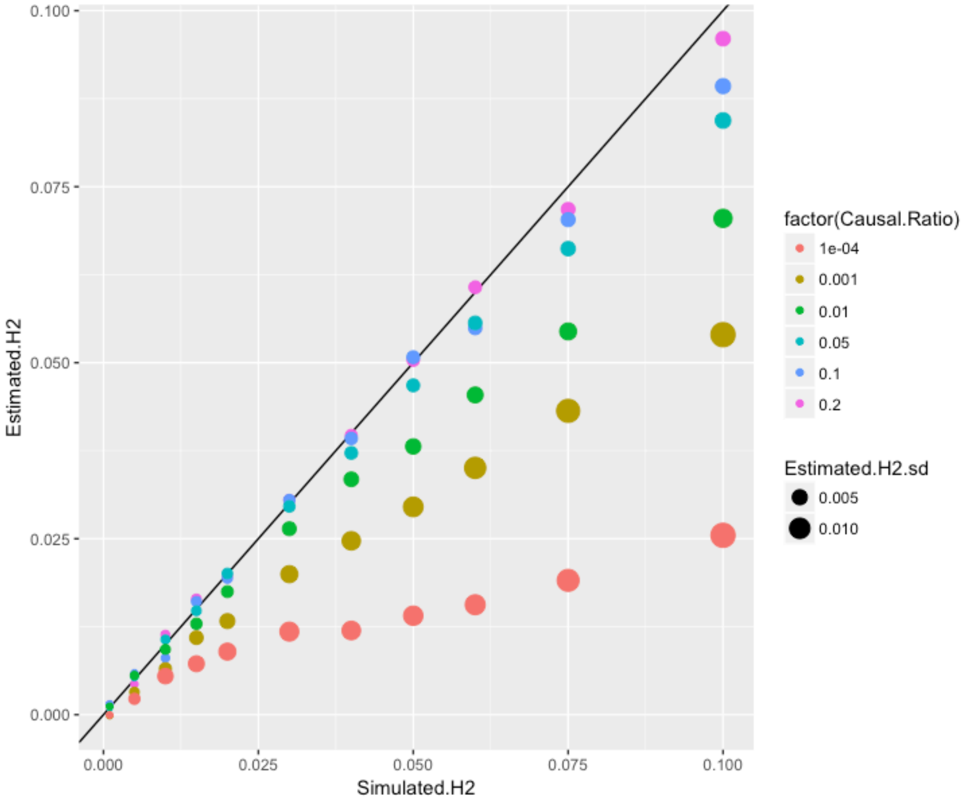


Figure 17 Estimated heritability of the simulated phenotype data based on UNSTANDARDIZED genotype data compared with the pre-set simulated heritability under different scenarias. The black reference line is y=x. Theoretically, the data points should locate on the reference line. The mean and sd heritability shown in the plot are based on 5 repetitions.

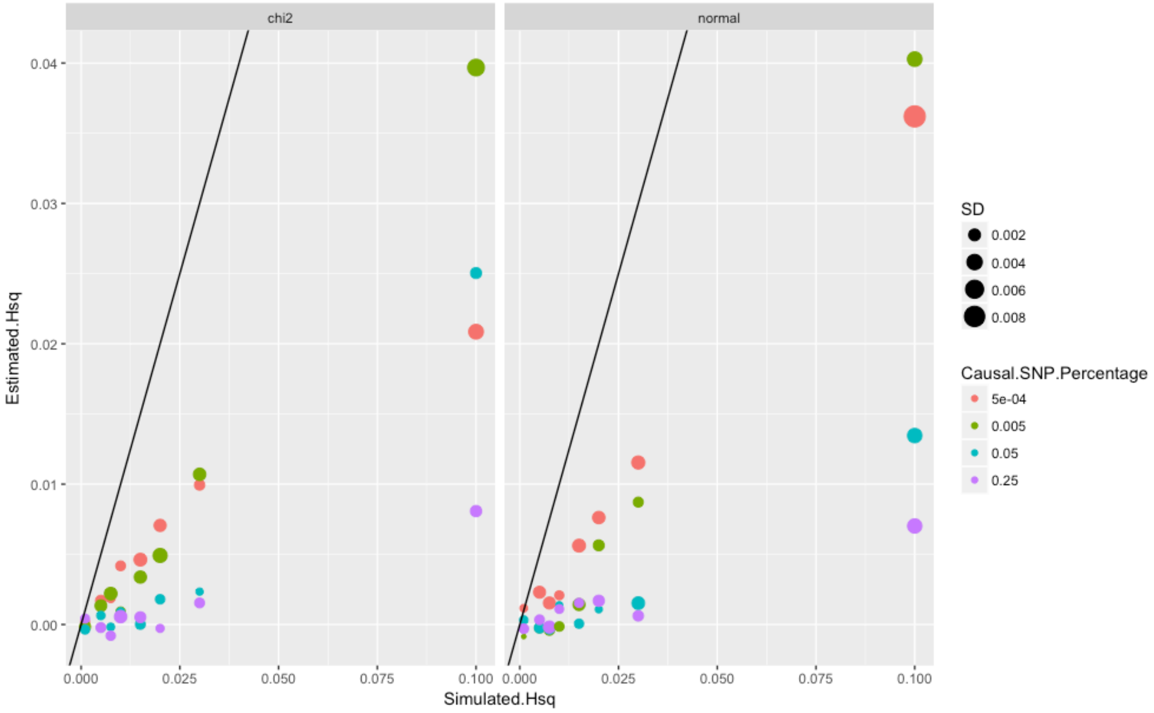


Figure 18 Estimated heritability of the simulated phenotype data based on STANDARDIZED genotype data compared with pre-set simulated heritability under different scenarios. Theoretically, the data points should locate on the reference line. The mean and sd heritability shown in the plot are based on 5 repetitions.

The PRS prediction based on unstandardized genotype simulation behaved similarly with real traits data. The relative increase was relatively high when the simulated heritability was lower and flatten out when the simulated heritability was high (Figure 19). The PRS prediction based on standardized genotype simulation were also much noisier and didn't show any clear relationship pattern either between relative PRS R^2 increase and simulated heritability (Figure 20) nor between relative PRS R^2 increase and estimated heritability (Figure 21).

The systematic bias observed can be caused by either the simulation methods or the heritability estimation method. However, testing the validity of these methods are far beyond the aim and the scope of this PhD project. Future work is needed to tackle this problem because the validity of simulation and heritability estimation can influence many genetic researches. Judging from the results observed in this project, the simulation based on the unstandardized genotype seemed more similar with the real data.

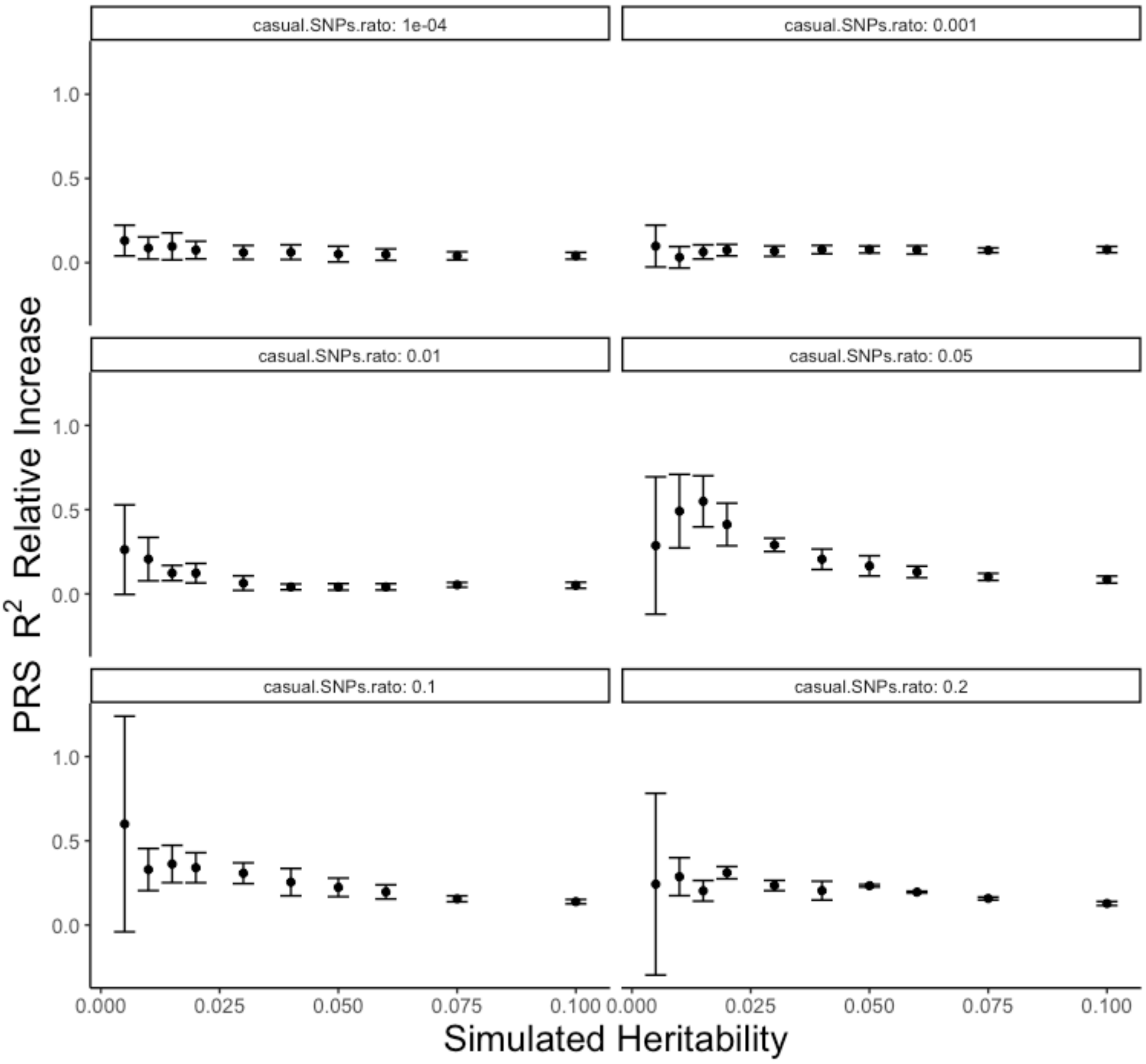


Figure 19 Relative increase of PRS R^2 calculated with simulated data based on UNSTANDARDIZED genotype data. Simulated GWAS data were divided into discovery, testing and validation datasets. The SNP effect size estimated were calculated and then corrected with discovery data; the P-value threshold were optimized with testing data and the PRS R^2 were calculated with validation data. Subplots show the scenarios under different ratio of causal SNPs. X-axis shows the heritability pre-set for simulation; y-axis shows the relative PRS R^2 increase after correcting the SNP effect size estimates in discovery data. The mean and sd PRS R^2 relative increase are based on 5 repetitions. The result generated with unstandardized genotype data have the similar pattern with real data.

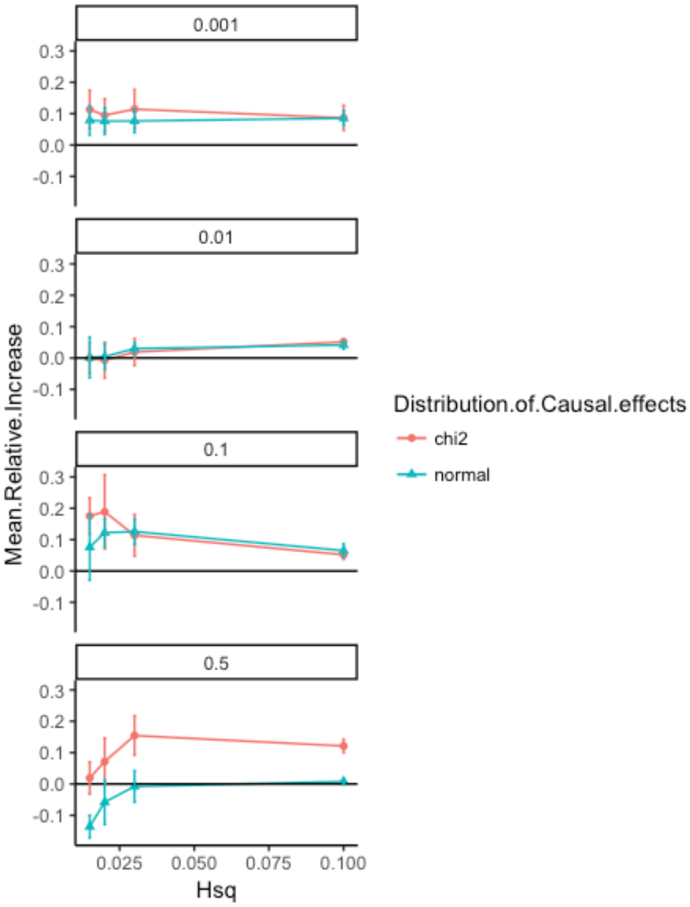


Figure 20 Relative increase of PRS R^2 in simulated data based on STANDARDISED genotype. Simulated GWAS data were divided into discovery, testing and validation datasets. The SNP effect size estimated were calculated and then corrected with discovery data; the P-value threshold were optimized with testing data and the PRS R^2 were calculated with validation data. The subplots show the scenarios under the ratio of causal SNPs of 0.001, 0.01, 0.1, 0.5. The colour indicates the distribution of simulated effect size. For chi squared distribution, the signs of effect size were randomly assigned. X-axis shows the simulated heritability; y-axis shows the relative PRS R^2 increase. The mean and sd of relative increase of PRS R^2 are based on 5 repetitions. The result generated with standardized genotype data did not follow the similar pattern as the real data, regardless the distribution of SNP effect sizes.

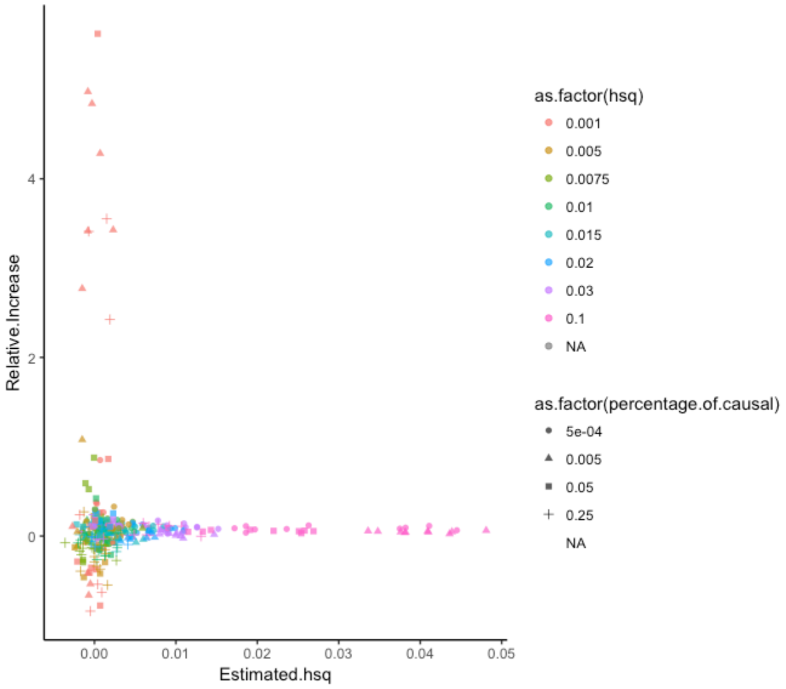


Figure 21 Relative PRS R^2 increase of simulated data based on STANDARDIZED genotype data. Simulated GWAS data were divided into discovery, testing and validation datasets. The SNP effect size estimated were calculated and then corrected with discovery data; the P-value threshold were optimized with testing data and the PRS R^2 were calculated with validation data. X-axis shows the estimated heritability estimated with the discovery data, y-axis shows the relative PRS R^2 increase. Different colours and shapes of data points shows the pre-set simulated heritability and the ratio of causal SNPs, respectively. Each data point is based on one single repetition. The plot shows that the estimated heritability did not match with the pre-set simulated heritability and the pattern of the relative increase of PRS R^2 did not accord with that of real data.

2.5.Conclusions and Discussion

GWAS results are now being used for prediction in both theoretical research and application such as precision medicine. Polygenic risk score prediction is heavily dependent on GWAS results. Spurious effect size estimates will make the PRS R^2 unreliable.

Permutation Shrinkage was developed to utilise the increasing availability of raw genotype data to increase the accuracy of GWAS and the prediction methods based on GWAS. It estimates the null distribution of the effect size and corrects the observed effect size by taking the ‘noise’ away from the raw estimates. Despite its simplicity, it increased the PRS prediction significantly in the tests.

Permutation Shrinkage assumes that the SNPs in the same MAF bin would have similar effect sizes or only a small proportion of SNPs have different effect sizes. This assumption is not always true but may be very close to the reality: the causal SNPs were estimated to only take a small proportion (<5%) even in polygenic traits[71].

Although permutation can be very computational expensive, it can run in parallel in a high-performance cluster and each string requires a relatively small RAM and is independent with each other. It can be faster and more robust than other methods using raw genotype data most of which compute the correlation matrix of the genotype data and take huge RAM at one time.

In the tests, the simulation based on standardised and unstandardized genotype data generated different result and the estimated heritability of both simulated data set were biased from the simulated heritability, which indicated that either the simulation methods or the heritability estimation method was problematic. However, this problem haven't been fully analysed and solved due to the limited time and scope of this project. Judging from the pattern of out-of-sample relative PRS R^2 increase, the simulation based on unstandardized genotype had a very clear pattern that was similar with real data while the simulation based on standardized genotype generated noisy estimated heritability and noisy out-of-sample relative PRS R^2 increase. This result indicated that simulation based on unstandardized was closer to the real data and low-MAF SNPs should not be assigned with higher effect size to the phenotype. Yet more investigated is needed to test these hypotheses.

One limitation of this project is that only the shrinkage method for quantitative traits have been developed and tested. In the future work, we would adjust the method so that it can work for binary traits. The preliminary plan is to use the same framework except that we estimate the null distribution of absolute value of log(odd ratio) and corrected the observed log(odd ratio). Log(odd ratio), instead of odd ratio, is used because its null distribution is more similar with normal distribution and therefore it is easier to evaluate whether the result is reasonable.

Another limitation is that Permutation Shrinkage only works on raw genotype data. Although raw genotype data have been increasingly available, the majority of available GWAS data are summary statistics. Therefore, a similar shrinkage method that only utilises summary statistics GWAS data was developed and tested in next chapter.

Chapter 3. Order Statistics Shrinkage method for GWAS data

3.1.Introduction

In the previous chapter, Permutation Shrinkage (PS) was developed to increase the accuracy of GWAS effect size estimates and, therefore, improve GWAS-based prediction models. The PS method estimated the null distribution of absolute SNP effect sizes by permuting the raw individual-level phenotype data and shrank the GWAS effect size estimates by adjusting for the ranked value of this approximated null distribution. In our tests on real and simulated data, the corrected GWAS effect size estimates significantly increased the predictive power of polygenic risk scores (PRS). However, the PS method depends on the use of raw genotype-phenotype data for the base GWAS, which are often not available for most users, who will often want to compute polygenic risk scores in their own target sample by exploiting large-scale GWAS summary statistics. Therefore, an alternative shrinkage method that can utilise GWAS summary statistics, and exploit the intuition underlying PS, could have the benefits of PS but have greater utility.

In PS, the null distribution of effect sizes is estimated from permuting the phenotype values in the raw genotype-phenotype data. This is the only step that requires the individual-level data. If the null distribution can be estimated from summary statistic data only, then an alternative shrinkage method could be formulated.

Assuming that the null SNP effect sizes follow a known distribution or a distribution that can be approximated from a known distribution, then the null distribution of the effect sizes can be generated using order statistics of this distribution. Order Statistics (OS) are the ordered or ranked values of a statistic from a sample. The k^{th} order statistic is the k^{th} smallest value in the sample. Order statistics offer us an alternative method to generate the null distribution of SNP effect sizes.

In this chapter, another novel shrinkage method, Order Statistics Shrinkage (OSS), is developed. OSS uses order statistics, instead of permutation based on individual-level genotype data, to generate a null distribution of GWAS effect sizes so that the GWAS results can be adjusted for inflation and deflation when only summary statistic data are available.

3.2. Methods

3.2.1. Overview of method

Order Statistics Shrinkage (OSS) uses the framework of Permutation Shrinkage (PS) (see Chapter 2, especially 2.3.1-2.3.3), except that the null distribution of the absolute value of the SNP effect sizes is estimated with order statistics. In OSS, SNPs are also binned into 1% minor allele frequency (MAF) intervals and the generation of the null distribution and the correction of effect sizes are all performed within MAF bins. After the generation of the null distribution in each MAF bin, the process of subtracting the estimated error quantity from the observed SNP effect size estimate and making the corrected value monotonically increasing with the original value is the same as that in PS. The following section will describe how OSS estimates the null distribution in a MAF bin in detail.

3.2.2. Generating the null distribution with order statistics.

The generation of the null distribution can be divided into the following four steps:

1. Estimating the effective number of independent tests

In order to use the order statistics, the effective number of independent tests, M_e, needs to be first estimated because an assumption in the definition of order statistics is that samples are drawn independently from the same distribution. There are two options:

(i) Assuming that all the SNPs within the MAF bin are independent of each other or only modestly correlated, then the effective number of independent tests, M_e, can be assumed to be the same as the number of SNPs in the MAF bin. This assumption should be close to the truth if the data are sparse and the SNPs in the 1% MAF bin are generally distant from each other, meaning that the pairwise correlation of their genotypes is likely to be low. In the subsequent steps, the order statistics corresponding to the same number of all the SNPs in the MAF bin is generated and the effect sizes of all the SNPs are corrected and used as the input of the PRS calculation. This option is called “all SNPs”.

(ii) However, if the SNPs are highly correlated, for example, when using imputed genotype data or using a very dense genotype microarray, then the assumption of SNPs being independent cannot hold. Since our aim is to improve the performance of PRS, it is not necessary to have all the corrected SNPs because only clumped SNPs are used in the PRS calculation. In this high-density SNP scenario, in order to avoid the collinearity problem, the SNPs are first clumped in each MAF bin. As discussed previously (see section 1.6.2) clumping thins SNPs according to Linkage Disequilibrium (LD), retaining the most associated SNPs, such that a subset of SNPs remain that are relatively uncorrelated. We clump using an r^2 threshold of 0.1. Thus, in each MAF bin, the clumped SNPs can be viewed as independent and the effective number of tests, M_e, is approximated by the number of clumped SNPs. Order statistics of the number of clumped SNPs are generated and only the clumped SNPs are corrected and then used as the input for the PRS calculation and subsequent analysis. This option is called "clumped SNPs".

2. *Generating the order statistics of the one-side Z score*

In this approach to shrinkage, the absolute effect size under the null hypothesis is estimated. For quantitative traits, an absolute t-statistic can be transformed to an effect size, beta value, if the standard error SE is known:

$$|beta| = SE * |\, t\,| \tag{1}$$

If the SNPs are independent and M_e is known or estimated (ii above), then the order statistics vector of an absolute t-statistic, $|\boldsymbol{t}|$, can be produced by generating an ordered vector, ***U,*** containing M_e elements drawn from a uniform distribution, U[0,1], and then converting ***U*** to an absolute t-statistic via the absolute value of the probit function of the t-distribution, with N degrees of freedom, at a quantile point given by ***U***. Given that N, the size of the sample in which the *beta* is estimated, is often very large, then if N is unavailable then the absolute t-statistic can instead be approximated by an absolute Z-score similarly.

3. Generating the order statistics of absolute effect size

When the quantitative trait is standardised, the standard error of the null effect size SE is a function of the sample size, N, and the MAF, p.

From Bernado and Smith[1] the regression coefficients θ in a linear regression model with

a reference prior (minimises prior information) follow a Student distribution:

$$\theta \sim \mathrm{St}\left(\hat{\theta}_N, \frac{1}{2}X^T X(N-k)\hat{\beta}_N^{-1}, N-k\right)$$

k is the number of parameters we are fitting =2, intercept and SNP effect

therefore since

$$\hat{\beta}_N = \frac{1}{2}\left(y - X\hat{\theta}_N\right)^T y$$

$$\mathrm{V}(\theta) = \frac{(N-2)\left(y - X\hat{\theta}_N\right)^T y}{(N-4)X^T X(N-2)} \qquad \text{(i)}$$

With mean centred genotype data:

$$X^T X = N\mathrm{V}(X) = 2Np(1-p) \qquad \text{(ii)}$$

And mean centred and standardized y:

$$\left(y - X\hat{\theta}_N\right)^T y = y^T y - \hat{\theta}_N^T X^T y$$

$$= n - k$$

$$= n - 2 \tag{iii}$$

Plugging (ii) and (iii) into (i) gives

$$SE = \sqrt{\frac{N-2}{(N-4)N \times 2p(1-p)}} \tag{2}$$

There are two options for estimating the *beta* under the null from (1) using (2): with or without considering the MAF variation within each MAF bin. If the MAF of all the SNPs is the same, the null distribution of beta can be calculated via (2) using a fixed p. However, in real GWAS data, each SNP has similar yet still slightly different MAF, even if the SNPs are binned into 1% MAF intervals.

If the slight variation of the MAFs in each MAF bin can be ignored, then the SNPs in the 1% MAF bins can be viewed as having a fixed MAF, which can be the middle point of the MAF bin, $\bar{p}$. All SNPs in the same bin then have the same fixed SE:

$$\overline{SE} = \sqrt{\frac{N-2}{(N-4)N \times 2\overline{p}(1-\overline{p})}}$$

The vector of order statistics absolute beta is then approximated, modifying (1), as:

$$|\boldsymbol{beta}| = \overline{SE} \cdot |\, t \,|$$

This is described as the "fixed SE" option.

An alternative implementation is to account for the MAF variation. Each element of a vector of the SE of all the SNPs, $\boldsymbol{SE}$, is calculated using the sample size N and the MAF of each SNP as:

$$se_i = \sqrt{\frac{N-2}{(N-4)N \times 2p_i(1-p_i)}}$$

in which p_i is the MAF of ith SNP. If the missingness of the SNPs varies considerably, then the sample size N should be N_i, the number of samples that have the i^{th} SNP genotyped. The $\boldsymbol{SE}$ vector is then formed from computing se_i for the M_e SNPs. Since this approach simulates the process of PS, where the i^{th} elements of the ranked *t*-statistic can be assigned to any of the SNPs in the MAF bin, the $\boldsymbol{SE}$ vector should be shuffled to calculate the $|\boldsymbol{beta}|$.

In this option, $|\boldsymbol{beta}|$ is a vector comprising the element-wise products of a randomly shuffled $\boldsymbol{SE}$ and the *t*-statistic (or Z-score). This option is described as the “random SE” option.

4. *Taking the average of ordered statistics of absolute effect size*

The order statistics directly generated in steps 2 and 3 are random variables that include stochastic variation due to random sampling of the uniform distribution in step 2, and random shuffling of the $\boldsymbol{SE}$ vector in step 3. Thus, the null distribution estimated by order statistics in one round can be viewed similarly to the null distribution generated in one round of permutation. In order to get a smoother estimate of the theoretical null distribution, step 2 and step 3 are repeated 100 times to get the average of the ordered effect size vectors as the resulting null distribution.

3.2.3. Alternative summary statistics

In the OSS method, the SNP MAF in the base data is necessary for binning the SNPs and calculating the null effect size distribution. However, the SNP MAF is not available in all summary GWAS data. Alternatively, the MAF of the SNPs from a reference data set can be used as a proxy. For example, if the base and target samples are drawn from the same or similar populations, which is often the case since PRS have been shown to only generalise well to similar populations[29,33,78], then the MAF of the target data SNPs can be used. Thus, the MAF of the target data set can be used as a proxy of the SNP MAF in base data to bin the SNPs into 1% interval bins of MAF and to calculate the standard errors of the effect size estimates. Alternatively, the standard error of the *beta* estimated in the GWAS can be used as a proxy of the standard error of the null effect, assuming that the standard error corresponding to the effect size estimate of a null SNP is the same as that of a SNP with the same MAF but non-zero effect size.

The focus of this chapter is to investigate the performance of OSS. The best strategy to take when the MAF of the SNPs in the base data are unavailable will be setting specific and is not investigated here, so hereafter we only test the performance of OSS when the MAF information is available.

3.2.4. Performance evaluation under simple simulation scenarios

In order to understand the general performance of the OSS approach compared to standard shrinkage approaches – the James Stein estimator, lasso and ridge regression – in simple generic scenarios, a set of 1000 SNPs were simulated in a sample of 10k individuals, each with a Minor Allele Frequency (MAF) of 0.25, and different fractions and effect sizes of causal SNPs with independent additive effects were modelled, with a total heritability of 0.1, and a Gaussian error term with variance 0.9 reflecting the residual trait variance. So, for example, in one scenario, 1000 SNPs were simulated to have genotypes 0, 1 and 2, each with MAF=0.25 and thus, assuming Hardy Weinberg Equilibrium (HWE), having expected frequencies of 0.5625, 0.375 and 0.0625; 100 SNPs had a causal effect of the minor allele of 1 standard deviation, 100 had an effect of -1 standard deviation, while the other 800 SNPs had no effect. In another scenario, all 1000 SNPs had the same effect size, thus the trait variance explained by each SNP was 0.1%.

The performance of the shrinkage methods was tested by comparing the mean squared errors (MSE) before and after the shrinkage:

$$\mathrm{MSE} = \frac{1}{N_{SNP}} \sum_{i=1}^{N_{SNP}} (Obersved\ beta_i - Simulated\ beta_i)^2$$

JS estimators were calculated with in-house R scripts based on the definition of the JS estimator as described in Chapter 1. Lasso and ridge estimates were calculated with R package “glmnet”. Glmnet gives the lasso estimator when alpha=1, and the ridge estimator when alpha=0.

3.2.5. Performance evaluation using real and simulated data based on the UK Biobank

The methods were tested using the same set of real data traits in UK Biobank used (see section 2.3.4) and with in-house software written by Dr. Shing Wan Choi described (see section 2.3.5) The heritability of the real traits was estimated with LD Score regression[30], using the default parameters. The GWAS data for testing the methods was split into 3 equal parts: discovery data, testing data and validation data. The PRS model was built and validated with the 3-dataset system as described in section 2.3.6.

3.3.Results

3.3.1. Comparison of OSS and PS for generating the null distribution

As discussed in 3.1, the only difference between PS and OSS is how the null effect size distribution is estimated. Therefore, the similarity of the null effect size distribution indicates the similarity of the corrected effect size estimate. In the simple simulated scenario, the SNPs are independent of each other and have the same MAF of 25%. The null effect size distribution in the simple simulated scenarios generated by permutation and prototype order statistics were almost identical, with the correlation coefficient = 0.9999 (Figure 22). The high similarity of the null distribution generated by PS and OSS means that the performance of these two methods should be almost identical, too. Therefore, in the subsequent testing in the simple simulated scenarios, we only compare the performance of OSS with the other approaches, and not with PS since OSS and PS give almost identical results

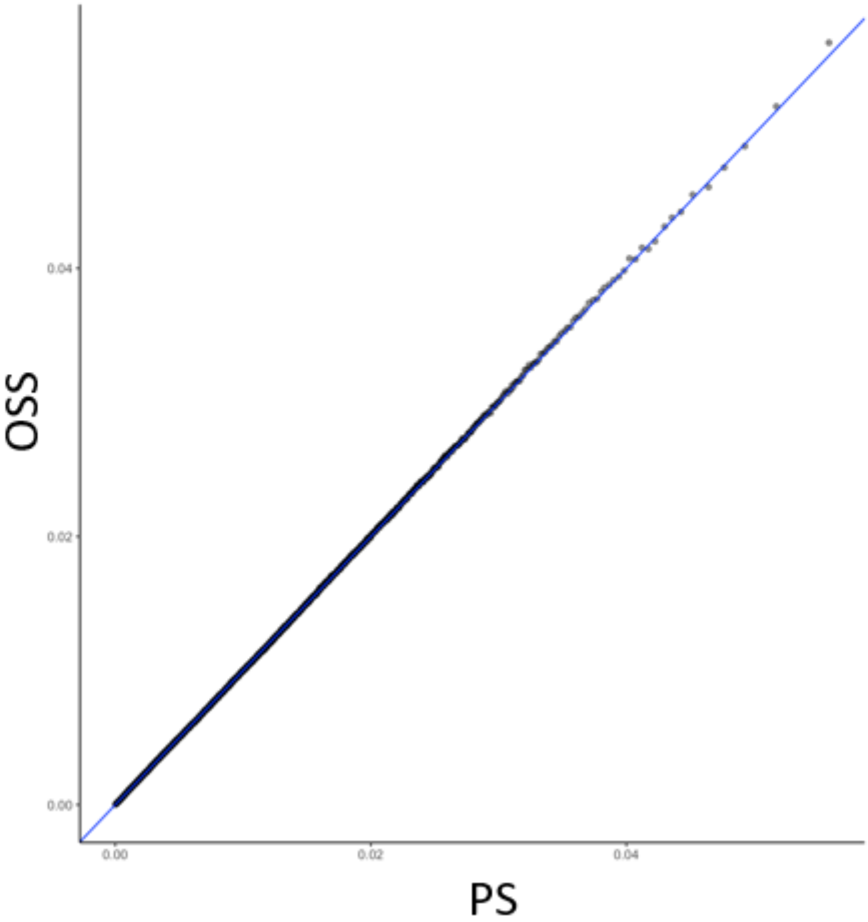


Figure 22 Compare the null distribution generated by PS and OSS in simulated GWAS data. The blue reference line is y=x.

3.3.2. OSS versus alternative shrinkage methods in simple simulation scenarios

Here, simple scenarios were simulated to test the performance of the basic performance of the OSS method and compared with other competing approaches: the James-Stein estimator (JS), lasso and ridge regression shrinkage methods. 10k individuals were simulated, each of which had 100 SNPs of Minor Allele Frequency (MAF) of 25%. A fraction, or all, of these 1000 SNPs had a causal effect on a standardised quantitative trait corresponding to a heritability of 0.1, while the remaining trait variance was modelled as a single Gaussian distributed residual error. Different proportions of causal SNPs, and the distribution of the effect sizes, were simulated to represent different genetic architectures (Table 3).

Table 3 Shrinkage methods performance under different simple simulated scenarios. The table is based on 10 repetitions.

Scenario	MSE （X10^{-5}）				
	Original observed	OSS	Ridge	Lasso	JS
All SNPs have the effect size =0	26.44±0.13	0.05±0.05	0.03±0.04	0.01±0.03	0.0001±0.0000
All SNPs have the effect size =1	26.90±4.43	**0.04±0.03**	13.13±5.22	18.58±8.38	26.53±3.28
50 SNPs have effect size drawn from *N*(0,1)	26.56±5.88	6.86±0.44	13.18±2.65	**4.31±0.45**	26.52±0.34
25 SNPs have effect size =1 and 25 SNPs have effect size =-1	26.83±2.03	8.85±0.74	12.91±0.18	**4.87±0.52**	26.73±0.30
100 SNPs have effect size =1 and 100 SNPs have effect size =-1	26.23±2.09	14.10±0.77	12.76±0.46	**11.70±0.65**	28.63±0.86
200 SNPs have effect size =1 and 200 SNPs have effect size =-1	26.29±4.00	15.97±0.54	**13.15±0.58**	15.54±0.70	26.66±0.45
400 SNPs have effect size=1 and 400 SNPs have effect size =-1	27.30±4.18	15.27±0.35	**13.21±0.52**	18.36±0.94	26.43±0.26
500 SNPs have effect size=1 and 500 SNPs have effect size =-1	26.41±1.74	15.38±0.45	**13.35±0.45**	18.97±0.51	26.90±026
100 SNPs have effect size =2 and 100 SNPs have effect size =-1	26.61±0.90	14.59±0.77	13.04±0.58	**10.90±0.56**	26.51±0.30
100 SNPs have effect size =2 and 100 SNPs have effect size =1	26.92±0.90	13.55±0.56	13.05±0.62	**10.85±0.60**	26.53±0.33

The methods with the best performance are marked with bold font.

In Figure 23- Figure 32 we illustrate each of the results from Table 3 in plots that show the unadjusted estimates of the effect sizes and the effect sizes estimates after adjustment via each of the shrinkage methods. These indicate some of the characteristics of each of the shrinkage methods.

Figure 23 shows the result for which none of the 1000 SNPs has effect on the output. All the four methods shrank the observed effect size to the true value, zero.

Figure 24 shows the results for which all 1000 SNPs have the same non-zero effect size. While the James Stein estimator shrinks all of the effects to close to 0, which continued to happen in all the scenarios, the effect sizes adjusted by the OSS approach estimates the simulated effects extremely accurately. Lasso and Ridge regression produce estimates that are slightly improved over the unadjusted estimates in this scenario. Lasso regression also had the tendency to shrink the effect size of non-zero SNPs to zero, which was also observed in the scenario described in Figure 27 - Figure 32.

The good performance of OSS in Figure 24 might result from the fact that this scenario perfectly accorded with the assumption of OSS that all the variable had the same effect size. However, in GWAS, it is unrealistic to assume that all the SNPs contribute equally to the phenotype. Therefore, more realistic scenarios (Figure 25 and Figure 26), in which the percentage of causal SNPs are 5% as the real scenarios[71] was simulated.

The performance of the four shrinkage methods were further investigated under the scenario where the percentage of causal variable increased in Figure 26 - Figure 30. Of the 4 summary statistics methods tested, the James-Stein estimator shows a consistent tendency to over-shrink the estimates to zero and therefore always has much higher MSE than the other methods. The OSS method preforms well in the scenario that accords with the assumption of PS and OSS, that is, that almost all the variables have the same effect sizes and the proportion of outliers from this are very small. As the percentage of causal variable increased, the performance of OSS and Ridge regression decreased while the performance of ridge regression was relatively consistent.

The scenarios where the effect size distribution of causal variables was dissymmetric were simulated in Figure 31 and Figure 32. Compared with the scenario in Figure 27 where the percentage of the causal variables was also 20%, the MSE of the four methods were similar but OSS had the tendency to “shrink” all the estimates to the average value of true effect size. When the effect size distribution is symmetric, the null variable will have corrected estimates that is close to zero and the causal variable will have corrected estimates that systematically smaller than their true effect size. However, when the distribution is dissymmetric, the null variables will have non-zero corrected estimates, which is obvious in Figure 32. When the majority of the variable are null and it is important to distinguish the null and causal variables, the latter scenario is worse than the former.

It is interesting to note, that the scenario here with 25 causal SNPs with effect 1 and 25 SNPs with effect -1, the performance of OSS is markedly better than the scenario where 100 SNPs have an effect of 1 and 100 SNPs have an effect of -1 (Table 3). This indicates the improved performance of the OSS method in scenarios where the effects overall are more homogenous; here, that more of the SNPs have no effect, since the fraction of causal SNPs here is 5% rather than 20%. It is also interesting that the OSS method performs well in the scenario where effects are drawn from a Gaussian distribution, demonstrating that the OSS method can perform well even when the causal effect sizes are not equal. Also, these scenarios, in which the OSS approach performs relatively well, may be more reflective of real data settings such as in GWAS of complex traits than those of the simple scenarios corresponding to Table 3 and Figure 23 - Figure 32.

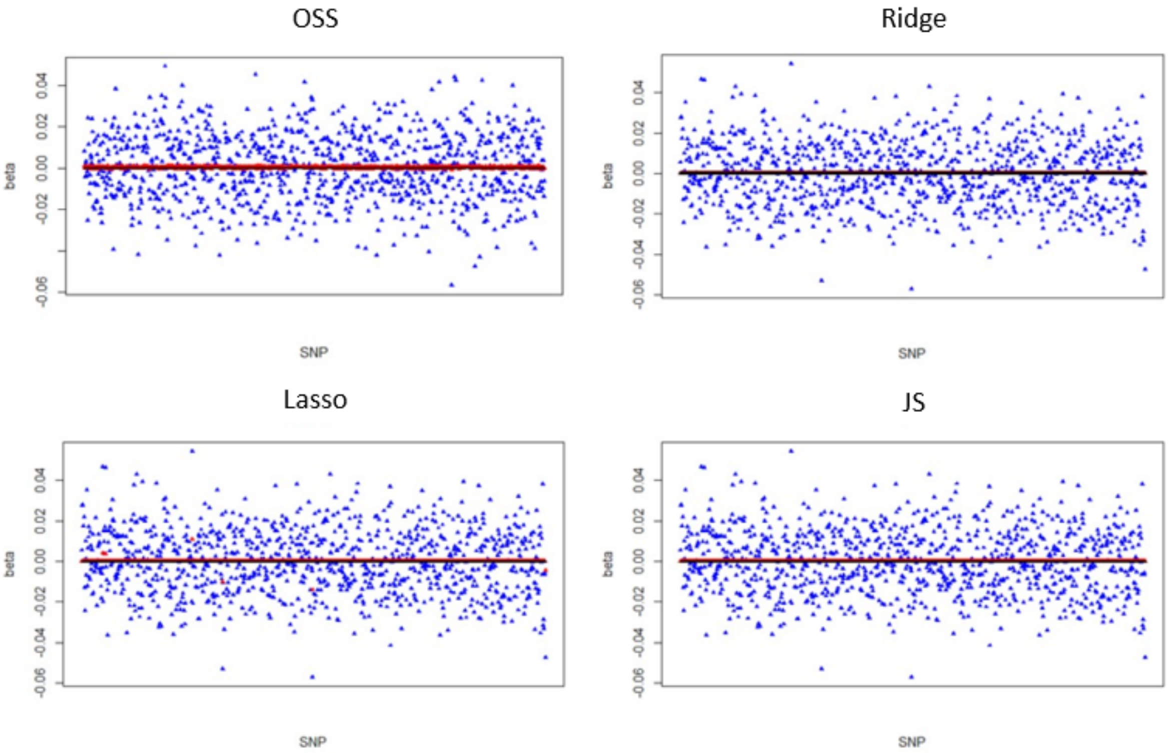


Figure 23 The performance of shrinkage methods in the scenario in which all the SNPs have zero effect size. The simulated, originally observed and corrected effect sizes are marked with black, blue and red colour. The plots are based on one round of simulation.

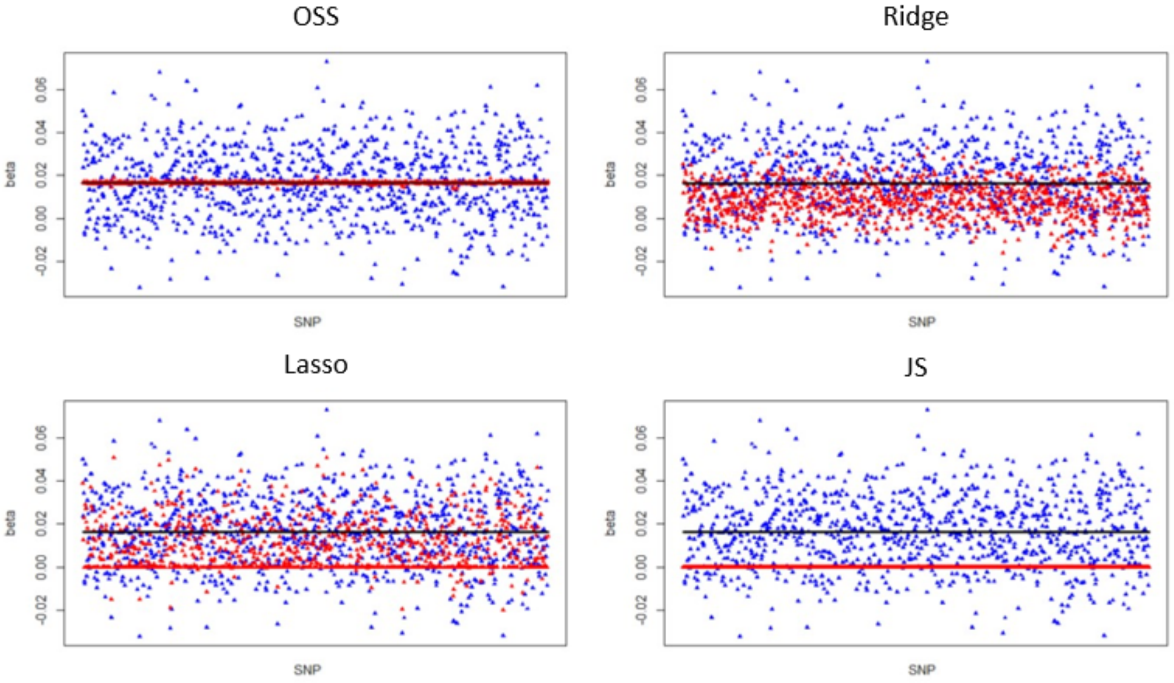


Figure 24 The performance of shrinkage methods in the scenario in which all the SNPs have the same non-zero effect size. The simulated, originally observed and corrected effect sizes are marked with black, blue and red colour. The plots are based on one round of simulation.

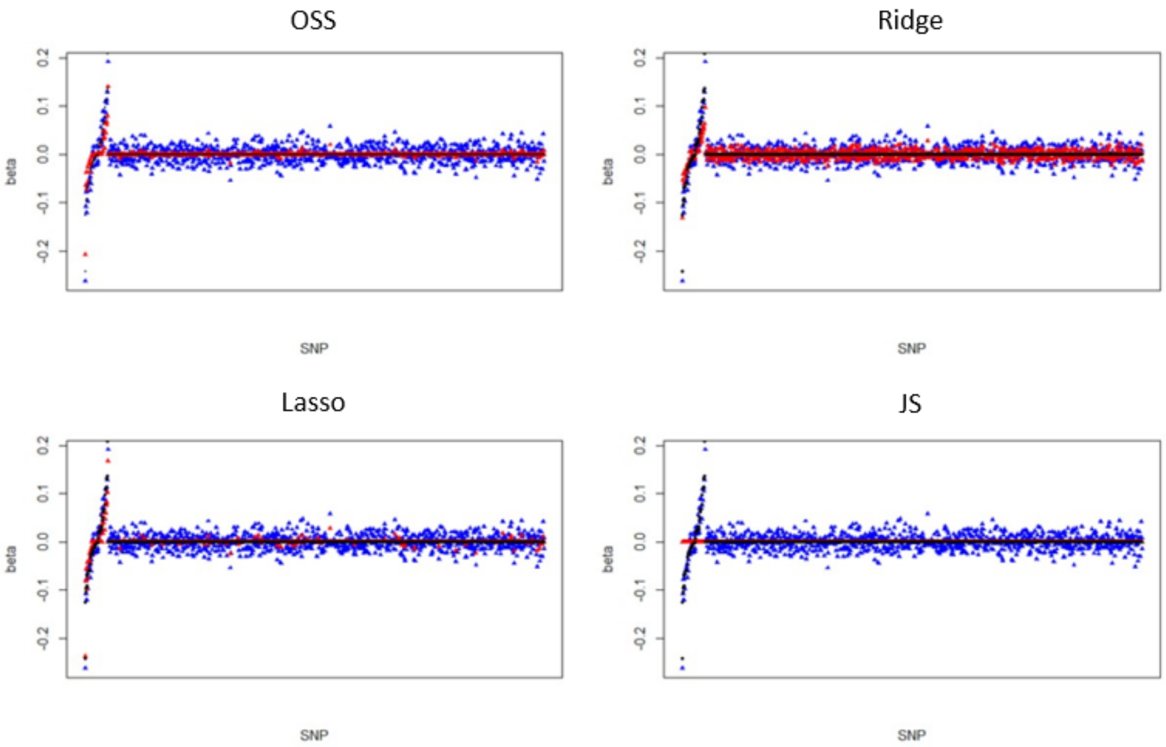


Figure 25 The performance of shrinkage methods in the scenario in which 50 SNPs have the effect sizes drawn from standard normal distribution. The simulated, originally observed and corrected effect sizes are marked with black, blue and red colour. The plots are based on one round of simulation.

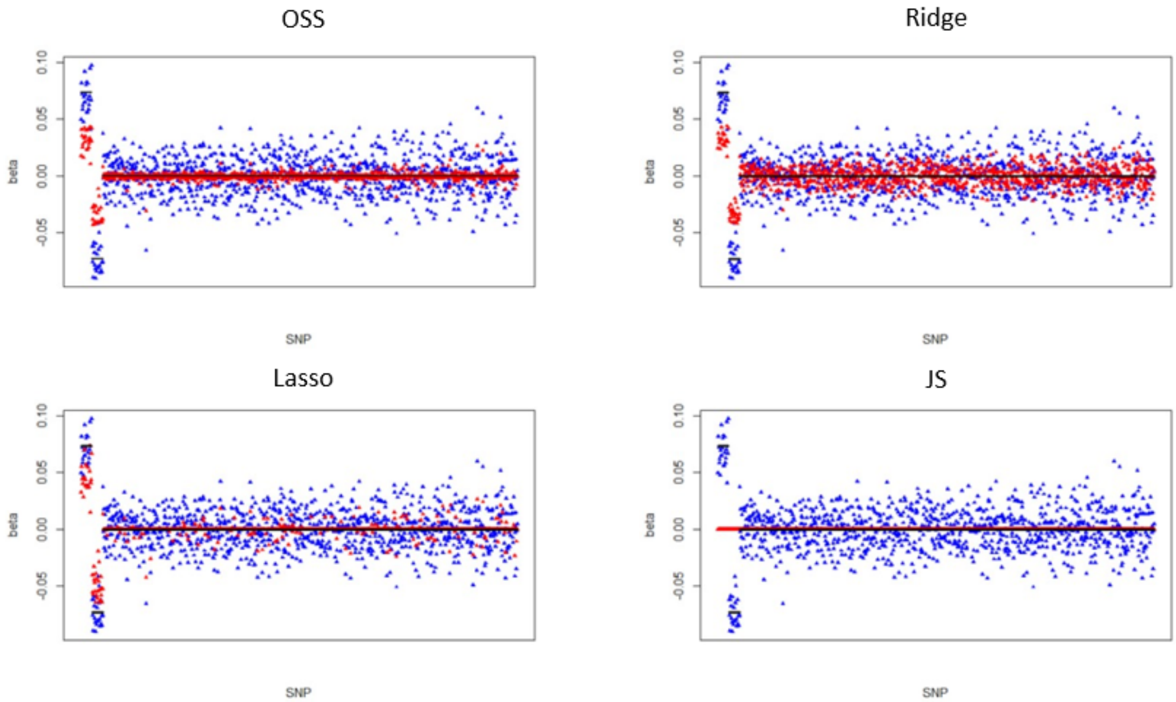


Figure 26 The performance of shrinkage methods in the scenario in which 25 SNPs have the effect sizes of 1 and 25 have the effect size of -1. The simulated, originally observed and corrected effect sizes are marked with black, blue and red colour. The plots are based on one round of simulation.

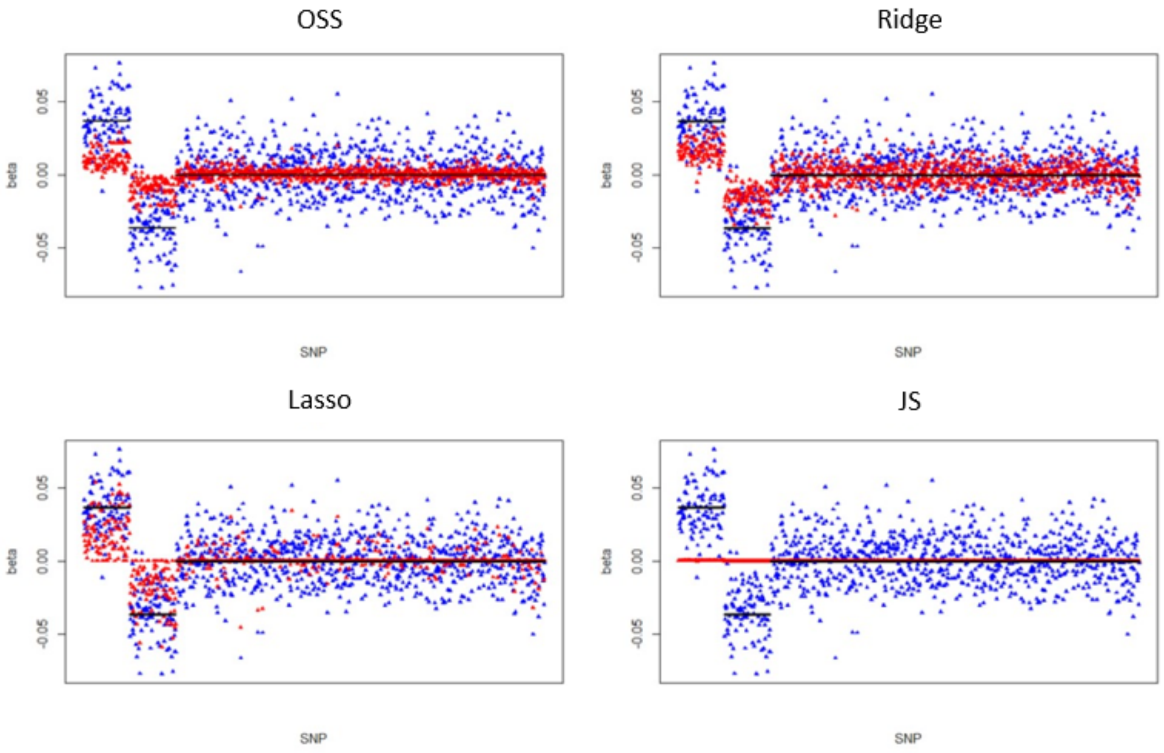


Figure 27 The performance of shrinkage methods in the scenario in which 100 SNPs have the effect sizes of 1 and 100 have the effect size of -1. The simulated, originally observed and corrected effect sizes are marked with black, blue and red colour. The plots are based on one round of simulation.

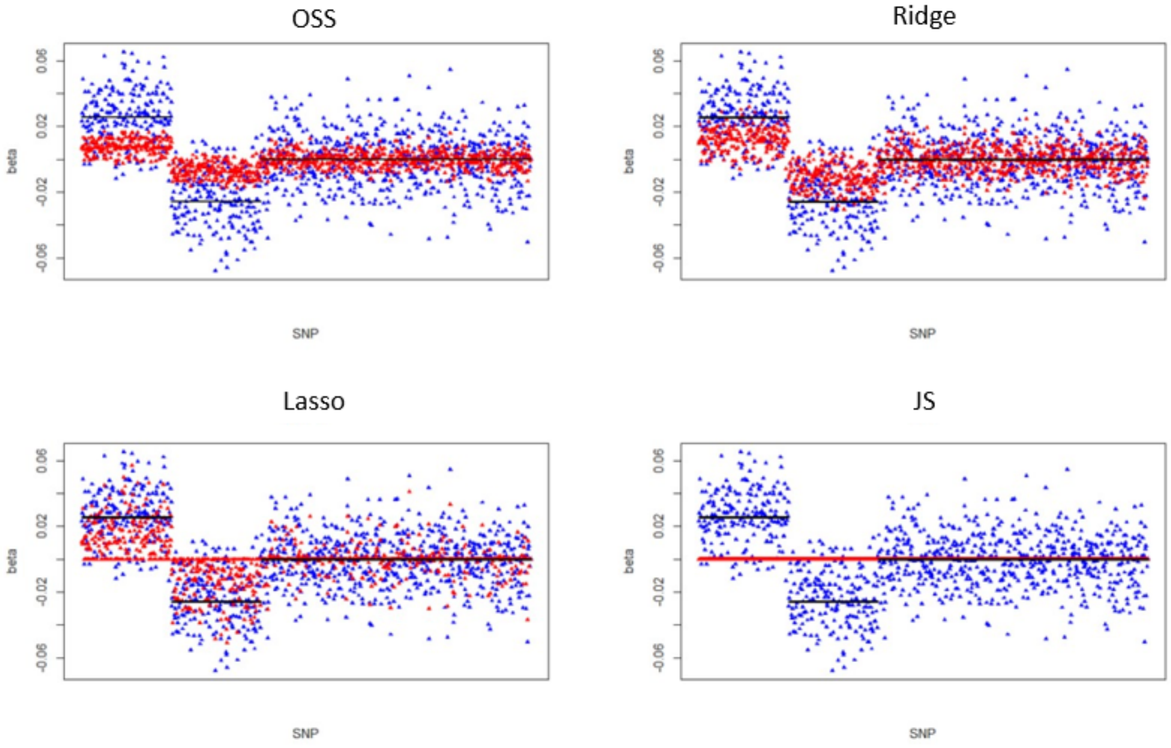


Figure 28 The performance of shrinkage methods in the scenario in which 200 SNPs have the effect sizes of 1 and 200 have the effect size of -1. The simulated, originally observed and corrected effect sizes are marked with black, blue and red colour. The plots are based on one round of simulation.

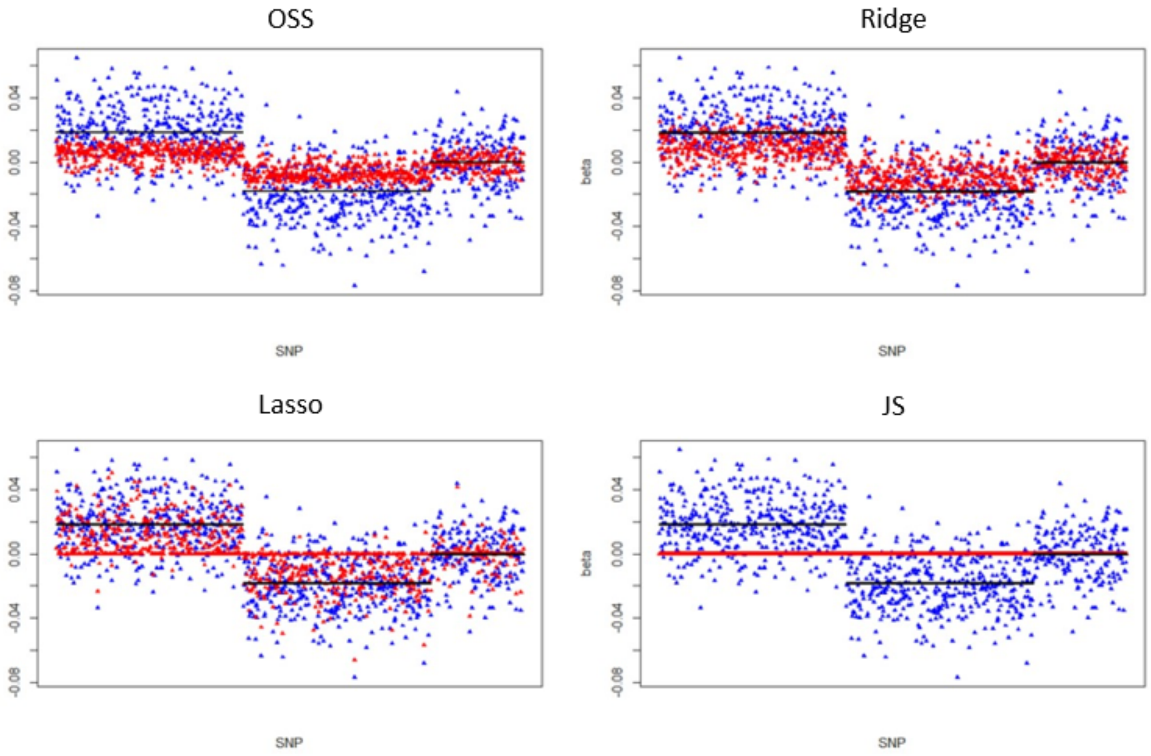


Figure 29 The performance of shrinkage methods in the scenario in which 400 SNPs have the effect sizes of 1 and 400 have the effect size of -1. The simulated, originally observed and corrected effect sizes are marked with black, blue and red colour. The plots are based on one round of simulation.

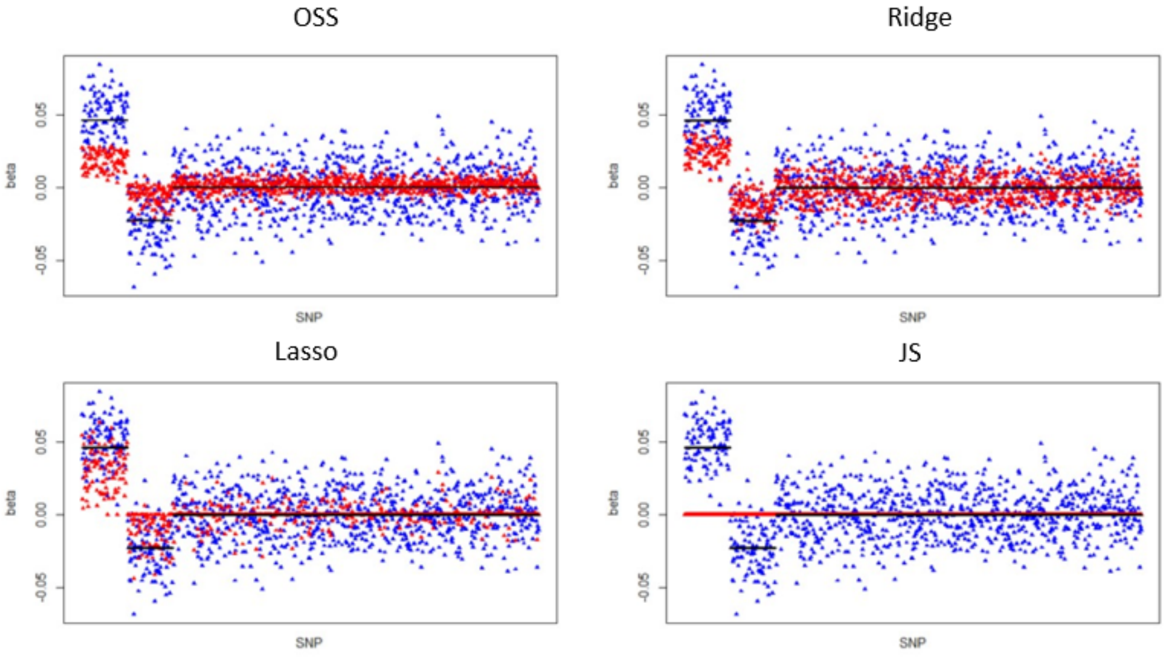


Figure 30 The performance of shrinkage methods in the scenario in which 500 SNPs have the effect sizes of 1 and 500 have the effect size of -1. The simulated, originally observed and corrected effect sizes are marked with black, blue and red colour. The plots are based on one round of simulation.

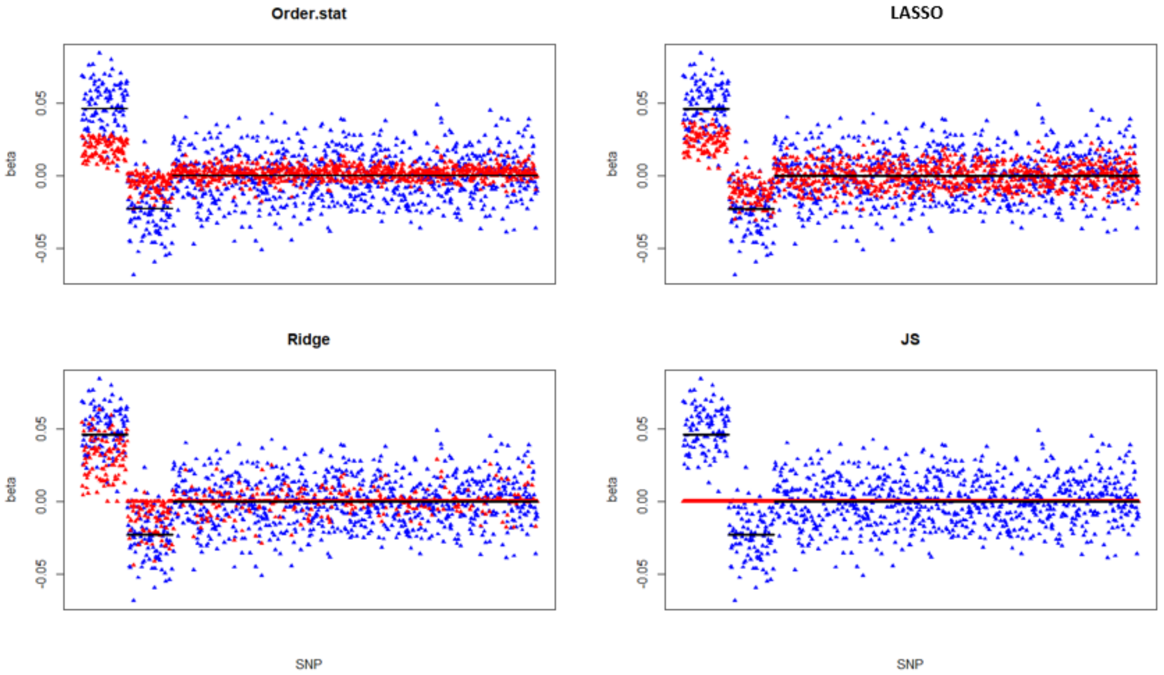


Figure 31 The performance of shrinkage methods in the scenario in which 100 SNPs have the effect sizes of 2 and 100 have the effect size of -1. The simulated, originally observed and corrected effect sizes are marked with black, blue and red colour. The plots are based on one round of simulation.

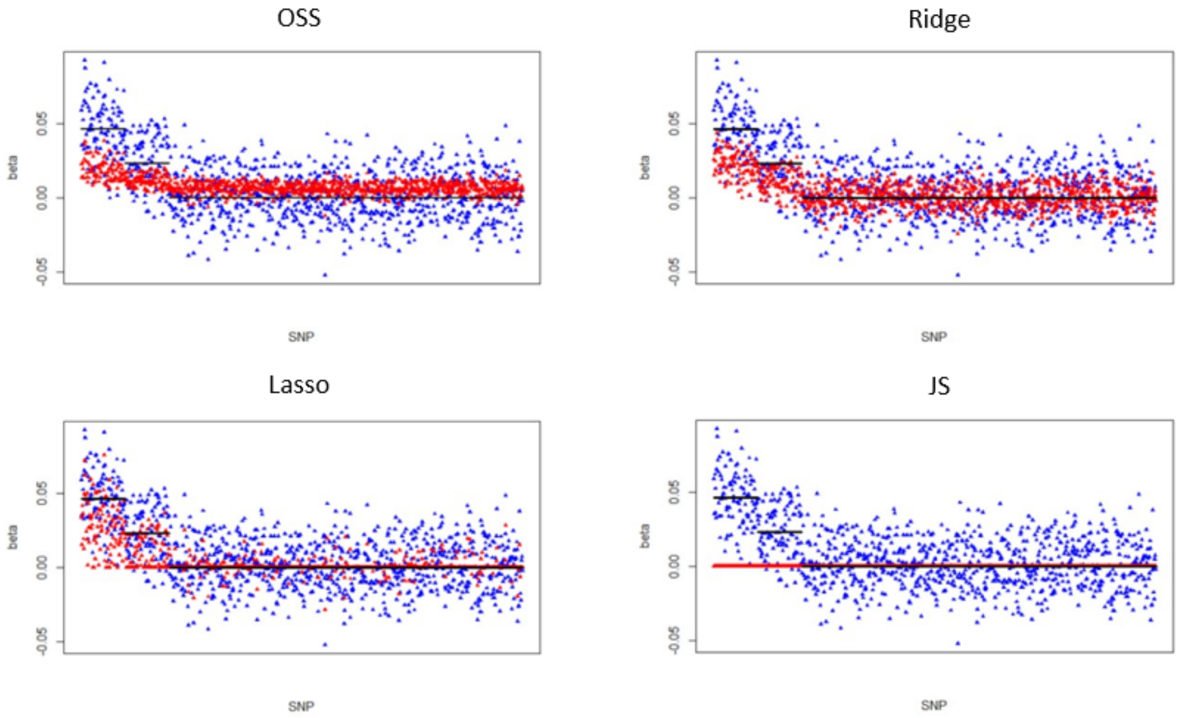


Figure 32 The performance of shrinkage methods in the scenario in which 100 SNPs have the effect sizes of 2 and 100 have the effect size of 1. The simulated, originally observed and corrected effect sizes are marked with black, blue and red colour. The plots are based on one round of simulation.

To summarise, the smaller the percentage of causal SNPs (or non-genotyped causal variants), the more the scenario will be in accordance with the underlying assumptions of the PS and OSS approaches. While these simulations did not include important aspects of real data, such as variation in MAF between SNPs, LD among SNPs and correlation between causal SNPs, which may improve the performance of more complex approaches such as Lasso and ridge regression, they also did not account for the potentially larger sample size available to the OSS approach in practical settings where summary statistics are more abundantly available than individual-level data, which is required for the standard implementations of Lasso and ridge regression tested here. Besides, the running time for correcting the simulated data of Lasso and ridge regression in R is approximately 40 times and 70 times more than that of OSS in out test. The running time depends on various factors such as the coding language that implements the algorithm, CPU, the size of data, etc. but it is clear that OSS is very computationally efficient. In conclusion, these simulations show that OSS has the potential to perform similarly as more complicated shrinkage methods, such as lasso and ridge regression, when correcting GWAS data results and thus may provide a useful alternative depending on data availability and time constraints given its computational efficiency.

3.3.3. Comparison of OSS and PS for generating the null distribution in real GWAS data

As in section 3.3.1, here we compare the generation of the null distribution using the Order Statistics Shrinkage (OSS) method with that based on the Permutation Shrinkage (PS) method (chapter 2). As described previously, the main difference between PS and OSS is how the two methods generate the null effect size distribution. If the null distributions generated by the two methods are similar, then their performance should be similar too.

There are different options to implement the OSS method, which we compare here. Firstly, in calculating the standard error (SE) of expected effect sizes, the MAFs of SNPs within each 1% MAF bin can be either treated as a having a fixed value corresponding to the middle point of the 1% interval ("fixed-SE") or as a random variable based on a random permutation of SEs calculated based the exact set of MAFs within each bin ("random-SE"). Secondly, in terms of the LD structure within the MAF bins, this can be either ignored so that all SNPs in the MAF bin are corrected as if they were independent of each other ("all-SNPs"), or taken in to consideration whereby the SNPs are first clumped and then corrected ("clumped").

Since under the "clumped" option, the number of SNPs to be corrected is smaller than the original number of SNPs, and the PS method corrects all the SNPs, there is no exact like-for-like comparison that can be made between the two methods. Therefore, only the null distribution generated by "all SNPs" OSS was compared with that generated by PS here, but we expect that the general findings should generalise to the "clumped" setting.

The pulse rate trait (estimated heritability=0.18) was used to illustrate the difference between the null distribution generated by OSS and PS (Figure 33 and Figure 34). The null distribution generated by the two methods were generally similar: almost all the data points were on the line $y=x$. For SNPs with MAF<4% the null effect size generated by "fixed SE" OSS were smaller than that generated by PS as shown in Figure 33 the data point is slightly below $y=x$. Considering the fact that that the effect sizes of the low MAF SNPs have higher variance, due to their larger standard error (corresponding to SE being a function of 1 / MAF(1-MAF) as given by eqn. 2 in 3.2.2), "fixed SE" might no be able to fully represent the deviation of the effect size distribution of low MAF SNPs.

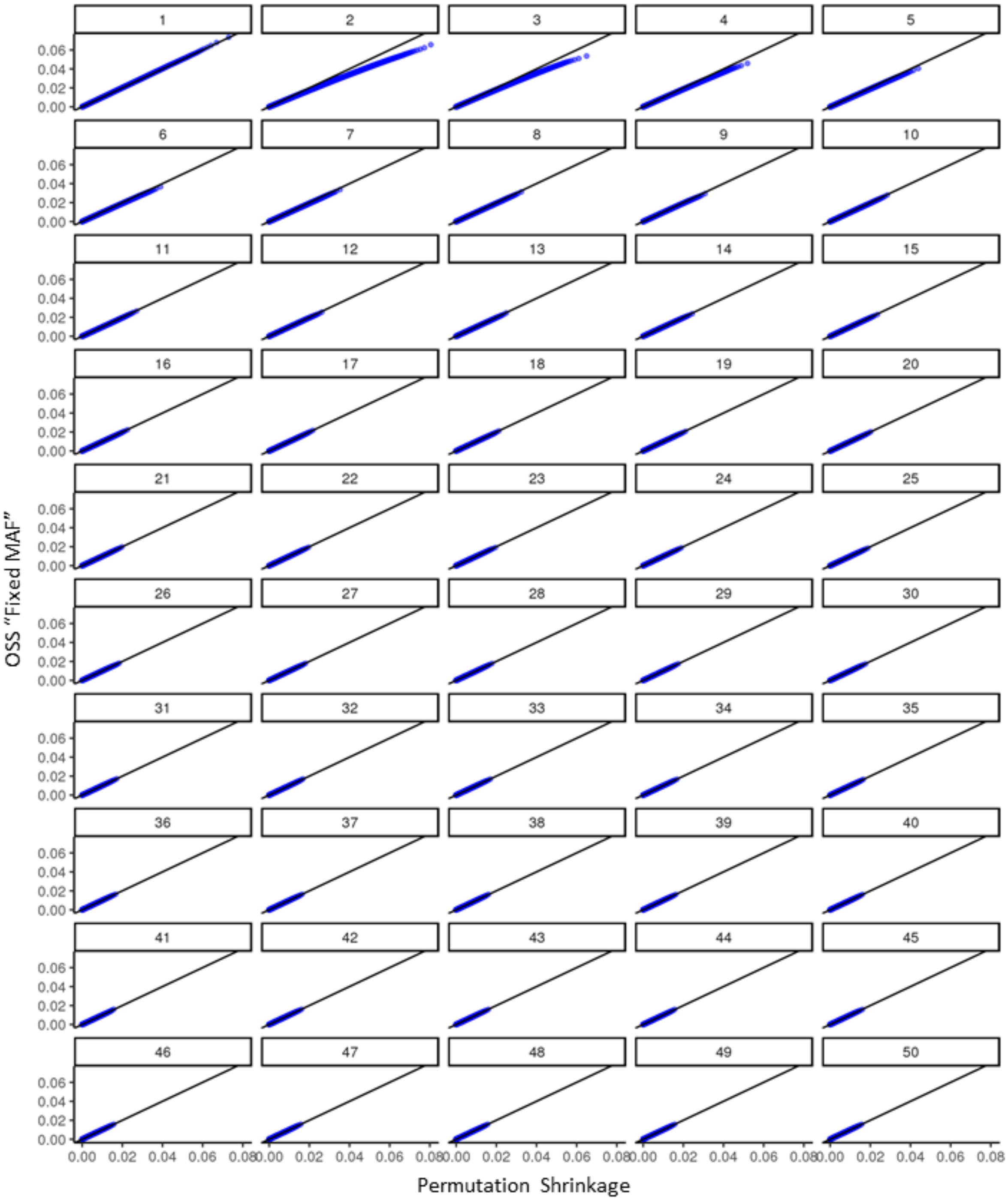


Figure 33 The null distribution generated by "All SNPs-fixed SE" OSS and PS with the same set of the UK Biobank QC'ed genotype data and standardized null quantitative data. PS: Permutation Shrinkage, OSS: "All SNPs-fixed SE" " Order statistics Shrinkage. The header of each subplot indicates the MAF bin, e.g. "1" means the MAF bin of 0-1%, 2 means the MAF bin of 1-2%

Figure 34, however, illustrates a smaller difference between the OSS and PS estimates, thus indicating the greater accuracy in using the exact distribution of MAF within each 1% MAF bin rather than approximating the MAFs in each bin by their mid-point value. Due to these results, we expect that the “random-SE” option for the OSS method will perform better than the “fixed-SE” option, since this more closely replicates the more exact procedure of the PS approach for generating the null distribution of effect size estimates.

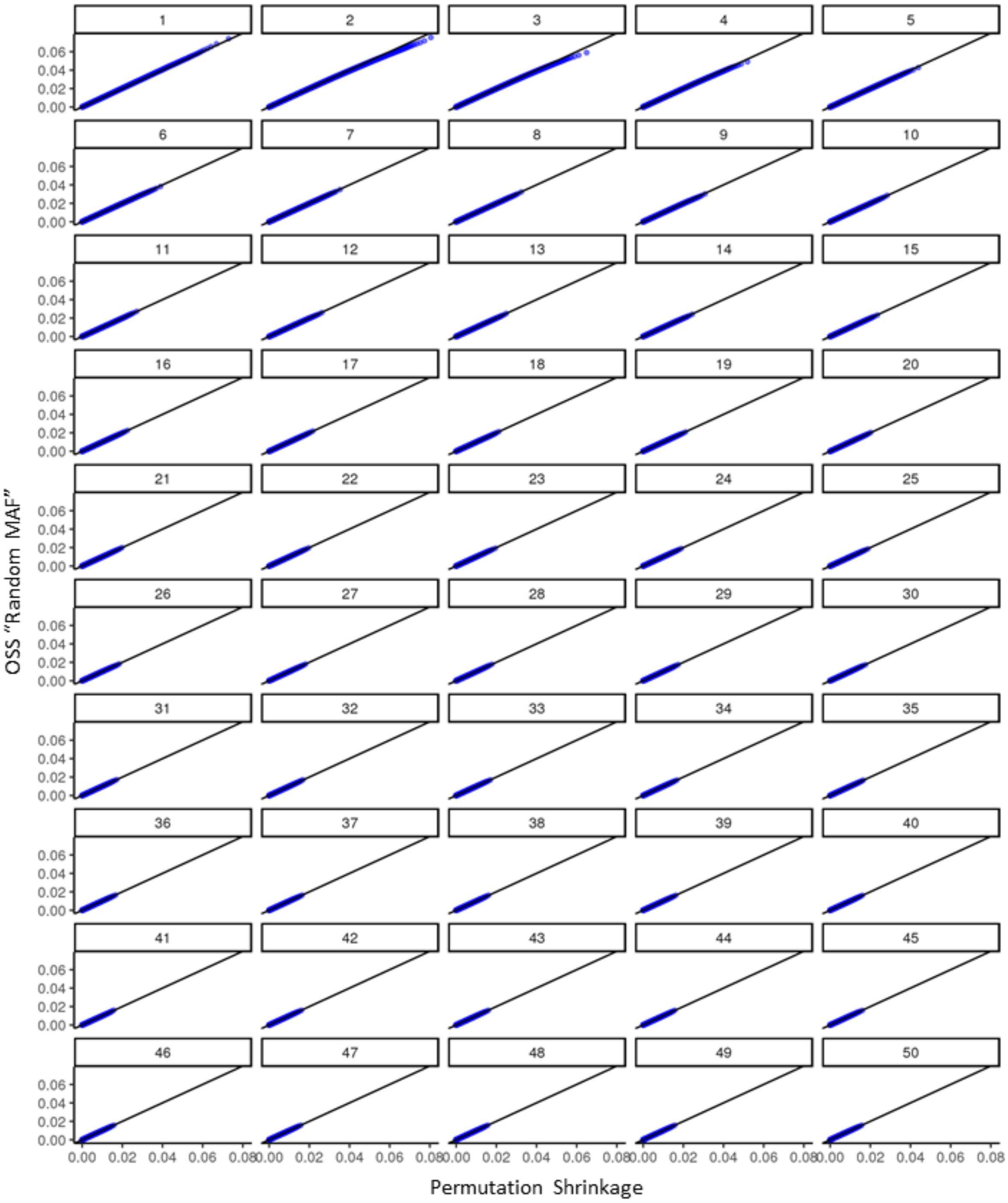


Figure 34 The null distribution generated by "All SNPs random SE" OSS and PS with the same set of UK Biobank QC’ed genotype data and standardized null quantitative data. PS: Permutation Shrinkage, OSs: "All SNPs random SE" Order statistics Shrinkage. The header of each subplot indicates the MAF bin, e.g. “1” means the MAF bin of 0-1%.

3.3.4. Testing performance of Order statistics Shrinkage on increasing PRS prediction with real data

Similar with PS, the performance of OSS was tested with a 3-dataset system using UK Biobank data (see section 3.2.5 and 2.3). The discovery data was used as the base data for PRS and corrected by Order Statistics Shrinkage. The testing data was then used as the target data for building a PRS model with *P*-value threshold optimisation. Then the optimised model was validated with the validation data using the *P*-value threshold used in the testing data model. The relative increase of PRS R^2 in the validation data after correction indicated how much the shrinkage method increase the power of PRS.

Similar with what happened in the tests of PS, when the estimated heritability <0.05, the relative increase of PRS R^2 were severely fluctuated since the PRS R^2 were around zero before and after the correction. Therefore, the results of these less heritable traits were omitted here. The performance of OSS was tested when the estimated heritability >0.05.

When tested with UK Biobank data, the way of processing the MAF variation with in a MAF bin (see "Generating the order statistics of absolute effect size" under section 3.2.2), made a significant difference to the performance of OSS (Figure 35). The "random SE" improve the PRS prediction much more than "fixed SE" as shown in Figure 35. The mean PRS R^2 relative increase of all the traits caused by "fixed SE" OSS is 5.3%, while the mean relative increase caused by "random SE" is 35%, which is close to the performance of PS (35%). The null distribution generated by "fixed SE" and "random SE" were similar with that generated by PS expect that for SNPs of MAF<4%, the null distribution generated by "fixed SE" is smaller than that generated by PS. The difference in correcting low MAF SNPs was the most likely explanation for the different performance of increasing PRS prediction, which indicated that low MAF SNPs might play an important role in predicting the phenotype.

The way of processing the LD structure (see "Estimating the effective number of independent tests" under section 3.2.2) would influence the OSS performance as well. As shown in Figure 36, for UK Biobank data, "all SNPs" mode still performed better than "clumped SNPs" mode. The mean PRS R^2 of the "all SNPs" 0.352 and the "clumped SNPs" 0.214. The "clumped SNPs" performed much worse than "all SNPs" and PS. There were 2 possible reasons for this: first, only the top SNPs were chosen in clumping, so expectation of clumped SNPs might not be zero, but in the following shrinkage the expectation of the SNPs were still assumed to be zero; second, in “clumped SNPs” option, the SNPs were clumped twice: when correcting the effect size estimates, and when calculating PS. The two tiers of clumping might lead to the loss of informative SNPs. Although “all SNPs” also had the risk of violating the assumption that all the SNPs in the 1% MAF bins were independent, at least in out tests with UK Biobank genotype data, the influence of violating the assumption of independent SNPs was low.

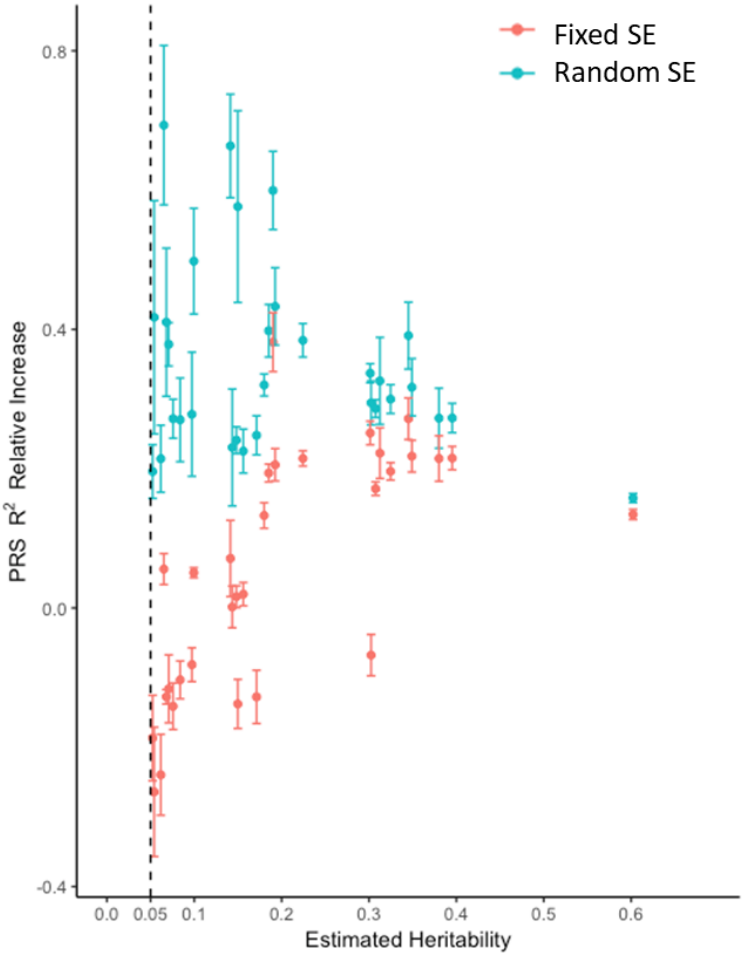


Figure 35 Comparison of PRS R^2 relative increase generated by different modes for processing the MAF. The dot shows the mean value of the relative increase and the error bar shows the 95% confident interval. Each data point shows the PRS R^2 relative increase of one UK Biobank quantitiative trait. The colour of the data point indicates the mode for porocessing the MAF. The mean value and error bar are based on the results of 5 repetition. "Random SE" OSS improved PRS prediction power much more than "Fixed SE" OSS. In this plot, only the result of “all SNPs” are shown.

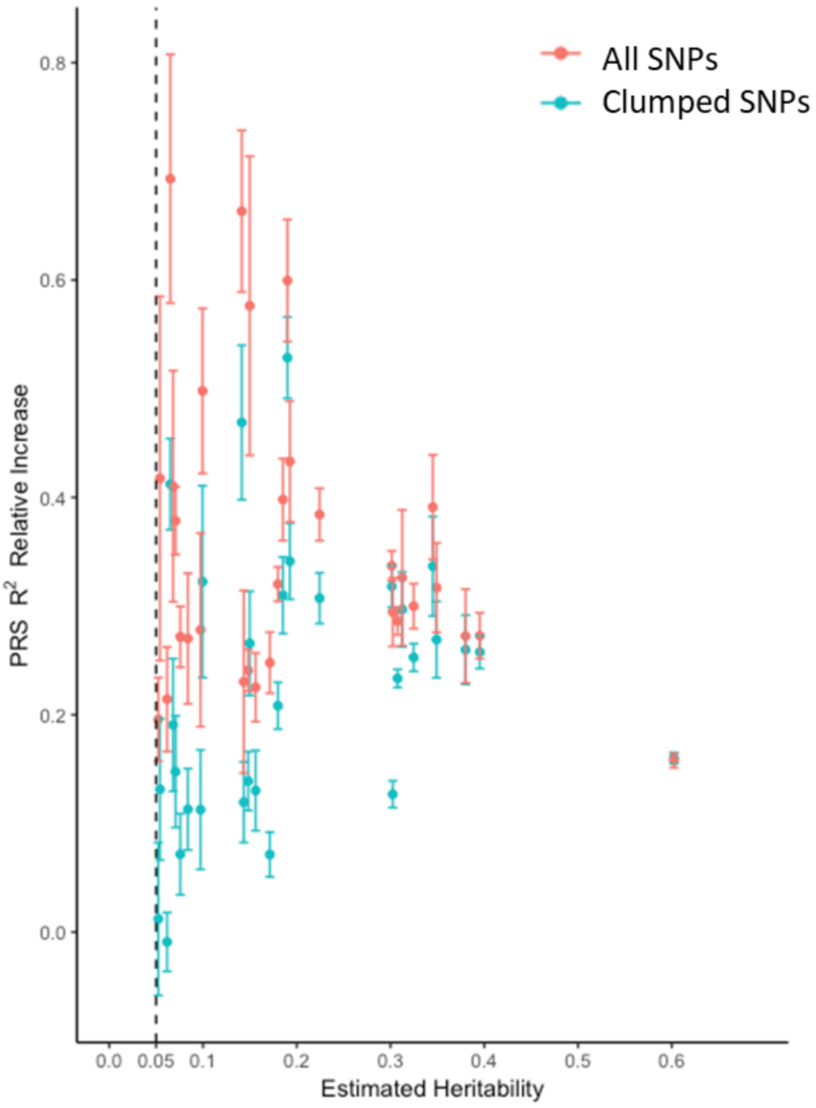


Figure 36. Comparison of PRS R^2 relative increase generated by different modes for processing LD. The dot shows the mean value of the relative increase and the error bar shows the 95% confident interval. Each data point shows the PRS R^2 relative increase of one UK Biobank quantitiative trait. The colour of the data point indicates the mode for porocessing LD. The mean value and error bar are based on the results of 5 repetition. "All SNPs" had better performance than "clumped SNPs" OSS. In this plot, only the result of "random SE" are shown.

We recommend that when correcting UK Biobank GWAS data or GWAS data of similar SNP density, the most optimized option is combination of "random SE" and "all SNPs". The PRS prediction relative increased caused by OSS "random SE - all SNPs" and PS were highly similar as shown in Figure 37. The correlation between the mean increase caused by the 2 methods is $r^2 = 0.97$. This result shows that OSS "random SE - all SNPs" can be used as an alternative method of PS with almost equal power.

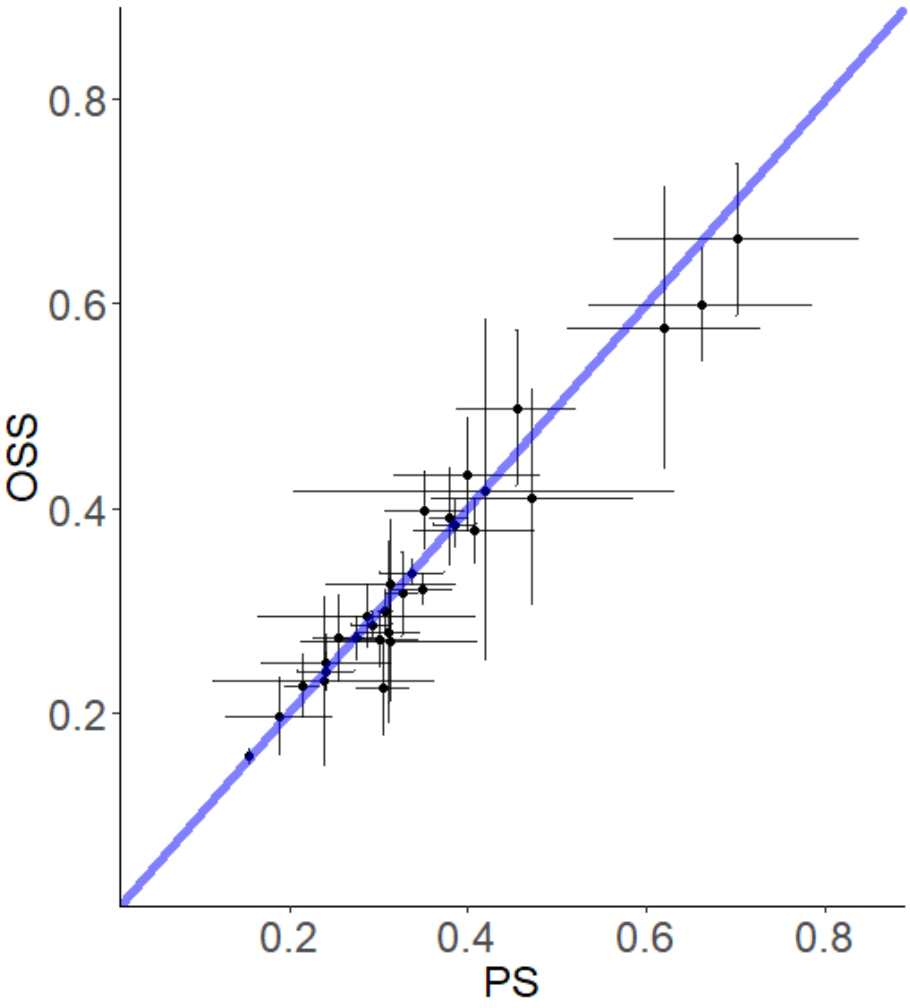


Figure 37 Comparing the PRS relative increase caused by PS and OSS "random SE - all SNPs". The dot shows the mean value of the relative increase and the error bar shows the 95% confident interval. The mean and error bar are based on the results of 5 repetitions of both methods.

In addition, more SNPs were included in the optimized PRS model after the correction of effect size estimates, as shown in Figure 38. Since the performance of Order Statistics Shrinkage and Permutation Shrinkage were highly similar, only one method, i.e. Order Statistics Shrinkage of “random SE – all SNPs” was used as an example here.

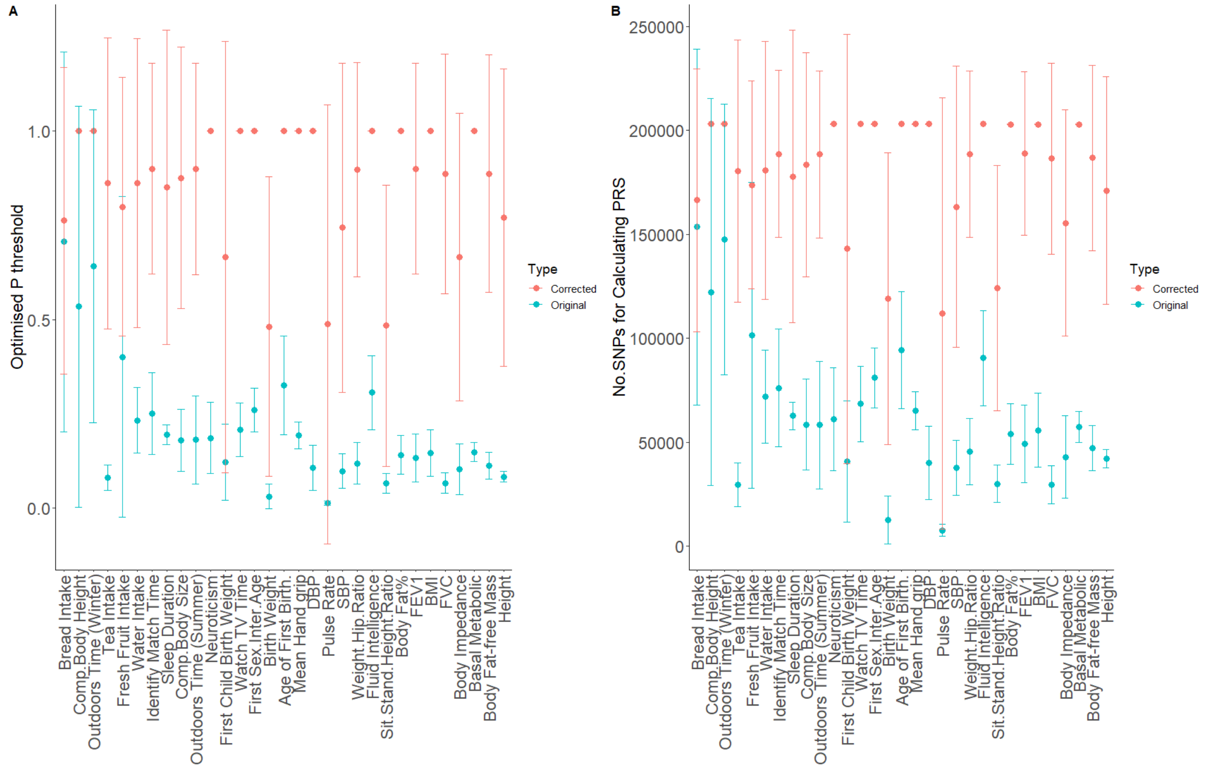


Figure 38 Comparing the optimised P-value threshold and the number of SNPs when using the original and corrected SNP effect size estimates of UK Biobank data. The SNP effect size estimates were corrected with Order Statistics Shrinkage of "random SE - all SNPs" mode. The phenotypes were ranked ascendingly from left to right according to their estimated heritability. The optimised P-value thresholds (A) and number of SNPs for calculating PRS were based on 5 repeats and shown in different colours. The dot shows the mean value of the relative increase and the error bar shows the 95% confident interval The plots shows that after correction more SNPs were included into the optimised PRS model.

3.3.5. Testing performance of Order statistics Shrinkage on increasing PRS prediction with simulated data

The similarity of PS results and OSS “random SE - all SNPs” results was repeated in simulated data as shown in Figure 39. The performance of “All SNPs” - “random SE” OSS were tested with simulated data based on the unstandardized genotype. In the simulation, OSS significantly increased the PRS prediction power when the traits were modestly heritable. The increase flatted out for traits with higher estimated heritability. The pattern of the result is highly similar with the result of PS shown in 2.4.4.

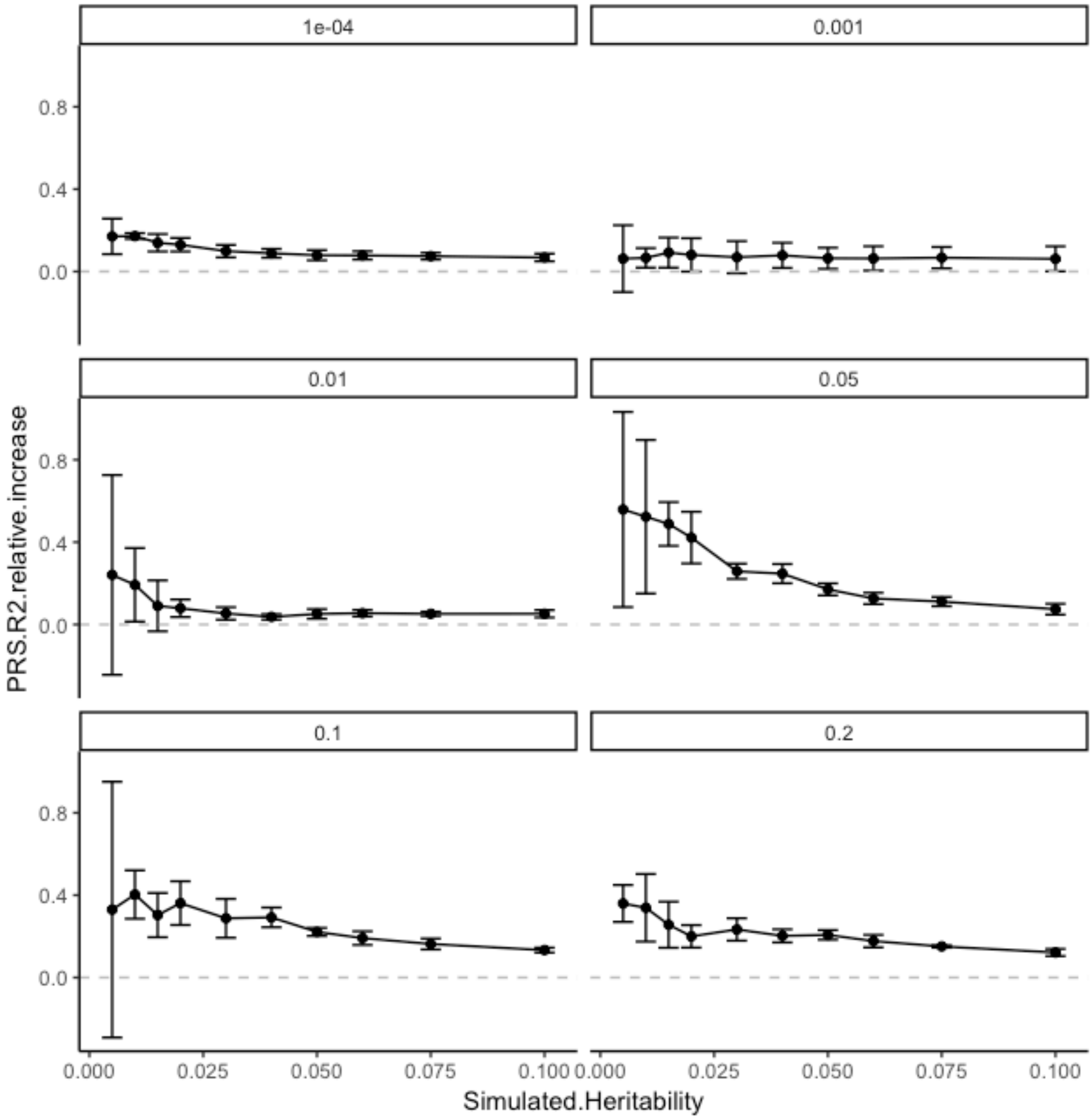


Figure 39 "Random SE-clumped" OSS improved the PRS performance in simulated data. The x-axis shows the simulated heritability and the y-axix shows the relative increase of PRS R^2 after the base data is corrected by "Random SE-clumped" OSS methods. The simulated data were based on Chromosome 1 of 20k QC'ed UK Biobank samples, so the heritability scale was approximately 8% of the genome-wide data. That is, a simulated heritability of 0.1 above corresponds to a real data heritability of approximately 80%. The phenotypes were simulated based on unstandardized genotype as mentioned in section 2.3.5. The dot shows the mean value of the relative increase and the error bar shows the 95% confident interval calculated from 5 repeats.

The number of SNPs included in the optimised PRS were compared before and after the correction of SNP effect size estimates when tested with simulated data. Only the result Order Statistics Shrinkage "random SE – all SNPs" mode was shown here because of the similarity of the shrinkage methods developed in this thesis. The increase of included SNPs was not as obvious as in the real data. The probably reasons are that some of the simulated scenarios were much less similar with a typical real trait and that simulation did not capture all the feature of real data.

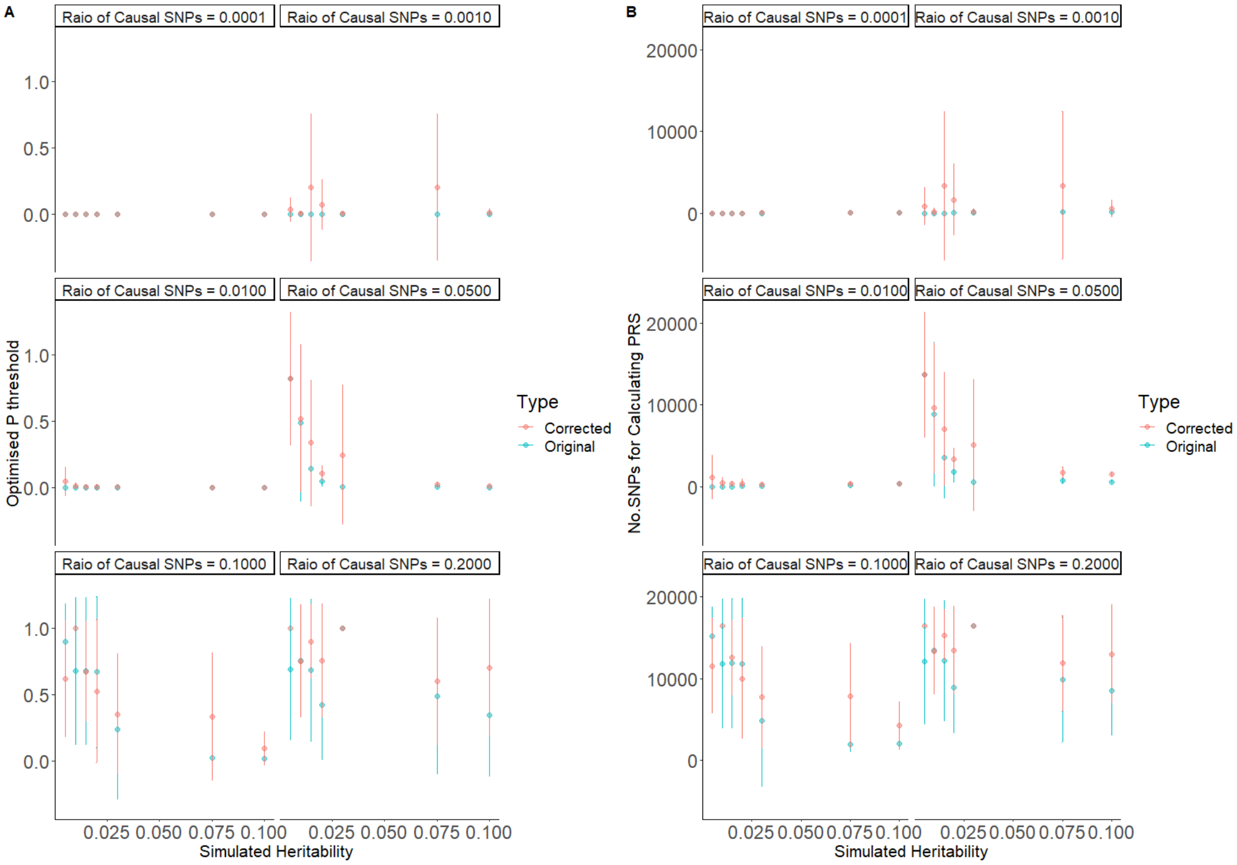


Figure 40 Comparing the optimised P-value threshold and the number of SNPs when using the original and corrected SNP effect size estimates of simulated data. The SNP effect size estimates were corrected with Order Statistics Shrinkage of "random SE - all SNPs" mode. The optimised P-value thresholds (A) and number of SNPs for calculating PRS were based on 5 repeats and shown in different colours. The increase of number of included SNPs in the optimised PRS is not as much as the that in real data, while the tendency is obvious yet. The dot shows the mean value of the relative increase and the error bar shows the 95% confident interval calculated from 5 repeats

3.4.Discussion

In this chapter, an alternative method for Permutation Shrinkage (PS) was developed. In our tests, Order statistics Shrinkage (OSS) can perform similarly with PS even when only summary statistics MAF and effect size calculated from standardised outcome trait are available.

This work not only provides a way of improving PRS prediction, but also reveals some statistical nature of effect size estimates in GWAS. That “Random SE” method performed much better than “Fixed SE” indicated that the when estimating the null distribution of effect sizes the differences in MAF should be taken into consideration even if the SNPs are binned in to small MAF intervals. This finding may help the future analytic work involved in the estimation of the null distribution of effect size.

The work for now has not fully solve the problem of processing the LD structure. It can provide a corrected effect size for every SNP using “all SNPs” mode only when using relative sparse GWAS data like UK Biobank. When correcting a denser dataset, i.e. imputed GWAS data, it is possible to use the “clumped” mode, which only provided an improved PRS prediction. However, in our test with UK Biobank data, the increase PRS prediction caused by “clumped” is less than that of “all SNPs”. A possible reason for this is that the first round of clumping in the “clumped” mode may delete some SNPs that may turn out to be informative in the PRS model. Two issues needed to be investigated to apply this method to more dense data: how to efficiently calculate the effective number of independent tests in a large-scale data, and how to extrapolate the null distribution according to the LD structure.

If the aim is to improve the GWAS result, the effect sizes of all the SNPs should be corrected. The most mathematically rigorous method is to first estimates the effective number of independent tests M_e of the base data, then generate the null distribution and finally extrapolate the distribution to all the SNPs. However, it is very complicated to implement this rigorous method.

Mathematically, M_e should be calculation by eigen value decomposition of the genotype correlation matrix[79,80]. However, the raw genotype data is not available because when using OSS, we assumed that only summary statistics data is available. Even if we can use the target genotype data as an proxy of the raw genotype data of the base data, the complexity of eigen value decomposition is $O(n^3)$ and it can be too computational expensive to calculate especially for large-scale dataset as UK Biobank.

A possible option for calculating M_e is to use the number of clumped/pruned SNPs as a proxy. In this case, the correction process can be first correct the clumped data and then extrapolate the data according to the LD structure and MAF.

Admittedly, the estimation of M_e and extrapolation of the corrected null distribution is important for the further improvement of the OSS method and many other methods in statistical genetics. However, the two questions are beyond the focus of this project, that is, to improve the performance of PRS. Besides, the method of estimating M_e and extrapolating the null distribution should be tested in an extra imputed data and extra simulated highly colinear data. Therefore, in this chapter, only "all SNPs" option and "clumped" option are tested.

Chapter 4. PRSet: Polygenic Risk Score Gene Set Analysis Method

4.1.Introduction

It has now been established that complex traits are influenced by hundreds or thousands of genetic variants, each of which only has a small effect[81–83]. However, polygenicity does not imply that all genetic variants or regions in the genome contribute equally or independently to the trait. It may be that individuals get a disease because they have a particularly high burden of risk alleles across a specific biological pathway, even if their risk across the rest of the genome is low. That is, it may be that genetic risk converges across functional groups in the genome, and that rather than being additive it instead involves high order interactions across many genetic variants and separate liabilities of risk.

Genes can be grouped into gene sets according to aetiological or biological factors. Identifying gene sets that contribute more to phenotypes of interest than other gene sets could help to reveal disease or trait aetiology[84,85] and stratification of risk, provide potential drug targets, and lead to better patient stratification[86]. Gene Set Analysis (GSA), or sometimes interchangeably called 'pathway analysis', exploit genetics and other 'omics' data, such as expression profile[87], proteome[88] and reactome[89], to gain insights into the aetiology and genetic basis of complex disease (see section 1.9). GWAS-based GSA methods[62,90] such as GSEA[59], FORGE[64], MAGMA[65], have been developed and most of them utilise the *P*-values of the SNP-phenotype associations and do not exploit or provide any individual-level information about the genetic burden enriched in the gene set.

Polygenic risk scores (PRS) have been widely-used in recent years to capture polygenic signal across the genome[52] (see section 1.6), but they are yet to be fully exploited as part of GSA methods[91]. If PRS-based GSA methods can have better or even comparable performance to the existing approaches in terms of assessing and ranking the enrichment of signal across different pathways, then they could have extremely high utility given that they can provide individual-level gene-set estimates of genetic propensity to phenotypes as well. With PRS for each gene-set in each individual, researchers may be able to identify specific sources of shared aetiology among different traits, identify reasons for differential treatment response and stratify cases of a disease into more homogeneous groups in terms of shared aetiology.

A group of PRS-based GSA methods that we collectively call "PRSet" is developed and tested in this chapter. The PRSet software has been incorporated into PRSice[77], the PRS analysis software suite that was developed by my research group, as a sub-function. The coding for PRSet was written both by myself and my supervisor Dr Choi. PRSet performs two types of gene set analysis: 1) self-contained: to test whether the gene set has an association with the phenotype; and 2) competitive: to test whether the gene set is any more associated with the phenotype under study than a random gene set from the genome with the same fundamental properties[62].

4.2.Methods

4.2.1. Calculation gene set PRS

Gene set PRS PRS_S is used to represent the genetic burden in the gene set. In order to develop PRS-based GSA methods, the first step is to calculate PRS_S. There are several technical aspects to be considered:

Definition of gene set SNPs: SNPs are defined as belonging to the gene if they fall into the range of the gene. The flanking area of the gene contains the regulatory sequence such as enhancer, silencer, promoter, 5' and 3' UTR, they also influence the expression of the gene, researchers may include the flanking region as part of the gene according to the need of the specific research. In PRSet, the range of a gene is the exon, intron and the flanking region whose size is defined by the user. All the genic regions that belong to the gene set are merged to construct the gene set region. If a SNPs belongs to more than one gene, it is calculated only once when constructing the PRS_S.

Clumping of the gene set SNPs: In calculating genome-wide PRS, clumping is used to extract independent signals across the whole genome. However, when calculating PRS_S, the SNPs inside the gene set may be clumped out because of a neighbouring SNPs outside the gene set if the clumping is performed across the genome. Therefore, the gene set SNPs are clumped only within the range of the gene set, i.e. to analyse k gene sets, k round of clumping within the gene set will be performed and each gene set is effectively treated as the whole genome for clumping purpose. Thus, the index SNPs outside the gene set will not cause the deletion of possible informative SNPs inside the gene set. Another advantage of clumping within the gene set SNPs is that the LD in each gene set will be naturally dealt with. After clumping, the gene set are treated as a group of independent SNPs. Even the genes in the gene sets are correlated with other genes, there is no need to estimate the correlation and LD of these genes like in MAGMA. However, independently performing clumping on each individual gene set is a computationally intensive process. To speed up the set-based clumping, we utilize the bit-flag system: when an index SNP clumped a target SNP, the bit-flag of the target SNP will be updated using the combination of bitwise AND and XOR operation such that the index SNP will "remove" gene set membership from the target SNP if and only if they fall within the same gene set (Figure 41).

	Set A	Set B	Set C	Set D
SNP 1	1	0	1	1
SNP 2	0	0	1	1
SNP 3	1	1	0	1

→

	Set A	Set B	Set C	Set D
SNP 1	1	0	1	1
SNP 2	0	0	0	0
SNP 3	0	1	0	0

Figure 41 Illustration of bit operation involved in PRSet clumping. Left: If a SNP is found in a gene set, it will be marked with 1, and 0 otherwise. Right: Assuming SNP 1 is the index SNP, it will "remove" set memberships from other SNPs that were clumped by.

***P*-value thresholding:** In a typical genome-wide PRS analysis, we usually aim to test the association between the genome PRS and the phenotype to test the existence of a polygenic basis or a genetic association between two phenotypes. To maximise the statistical power to detect the association between PRS and the target phenotype, a series of PRS are usually calculated using different *P*-value thresholds to get the most optimised model[52,77]. However, the aim of GSA studies is to represent the association between the gene set and the phenotype without bias and every gene set should be tested in the same way. Therefore, it will be inappropriate to calculate gene set PRS with optimizing the *P*-value threshold for each gene set like calculating the typical genome-wide PRS. Besides, after optimization for SNP *P*-value threshold, only part of clumped SNPs are included in the PRS. It is difficult to justify whether this part of SNPs is a fair representative of the whole gene set. Therefore, I calculated gene set PRS without optimising *P*-value threshold. All the clumped SNPs are included in the PRS.

To sum up, PRS_S is the weighted sum of risk alleles of all the clumped SNPs and the clumping is performed only within the gene set. PRS_S represent the genetic burden that enriched in the gene set.

4.2.2. PRSet methods

1) Self-contained test

The null hypothesis of self-contained GSA methods is that the gene set has no overall association with the phenotype. It is the most basic gene set test in which the representative of the genetic burden in the gene set is calculated to test its association with the trait.

According to the null hypothesis, self-contained test in PRSet is simply to test the association between the phenotype and the gene set PRS_S controlling for covariates such as age, sex, top principal components (PCs):

$$Phenotype \sim PRS_S + covariates + \varepsilon$$

in which ε is the residuals

2) Competitive test

The null hypothesis of competitive test is “The genes in a gene set is not more association with the phenotype than the genes not in that gene set”. An equivalent expression is that “the gene set is not more association with phenotype than the random gene sets of the similar features.” For polygenic traits, any region in the genome, even those of no biological importance, can be self-contained significant because the signal can be distributed along the whole genome.

The competitive test can be implemented in two ways: First, the statistics of the genes inside the gene set are to be compared with those of the genes outside the gene set. In PRS-based GSA, this implementation is to compare the PRS of genes inside the gene set with those outside the gene set. The framework of MAGMA-geno is borrowed to implement this approach (marked as “PRSet-MAGMA.like”); Second, the statistic of the gene set is to be compared with those of random gene sets from the genome. In PRS-based GSA, this implementation is to compare the PRS of the observed gene set with those of random gene sets of the similar properties. This approach was marked as “PRSet-perm”.

PRSet-MAGMA.like method:

This method compares the PRS of genes inside the gene set with the PRS of the genes from the outside. It uses the 2-tier framework of MAGMA’s competitive method (see section 1.9.5):

In the first tier, the gene-phenotype association is calculated. PRSet-MAGMA.like first calculates the PRS of each gene. The calculation of gene PRS is similar with calculation of gene set PRS. The PRS of gene *g* (PRS_g) is the weighted sum of risk alleles of clumped SNPs belonging to the gene *g*. The SNPs were clumped only within the range of the gene. The phenotype is then regressed on the PRS of each gene:

$$Phenotype \sim PRS_g + covariates + \varepsilon$$

in which covariates can be individual-level data such as sex, age, top genome principal components, etc. and ε is the residue.

The *P*-value of this regression model p_g is converted into one-sided Z score $z_g = \Phi^{-1}(1 - p_g)$, in which Φ^{-1} is the probit function. The Z score z_g represents the association between the gene *g* and the phenotype.

In the second tier, PRSet-MAGMA.like compares the phenotype-gene association inside and outside the gene set with a regression model of the gene information: the phenotype-gene association is regressed on an indicator of whether the gene is in the gene set or not, controlling for the covariates of the gene, such as the gene size, gene density and minor allele account (MAC) of the SNPs in the gene and corrected for the correlations between the genes:

$$Z \sim S_s + gene\ size + gene\ density + MAC + \ \varepsilon$$

where vector Z consists of the Z scores z_g of the genes. Vector S_s consists of element that denotes the whether the gene g is in the gene set s: element s_g =1 if gene g is in the gene set; else s_g = 0.

The correlation between the genes are corrected for in the error term ε. The statistics of the genes can be correlated because these genes share SNPs or have SNPs in LD. This breaks the assumption of standard linear regression that the error terms should be independent. MAGMA provide a solution for this problem by adding a generalized least squares error term that takes the correlations of the gene into account and PRSet-MAGMA.like borrows this term:

$$\varepsilon \sim \mathrm{MNV}(0, \sigma^2 R)$$

in which R is the matrix of gene-gene correlations.

The covariates used by MAGMA is gene size, gene density, minor allele account (MAC). Gene density in MAGMA-geno is the number of pruned PCs divided by the number of SNPs in the gene. The counterpart of gene density in PRSet-MAGMA.like is the number of clumped SNPs divided by the number of raw SNPs in the gene. In the pilot tests the value of the two gene density is highly correlated and the meaning of the gene density is the ratio of independent signals versus the genotyped SNPs in the gene. Therefore, the gene density in the two approaches are exchangeable. Although the gene-gene correlation matrix R in MAGMA is calculated using gene PCs while theoretically the counterpart matrix in PRSet-MAGMA.like should be calculated using gene PC. Since developing the matrix R in PRSet-MAGMA.like is time-consuming and the matrix in the two methods might be very similar since they represent the same relationship. PRSet-MAGMA.like is implemented by plugging the Z scores of gene PRS from the first tier into the second tier of MAGMA flamework to get a quick estimate of what the power of PRSet-MAGMA.like could be.

Permutation method:

This method compares the observed PRS_S with the permuted null distribution and uses the empirical P-value as the competitive P-value. The null gene sets should be comparable with the tested gene set in terms of gene set size controlling for gene-gene correlation and LD structure.

As to constructing the null gene set PRS, randomly choosing the same number of raw genotyped SNPs or using SNPs from a region of the same number of base pairs can generate biased results because the LD structure and genes are not evenly distributed across the genome. A null gene set with the same number of raw genotyped SNPs or SNPs from region of the same number base pairs may have different amount of independent genetic signals and different LD structure.

Instead, PRSet-perm uses the same number of random clumped SNPs from genic region as a null gene set. The rationale is that PRS is actually the weighted sum of clumped SNPs. The clumped SNPs are independent or of negligible correlation. They represent the independent genetic signals free of LD structure.

To generate the null distributions for the observed gene sets, a gene set containing all the annotated genes is first constructed. The SNPs are then clumped within this gene set and the clumped SNPs are the pool for generating the null gene sets. To run the competitive analysis test for a gene set containing *n* clumped SNPs, *n* SNPs will be randomly drawn from this pool. This set of randomly chosen SNPs is a null gene set that contains the same amount of genetic information but was expected to not be enriched with association signals. Then PRS of this null gene set PRS_{null} is calculated and the phenotype is regressed on PRS_{null}. This process is repeated for multiple times to get the distribution of PRS_{null}. The empirical P-value is generated by comparing the observed PRS_S with the PRS_{null} distribution. This empirical P-value is the competitive P-value. In this chapter, the permutation number was 10^4. This number can be adjusted according to each specific analysis.

4.2.3. *Competitive* R^2

Since GSA methods based on permutations may have truncated P-value, an alternative method that overcomes the problem of truncated P-values is desirable. It is intuitive that the more important a gene set, the more per clumped SNP can explain the phenotype. Here, a statistic is introduced:

$$Competitive\ R^2 = \frac{R^2_{PRS\ gene\ set}}{Num(clumped\ SNPs)}$$

$Competitive\ R^2$ can have two possible application: First, neither competitive MAMGA nor competitive PRSet generated R^2, $Competitive\ R^2$ can be used as the counterpart of self-contained R^2 in competitive tests; Second, an analytical way of calculating competitive gene set association P-value can be developed based on $Competitive\ R^2$.

For the application of analytical association test, the rationale is very straightforward: if the $Competitive\ R^2$ is significantly higher than expectation, the gene set are significant in the competitive test. Although the $Competitive\ R^2$ is easy to define, the definition of null expectation and standard error is not as self-evident. There are two possible ways to calculate the competitive P-value: One way is almost the same as the permutation method: In order to test a gene set of n clumped SNPs, n clumped SNPs are randomly chosen from all the genic regions for multiple times and calculate the $Competitive\ R^2$ use the empirical P-value based on the $Competitive\ R^2$ distribution as the competitive P-value; The other way is to derive an analytical description of the distribution of $Competitive\ R^2$ empirical P-value based on.

Admittedly, the analytical description of $Competitive\ R^2$ distribution is difficult to derive due to the complex structure of the genome. However, an empirical estimation is possible: Assume that the clumped SNPs are independent, then according to the additive assumption of GWAS, we will have:

$$R^2_{gene\ set\ PRS} = \sum R^2_{clumped\ SNP}$$

Therefore, $Competitive\ R^2$ is the mean value of $R^2_{clumped\ SNP}$ in that gene set. Since $R^2_{clumped\ SNP}$ are independent, too. According to the central limit theorem, the mean value of independent random variables tends toward a normal distribution.

As a preliminary attempt to describe $Competitive\ R^2$ in an analytical way, I assume that: 1) given the number of clumped SNPs, the $Competitive\ R^2$ follows normal distribution; 2) $Competitive\ R^2$ of an "background" gene set that contains all the genic region, R^2_{genic} , is the expectation of null gene sets of different size; 3) the standard deviation of the $Competitive\ R^2$

is correlated with the number of clumped SNPs. sd($Competitive\ R^2$) can be described with a function containing the number of clumped SNPs.

In 4.3.4, the properties about $Competitive\ R^2$ were investigated, aiming to generate an analytical way of calculation competitive gene set association P-value.

4.2.4. Data

The PRSet methods were tested with UK Biobank data and simulated data. The raw UK Biobank genotype data were cleaned and quality-controlled as described in 2.3.4 and 20k cleaned samples were randomly chosen from the QC'ed samples for the tests in this chapter in order to ensure precise estimates of power differences between the methods requires as many permutations of simulated data as possible. We should expect the results to be qualitatively similar on a larger sample size. When these methods are performed in real data settings then they typically only need to be run once and thus real analyses can be scaled up to much larger sample sizes. Two types of simulated data were used: one was simulated based on the 20k UK Biobank samples and the other was simulated entirely in R.

Real data

Quantitative trait BMI (estimated h^2_{SNP} =0.32), Height (estimated h^2_{SNP}=0.60), Forced Vital Capacity (FVC) (estimated heritability=0.35), and Fluid Intelligence (Gf score) (estimated heritability=0.31) of the 20k QC'ed UK Biobank samples were used. As described in 2.3.4, the residual after regressing out the covariates (top 15 PCs, sex, age, assessment centre) and standardization was used as the phenotype. The heritability of these traits was estimated in 2.3.6.

The gene set information was downloaded from MSigDB[59]. Only the gene sets which contain 10-1000 genes were analysed. The definition of gene locations was downloaded from Gencode Release 19 (GRCh37.p13) (https://www.gencodegenes.org/releases/19.html). In this chapter, only SNPs in the range of gene were defined to belong to the gene. SNPs in the 5' or 3' flanking regions were not included.

When testing PRSet methods, the samples were randomly divided into 2 equal parts, and the two parts were used as base dataset and target dataset respectively. MAGMA (see section 1.9.5) have 2 options: using the raw genotype data (Marked as "MAGMA-geno") or using the summary statistics GWAS data (marked as "MAGMA-sum"). MAGMA-geno used the raw genotype data of all the 20k samples as the input; MAGMA-sum used the summary statistics GWAS result of all the 20k samples as the input. The permutation time for PRSet-perm was 10k.

Simulated data

UK Biobank simulation: In order to evaluate the statistical power of GSA methods here, phenotype data with known heritability and causal SNP distribution were simulated based on the QC'ed 20k UK Biobank samples mentioned in real data tests. Except using simulated phenotype data, the gene location and gene set annotation were the same as in the real data tests. The phenotypes were simulated from the UK biobank genome-wide genotype data using GCTA[76] when simulation based on standardized genotype data were modelled, and using the in-house method SGCP when simulation based on unstandardized genotype data were modelled (see section 2.3.5). A range of heritabilities, causal SNP percentages, as well as different effect size distributions and choice of causal gene sets and causal SNPs were tested, and the specific parameters are detailed in 4.2.4 where the tests are introduced. When testing PRSet methods, 10k of the QC'ed samples and their corresponding simulated phenotype data were used as base data and the summary statistics of another 10k samples as target data; when testing MAGAM-geno, the raw genotype data and simulated phenotype data of the 20k samples were used as input; when testing MAGMA-sum, the summary statistics of the 20k samples were used as input. The permutation time for PRSet-perm was 100k.

Simulation in R: Besides simulated data based on UK Biobank genotype data, data that were entirely simulated were used. These simulations were performed to investigate the impact of gene set size and causal SNP percentage on GSA result (see section 4.3.3). For this investigation, the following data were simulated in R: a "genome" containing 1000 independent genes, each of which contains one single SNP, was simulated. For all the SNPs, MAF=0.25; 20% of the SNPs were causal and had the same effect size of 1 that all together contribute to the heritability of 0.1. 10, 100, 250 SNPs were selected to form the gene set to be tested. 20% or 50% of them were randomly draw from the causal SNPs and the rest were draw from null SNPs. When analysing the simulated data with PRSet-perm (see section 4.2.2), the permutation was repeated 10k times.

I introduce a new statistic, *Competitive* R^2 (see section 4.2.3 and 4.3.4), and explore the possibility of developing an analytical GSA method based on *Competitive* R^2. Besides the UK Biobank data described above, the properties of the $Competitive\,R^2$ were tested with 'simulation in R' data as described above.

Critical factors in real GWAS-based gene set analysis, such as population structure, LD structure, are not included in the data simulated in R. However, even this basic simulation should reveal some of the properties of how the competitive GSA method behave and so act as a useful complement to the UK Biobank simulated and real data analyses.

4.3. Results

4.3.1. Self-contained test

Type 1 error rate test

Type 1 error is false positive, that is, the P-value is smaller than the threshold α under the null. Ideally, the type 1 error rate should be equal to the threshold α. The type 1 error rate is defined as below in this chapter:

$$Type\ 1\ error\ rate = \frac{N_{P-value<\alpha}}{N}$$

where N is the number of gene sets to be tested; $N_{P-value<\alpha}$ is the number of gene sets whose P-value pass the significance threshold α. The threshold for calculating type 1 error rate in this chapter was 0.005 and 0.05.

For the self-contained test, the null phenotype data should be irrelevant with the genotype. Permuted height of UK Biobank data was used as the null to test the self-contained PRSet and self-contained MAGMA. The simulation-testing process was repeated for 100 times. Table 4 shows that MAGMA-Geno and PRSet behaved well while MAGMA-sum was over-conservative. Type 1 error rate test showed that self-contained PRSet behaved as good as MAGM-geno. MAGMA-sum had less power probably because using summary statistics datamay cause the loss of information.

Table 4 Type 1 error of self-contained GSA methods

	Alpha=0.05	**Alpha=0.005**
MAGMA-geno	0.0488±0.0142	0.0048±0.0026
MAGMA-sum	0.0408±0.0124	0.0038±0.0023
PRSet	0.0503±0.0084	0.0052±0.0020

Testing on real data

The result of real data can be used as an approximate measure of the methods' power. The observed value can be viewed as the combination of the underlying true value and the error term. Assuming that the error terms of different methods are independent, a high correlation indicate that the true values contribute a high proportion of the observed values. Therefore, the correlation of two different measures indicate their power. The main statistics of both self-contained and competitive methods is the P-value. Therefore, we compare the -$\log_{10}$(P-value) of MAGMA and PRSet. In our previous test, MAGMA-geno behaved better than MAGMA-sum, so I compared the PRSet method with MAGMA-geno.

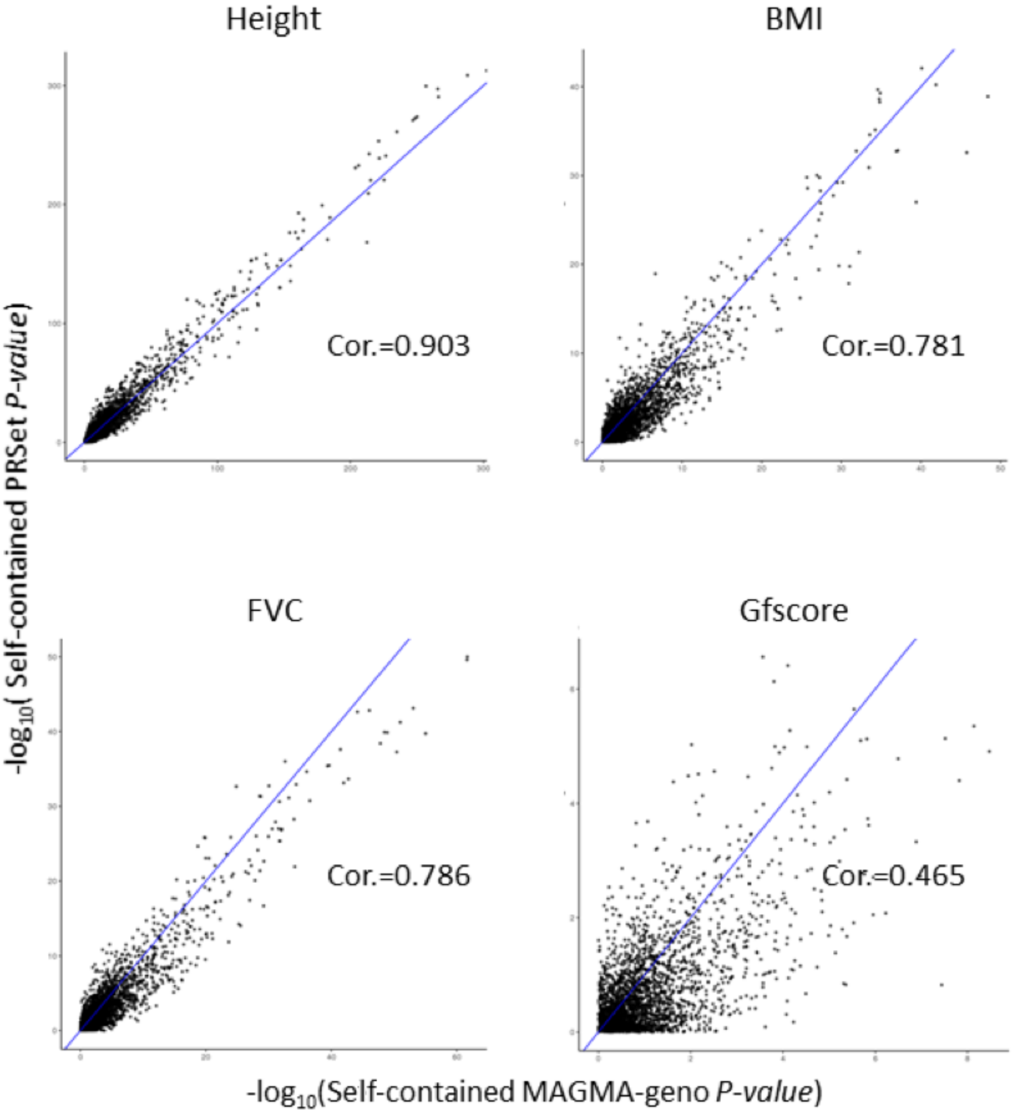


Figure 42 Comparison of self-contained MAGMA and PRSet result based on same sets of UK biobank data. Each data point indicates a real pathway in MSigDB. The x- and y-axis show the -$\log_{10}$(P-value) given by the two methods. Pearson Correlation Coefficient of -$\log_{10}$(P-value) of height, BMI, FVC and fluid intelligence (measured by Gf score) were calculated. The blue reference lines are y=x

The self-contained PRSet and MAGMA-geno results were highly correlated. The correlation of Height was highest. BMI, FVC and fluid intelligence have similar estimated heritability (0.32, 0.35 and 0.31 respectively), but the correlations of these traits were quite different.

The most heritable trait, height, had the highest correlation. This may be due to the fact that both methods have higher power when analysing a more heritable trait, and thus their true overlap in results is apparent for highly powered traits. However, the difference in correlations among BMI, FVC and fluid intelligence does not correspond well to the differences of their estimated heritabilities, since these were all similar to each other. Assuming that the heritability was correctly estimated, then the difference between the correlation in their results may be partially caused by the different genetic aetiologies of these traits. Thus, the performance of self-contained GSA methods may be sensitive to the genetic aetiology of traits.

However, the principal factor affecting the results of the self-contained methods is likely to be confounding by the size of the pathways, since this is not accounted for in the self-contained analyses. That is, large gene sets are likely to present strong overall associations with the phenotype because overall, they contain more true casual variants, on average, than small gene sets. This confounding is likely to be an important explanatory factor in the high correlations observed in the results between PRSet and MAGMA here. Therefore, more attention should be focused on the following analysis of competitive methods, since it is the results from *enrichment* of signal across gene sets above that expected from a random such gene set, rather than total signal, that most researchers in the field are most interested in.

4.3.2. Competitive tests

Type 1 error rate test

For the competitive tests, the null hypothesis is: gene sets are not more associated with the phenotype than gene sets with similar properties. Therefore, the null phenotypes should be simulated so that the causal SNPs are evenly distributed along the genome without being enriched in any genes or gene sets. The phenotype was simulated with SGCP based on the QC'ed UK Biobank genotype data. All the SNPs had an effect size drawn from normal distribution *N(0, 1)*. Four data sets in which the heritability of the whole genome was 0.1, 0.3, 0.5, 0.7 were simulated. The association of gene sets were tested with MAGMA-geno, MAGMA-sum, PRSet-MAGMA.like and PRSet-perm. The process of simulation and testing were repeated 100 times.

In general, competitive MAGMA methods were over-conservative while competitive PRSet methods were inflated, except that MAGMA-sum appeared to be slightly inflated when the simulated heritability was low (Table 5). Both PRSet methods were more inflated as the simulated heritability increased. Notably, MAGMA-geno and PRSet-MAGMA.like used the same 2-tier framework and the only difference of these two methods is the Z score plugged into regression model in the second tier. The inflation of the PRSet methods when alpha=0.005 was more severe than that when alpha=0.05.

Both competitive PRSet methods, PRSet-MAGMA.like and PRSet-perm, were inflated especially for the results of small α threshold and the results of high simulated heritability, which indicates that the confounding factors were not perfectly accounted for by the PRSet methods, especially when the genetic signal was strong, that is, when the simulated heritability was high and when the α threshold was small.

It is also notable that PRSet-MAGMA.like and MAGMA-geno used the same framework as each other, except that the gene association was calculated in different ways. While PRSet methods were inflated, the competitive MAGMA methods were slightly conservative. The framework of MAGMA (included controlling for gene size, gene density, MAF and correcting for the gene-gene correlation calculated using the SNP PCs of the genes) may over-correct the gene association calculated by MAGMA using SNP PCs, but not sufficiently to correct for the gene association calculated by PRSet using gene PRS.

The fact that PRSet-perm was also inflated, especially when a small α threshold was used, shows that the current permutation method could not sufficiently correct for confounding factors, such as LD and gene set size. This inflation may result from the fact that in the observed gene sets the clumped SNPs are from a group of real genes and may still be in LD; while in the null gene sets, which consisted of clumped SNPs randomly chosen across all the genic regions, the SNPs in the null gene sets were therefore less likely to be in LD. Therefore, the permutation on clumped SNPs can capture the majority of the correlation or LD structure, but there was still residual LD. Future investigation is needed to find a way that more properly corrects for this inflation, but in the following analyses we evaluate the performance of the different methods according to sensitivity conditioned on a fixed specificity, as well as by comparing Area under the Curve (AUCs) from ROC analyses, and therefore these results incorporate the effects of this inflation.

Table 5 Type 1 error rate of competitive GSA methods

Simulated h2 = 0.1	**Alpha=0.05**	**Alpha=0.005**
MAGMA-Geno	0.0486±0.0103	0.0047±0.0024
MAGMA-sum	0.0512±0.0113	0.0053±0.0024
PRSet-MAGMA.like	0.0551±0.0099	0.0069±0.0028
PRSet-perm	0.0510±0.0090	0.0052±0.0027
Simulated h2 = 0.3	**Alpha=0.05**	**Alpha=0.005**
MAGMA-Geno	0.0486±0.0102	0.0047±0.0020
MAGMA-sum	0.0495±0.0102	0.0054±0.0022
PRSet-MAGMA.like	0.0516±0.0111	0.0067±0.0039
PRSet-perm	0.0518±0.0104	0.0068±0.0032
Simulated h2 = 0.5	**Alpha=0.05**	**Alpha=0.005**
MAGMA-Geno	0.0452±0.0089	0.0043±0.0019
MAGMA-sum	0.0482±0.0105	0.0049±0.0023
PRSet-MAGMA.like	0.0528±0.0096	0.0072±0.0038
PRSet-perm	0.0562±0.0100	0.0090±0.0049

Simulated h2 = 0.7	Alpha=0.05	Alpha=0.005
MAGMA-Geno	0.0456±0.0097	0.0047±0.0022
MAGMA-sum	0.0457±0.0096	0.0048±0.0022
PRSet-MAGMA.like	0.0589±0.0112	0.0084±0.0036
PRSet-perm	0.0584±0.0108	0.0110±0.0049

Ranking test

Ranking the gene sets according to their relative importance is one of the main applications of GSA methods. Here, a ranking test was performed to compare the ability of the competitive methods to distinguish the relative importance of gene sets: 10 real gene sets from KEGG-defined pathways were chosen randomly and assigned to be causal in the UK Biobank data. 50% SNPs of the first gene set, 45% SNPs of the second, 40% SNPs of the third, etc. were assigned to be causal. All the causal SNPs contributed equally to the phenotype. All other pathways were modelled as harbouring no causal genetic variants, although some contain causal variants due to overlap with one or more of the ten pathways modelled as causal. The phenotype was simulated based on a standardized genotype matrix with GCTA[76] (see section 2.3.5). Then the data were analysed with competitive MAGMA and competitive PRSet. The gene sets were then sorted by their competitive *P*-value. If multiple gene sets tie, the ranking is the mean rank of the tying gene sets.

Ideally, the estimated rank should be same as the simulated rank. The closer the two ranks are, the more power the method has. However, it is very likely that the gene sets that were not assigned to be causal can be associated with the phenotype because of the genes shared with other causal gene sets or SNPs in LD with causal gene sets. LD structure and stochastic variation, which mean that the index SNPs in GWAS are not necessarily the causal SNPs, also mean that the gene sets with the smallest *P*-value are not necessarily the gene sets simulated to be causal. However, on average, over many permutations, the causal gene sets are more likely to rank higher than non-causal gene sets. The simulation and analysis process was repeated

100 times, with ten causal gene sets randomly selected each time, and the medium ranking of each of the ten causal gene sets was calculated for the different PRSet and MAGMA methods.

The ranking of the causal gene sets showed that when the simulated heritability was 0.1, PRSet-perm had similar performance to the MAGMA methods. As the simulated heritability increased, the ranking results of the PRSet-perm became worse relative to MAGMA. When the simulated heritability was higher than 0.1, PRSet-perm ranked the top gene sets to the same or similar rank and the estimated ranking, which was a mean of all tied ranks, was lower as the heritability increased.

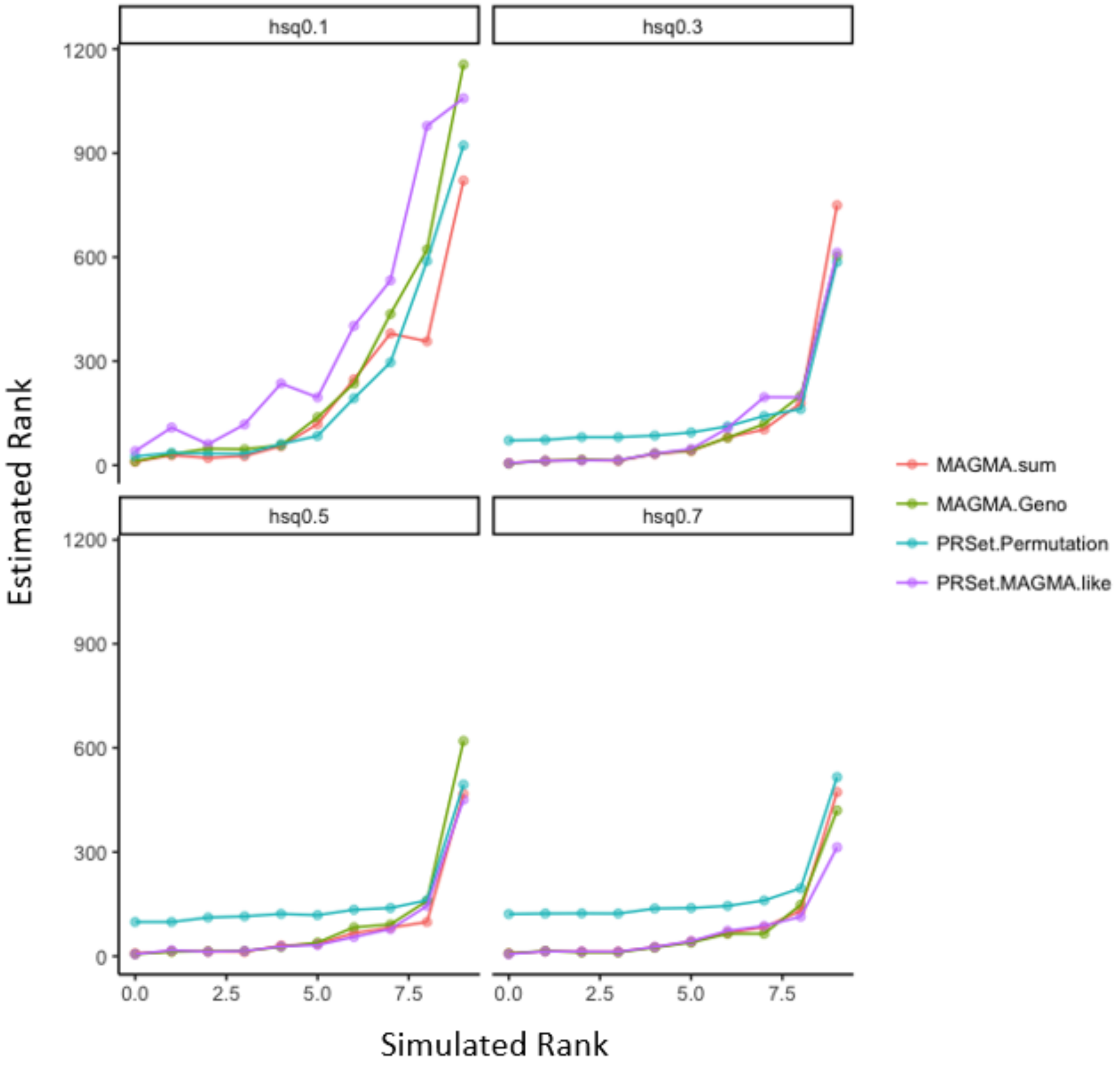


Figure 43 Ranking test of GSA methods. The estimated ranks were compared with the simulated ranks. The more similar the two rankings are, the better performance. Each plot shows a scenario under a simulated heritability. The plot shows that when the heritability is low, PRSet-perm performed well in terms of ranking the causal pathway according to the pre-set order, however, as the heritability increased, the PRSet-perm lost power because of the truncated P-value (as shown in Figure 44)

To investigate the causes underlying the ranking results, the competitive *P*-values of the causal gene sets corresponding to MAGMA-geno and PRSet-perm analyses were compared (Figure 49). If the *P*-values calculated by the two methods were highly correlated, then these two methods should generate similar ranking results. Figure 44 showed that PRSet-perm tended to give smaller P-values (most points above the Y = X) but the top results of PRSet-perm were truncated because of the limited number of permutations. The higher the simulated heritability, the higher power both PRSet-perm and MAGMA would have. This means that the overall correlation should be higher. However, since the more gene sets had truncated PRSet-perm *P*-values (Figure 44), which counteracted the increase of the power, as an overall result, the correlation between PRSet-perm and MAGMA-geno results were therefore slightly less correlated. The correlation between the results of the two methods was approximately 0.6 in all the simulated scenarios.

This investigation indicated that competitive MAGMA methods and PRSet methods had similar performance when PRSet-perm result did not reach the upper limit of the *P*-value. However, as the heritability increased, which means all the methods should have more power to detect the gene sets assigned to be causal, the performance of PRSet-perm was hindered by the truncated *P*-value: the top gene sets could not be distinguished from the less significant gene sets. This caused PRSet-perm to rank the most enriched gene sets as having an equal or similar rank to those gene sets with substantially lower enrichment. If the number of permutations were increased, we may expect that PRSet-perm would have similar performance to MAGMA even for traits with high heritability.

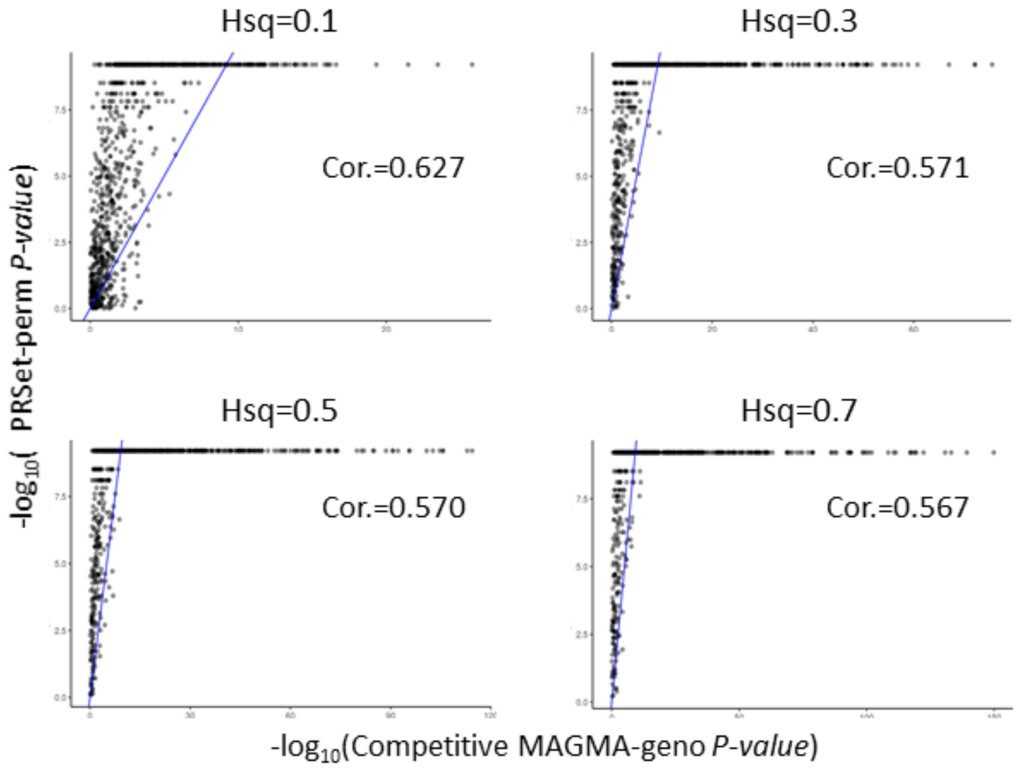


Figure 44 Comparing the competitive P-value of causal gene sets estimated by MAGMA-geno and PRSet-perm with the same set of simulated data simulated. Each data point indicates a real pathway in MSigDB that is assigned to be causal in the simulation. The x- and y-axis show the -log$_{10}$(P-value) of the causal pathway given by the two methods. Correlation Coefficient of -log$_{10}$(P-value) in scenario of simulated heritability 0.1, 0.3, 0.5 and 0.7 were calculated. The blue reference lines are y=x. The plots show that as the heritability increased, more pathways reached the PRSet-perm truncated P-value threshold due to the limited permutation times. This caused the loss of power of PRSet-perm in Figure 43

Sensitivity and specificity test.

The simulated data generated in the ranking test were also used for sensitivity and specificity test. Sensitivity is the ratio of causal gene sets that have a positive result (exceed *P*-value threshold) versus all the causal gene sets. Specificity is the ratio of non-causal gene sets that obtain a negative result (have larger *P*-value than *P*-value threshold) versus all the non-causal gene sets. The receiver operating characteristic (ROC) is created by iterating through *P*-value thresholds and plotting (1 – specificity) versus the sensitivity.

Competitive MAGMA and PRSet have similar sensitivity and specificity and ROC curve (Figure 45). Table 6 shows that PRSet-perm performed similarly or better than MAGMA methods for heritabilities of 0.1 and 0.3 but its performance declined for heritabilities of 0.5 and 0.7.

As discussed in the previous section, the truncation of empirical *P*-values due to finite permutations makes it impossible to distinguish the most enriched gene sets from those with lower enrichment. Thus PRSet-perm ROC had a linear segment between (0,0) and the first data point with a non-truncated *P*-value, while the ROCs of other methods have more data points to form a convex segment. Therefore, the AUC of PRSet-perm was smaller than it could have been if the *P*-values were not truncated.

Table 6 The AUC of ROC curve and sensitivity and specificity of GSA method. The results were based on the UK Biobank genotype data and simulated phenotype data where 10 of real pathways in MSigDB were assigned to be causal.

Simulated heritability =0.1	AUC of ROC curve	Sensitivity when specificity = 0.9
MAGMA-geno	0.833	0.618
MAGMA-sum	0.843	0.626
PRSet-MAGMA.like	0.772	0.494
PRSet-perm	0.826	0.632
Competitive R^2	0.829	0.577
Simulated heritability =0.3	AUC of ROC curve	Sensitivity when specificity = 0.9
MAGMA-geno	0.910	0.758
MAGMA-sum	0.911	0.771
PRSet-MAGMA.like	0.880	0.723
PRSet-perm	0.910	0.775
Competitive R^2	0.932	0.812
Simulated heritability =0.5	AUC of ROC curve	Sensitivity when specificity = 0.9
MAGMA-geno	0.931	0.820
MAGMA-sum	0.931	0.819
PRSet-MAGMA.like	0.917	0.789
PRSet-perm	0.916	0.767
Competitive R^2	0.946	0.851
Simulated heritability =0.7	AUC of ROC curve	Sensitivity when specificity = 0.9
MAGMA-geno	0.933	0.832
MAGMA-sum	0.934	0.833
PRSet-MAGMA.like	0.930	0.805
PRSet-perm	0.924	0.751
Competitive R^2	0.951	0.863

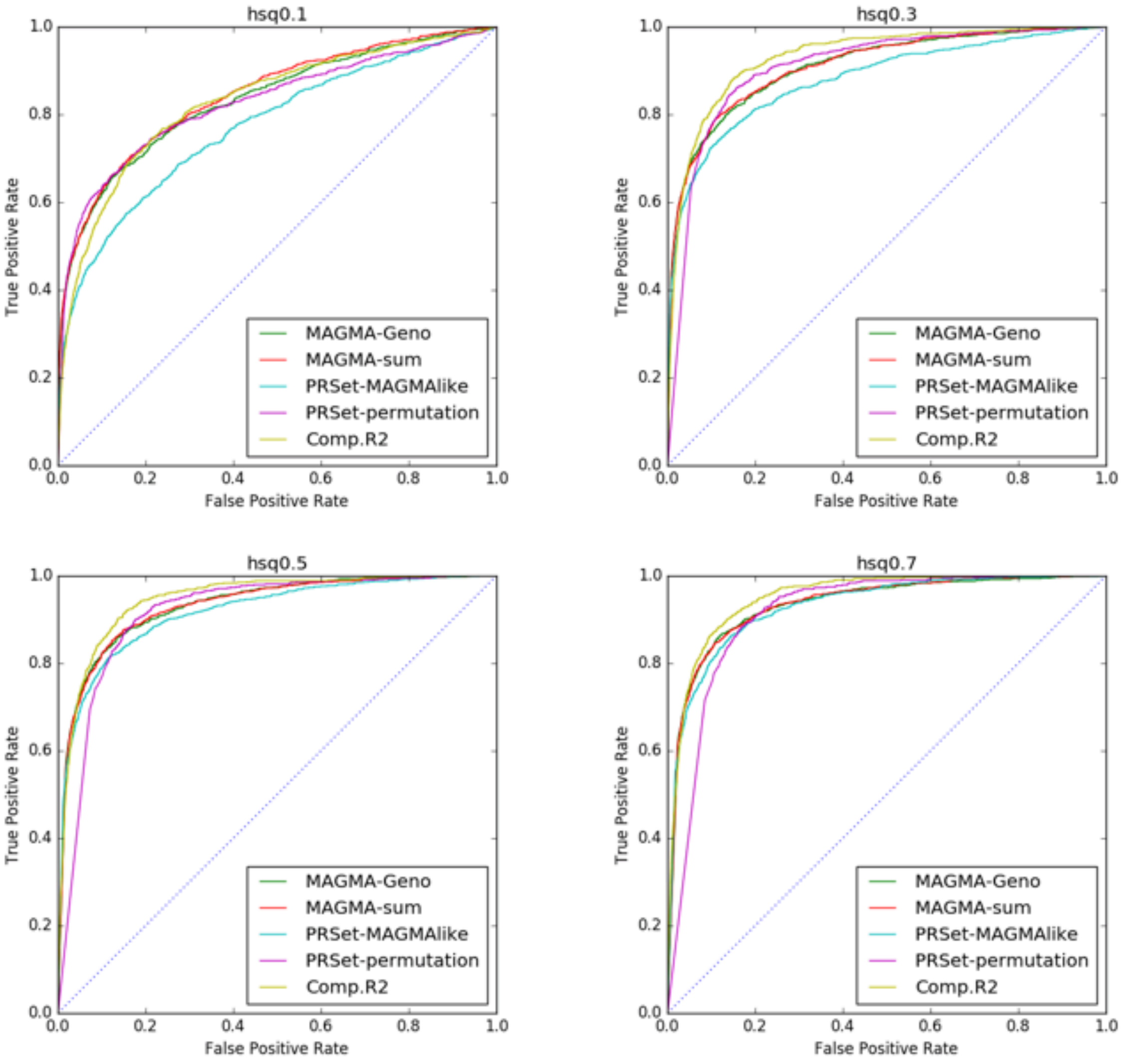


Figure 45 ROC curves of GSA methods tested with simulated data. The results were based on the 100 repetitions of tests based on UK Biobank genotype data and simulated phenotype data where 10 of real pathways in MSigDB were assigned to be causal.

Test on real traits

In this section, the results of competitive PRSet methods are compared with MAGMA-geno using real phenotype data from the UK Biobank on height, BMI, FVC and fluid intelligence (see section 4.2.4).

The same pattern observed in the self-contained tests were observed in the competitive tests too: in both the comparison of MAGMA-geno versus PRSet-MAGMA.like and the comparison of MAGMA-geno versus PRSet-perm, the most heritable trait, height, had the highest correlation in results among the methods; BMI, FVC and fluid intelligence had similar estimated heritability but different correlation in results. However, the result of competitive PRSet methods and MAGM-geno were less correlated compared with the self-contained results. As mentioned in section 4.3.1, the correlation of self-contained results might be inflated because of confounding factors, especially gene set size. However, it is interesting that the correlation in results between PRSet-MAGMA.like and MAGMA was much lower than self-contained results, given that both methods are based on gene association Z scores and use the same framework (Figure 46). The correlation of the PRSet-perm result and MAGMA-geno results was even lower (Figure 48).

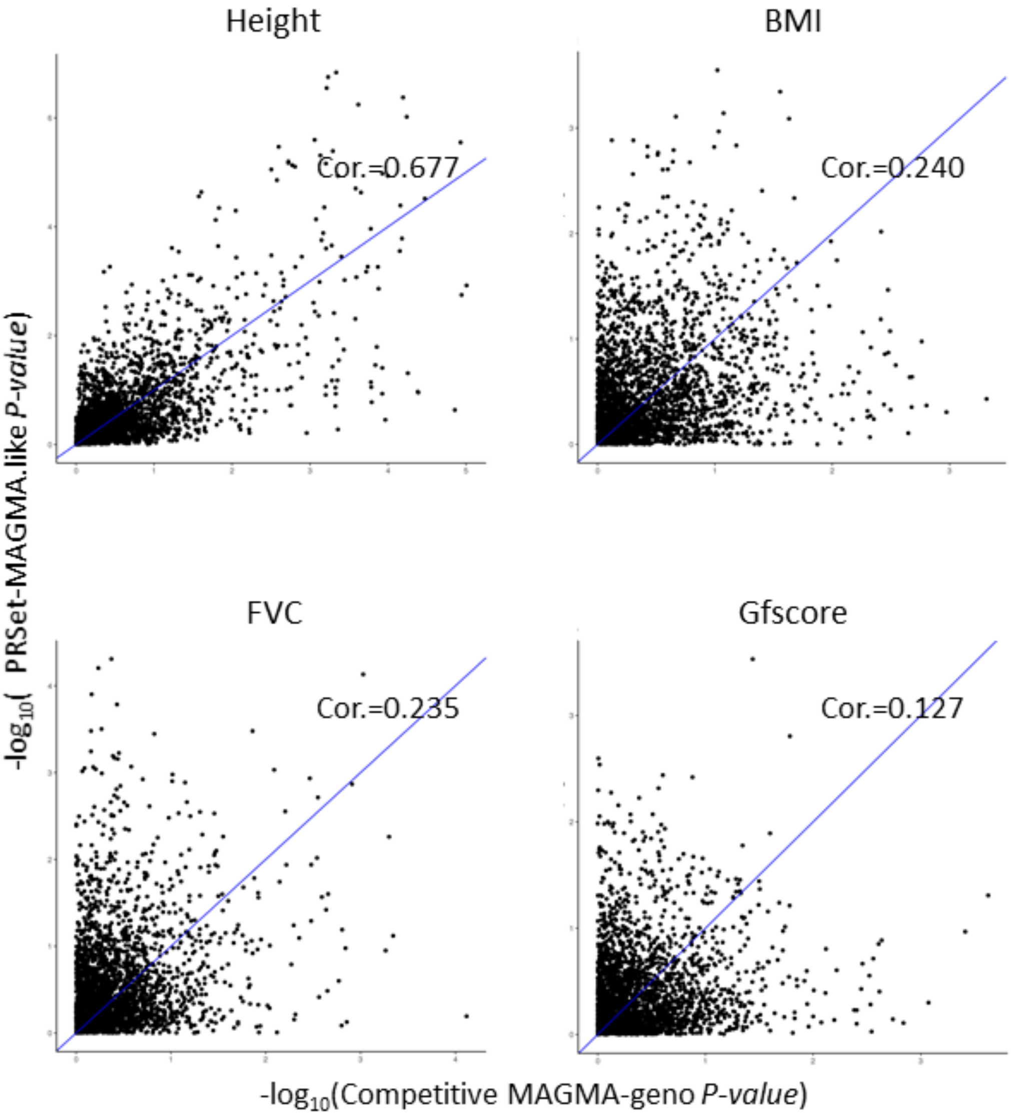


Figure 46 Comparison of competitive MAGMA-geno and PRSet-MAGMA.like result with the same set of UK Biobank data. Each data point indicates a real pathway in MSigDB. The x- and y-axis show the -log$_{10}$(P-value) given by the two methods. Pearson Correlation Coefficient of -log$_{10}$(P-value) of height, BMI, FVC and fluid intelligence (measured by Gf score) were calculated. The blue reference lines are y=x

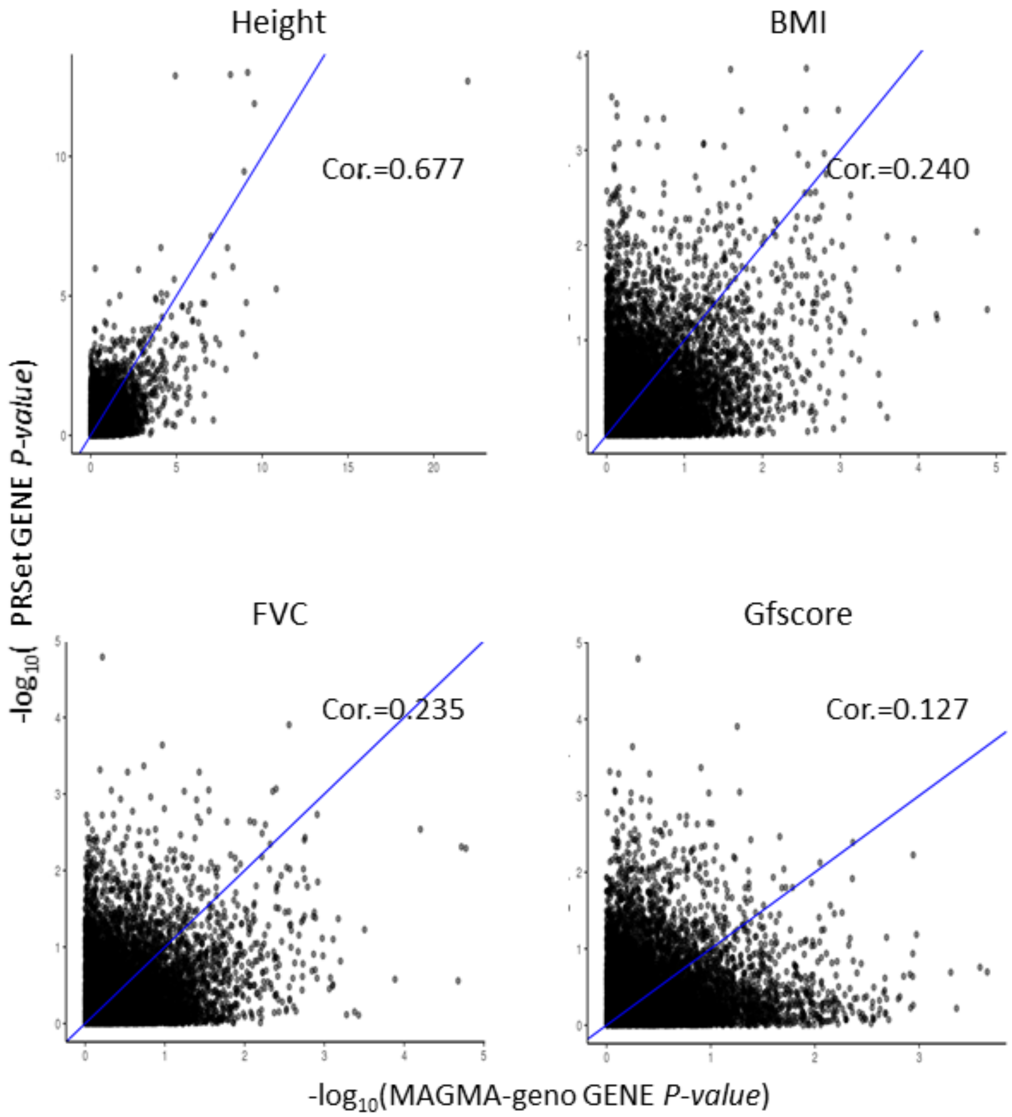


Figure 47 Comparison of gene P-value of MAGMA-geno and PRSet on the same set of UK Biobank data. Each data point indicates a real gene. The x- and y-axis show the -log_{10}(P-value) given by the two methods. Pearson Correlation Coefficient of -log_{10}(P-value) of height, BMI, FVC and fluid intelligence (measured by Gf score) were calculated. The blue reference lines are y=x

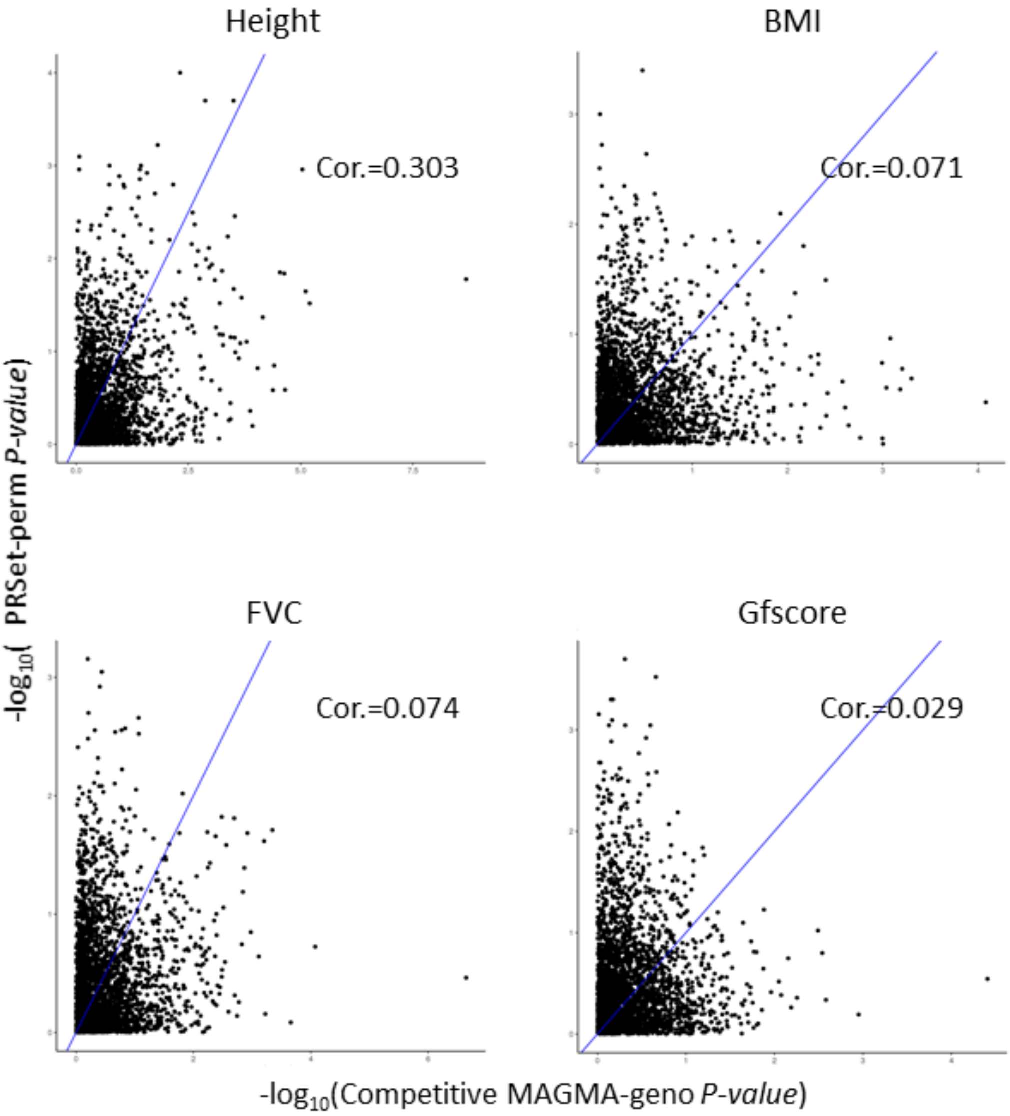


Figure 48 Comparison of competitive MAGMA and PRSet-perm result. Each data point indicates a real pathway in MSigDB. The x- and y-axis show the -log$_{10}$(P-value) given by the two methods. Pearson Correlation Coefficient of -log$_{10}$(P-value) of height, BMI, FVC and fluid intelligence (measured by Gf score) were calculated. The blue reference lines are y=x

As an investigation into the PRSet-MAGMA.like method, the *P*-values of the gene-phenotype associations calculated by PRSet and MAGMA were compared (Figure 47). For MAGMA, the gene-phenotype association is calculated using the PCs of the SNPs in the gene; for PRSet, the association is calculated using the PRS of the SNPs in the gene. Then the phenotype is regressed on the gene PCs or PRS, respectively. The gene-phenotype association can be viewed as the self-contained result of a gene set that only contains one gene. The correlation of the gene-phenotype association *P*-value was smaller than the correlation of self-contained gene set results between PRSet and MAGMA-geno, but larger than the correlation of the result of competitive MAGMA-geno and PRSet-MAGMA.like. This indicated that the self-contained results of smaller regions, for instance single genes, were more sensitive to chance variation than the results of larger regions, such as entire gene sets. PRSet-MAGMA.like and MAGMA-geno were based on self-contained gene association, which was more sensitive to chance variation. Therefore, the downstream competitive GSA methods based on these gene associations were likely more sensitive to chance variation than the self-contained gene set associations.

The correlation between the PRSet-perm and MAGMA-geno results was even lower. A possible reason for this was that the competitive tests were more complicated and less confounded by other factors, and therefore more prone to differences. Another reason was that PRSet-perm and MAGMA-geno used different frameworks, so they were powered to detect different gene sets.

Although the truncation of PRSet empirical *P*-values led to a power decrease in the ranking test and the sensitivity and specificity comparisons, this may not explain the low correlation observed in the real data results because the real traits were much more polygenic, so the gene sets in the real data did not reach the *P*-value truncation threshold.

In summary of competitive PRSet method, PRSet-perm was demonstrated to have similar performance for ranking gene sets to MAGMA when results were not truncated. Since the main use of PRSet is likely to be in exploiting the scores themselves, then it has been valuable here to show that our approach to measuring PRS across different gene set is powerful and thus is likely to have utility for researchers wanting to use the gene-set PRS for various applications.

4.3.3. The impact of gene set size and causal SNP fraction on competitive results.

In the previous section, it was suspected that different competitive methods capture gene sets of different features. Admittedly, real data were more complicated than simulated data and the low correlation of MAGMA and PRSet-perm observed in 4.3.2 may be largely caused by chance variation. It still worthwhile to investigate what factors may influence the GSA result. Here, two factors, the size of gene set and the fraction of causal SNPs, were investigated using a simple simulation where all the SNPs were independent (see section 4.2.4) as a preliminary step for further study of how different methods behave under various scenarios. The power of PRSet-perm and MAGMA to detecting gene set of different sizes (i.e. containing different number of SNPs) and causal SNP factions were compared using the simulated data.

The two competitive methods, PRSet-perm and MAGMA, gave similar *P*-values when the gene set had the same percentage of causal SNPs (20%) or a higher percentage of causal SNPs (50%). However, the PRSet-perm result had a limit for the *P*-value ≥ 1.0e-5 due to the number of permutations and the gene set of 250 SNPs and 50% causal SNPs reached the *P*-value limit. When the percentage of causal SNPs was 50%, larger gene sets had smaller competitive *P*-values. This result showed that the gene set size mattered for both self-contained and competitive tests. Given a certain percentage of causal SNPs, it is more probable that a random small gene set will have a strong signal or higher percentage of causal SNPs due to overlap with causal gene sets just by chance, compared with a large gen set. Therefore, smaller gene sets were less likely to have significant *P*-value for enrichment when compared with other gene sets of the same size. Overall, PRSet-perm methods gave slightly higher *P*-values than MAGMA in the simulated data, which accorded with the UK Biobank simulation as shown in Figure 44.

Table 7 Competitive P-value of gene sets containing 10, 200, 250 SNPs under different simple simulated scenarios given by MAGMA

	10 SNPs	**100 SNPs**	**250 SNPs**
20% Casual SNPs	0.5702±0.2784	0.4925±0.2510	0.5732±0.2511
50% Causal SNPs	0.2077±0.2222	6.5e-05±0.00039	3.7e-14±1.8e-13

Table 8 Competitive P-value of gene sets containing 10, 200, 250 SNPs under different simple simulated scenarios given by PRSet-perm

	10 SNPs	100 SNPs	250 SNPs
20% Casual SNPs	0.4287±0.2339	0.4676±0.2334	0.4676±0.2549
50% Causal SNPs	0.1518±0.1419	9.7e-05±0.00025	1.0e-5±0

The simulation showed that the size of gene set and the fraction of causal SNPs influence the power of PRSet-perm and MAGMA very similarly; large gene sets with high causal SNP fraction tended to have high significance. However, in the real trait, relative enrichment can be low due to the polygenicity and complexity of the real trait. Thus, the correlations of the real data results were low because of the low power.

Admittedly, LD structure was not included in this simulation. It is possible that MAGMA and PRSet-perm are differently powered when dealing with data with LD. The influence of LD on different GSA methods can be further investigated in the future work.

4.3.4. Investigating Competitive $\mathbf{R^2}$ and possibility of analytical GSA test based on it

As a possible alternative method of PRS-based competitive GSA method that may overcome the problem of truncated P-value of permutation method (see 4.2.3), $Competitive\ R^2$ were invested as an extension of PRSet methods.

Comparison of $\boldsymbol{Competitive\ R^2}$ and MAGMA statistics

$Competitive\ R^2$ and competitive MAGMA-geno P-value were compared. The correlation between these two results behaved similarly as the correlation between results of PRSet methods and MAGMA-geno. The correlation value was slightly higher than that of PRSet-perm and MAGMA-geno.

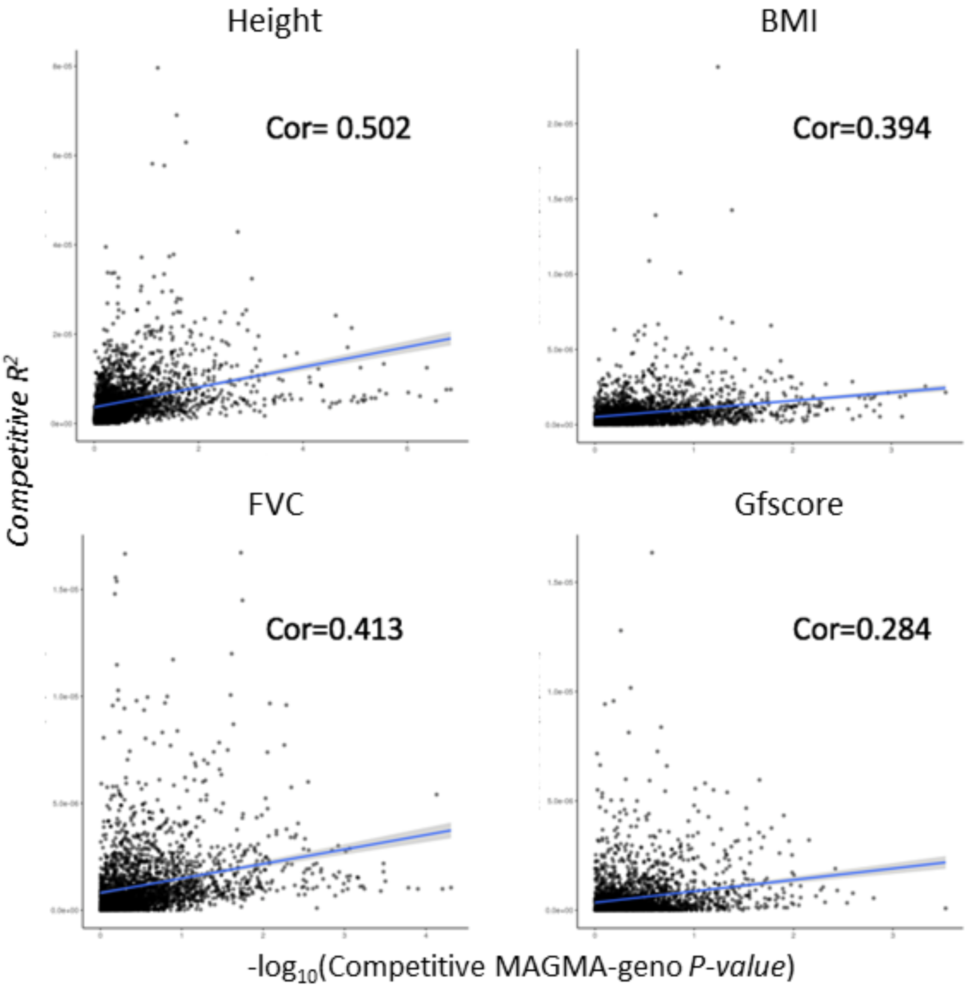


Figure 49 Comparison of competitive MAGMA and Competitive R^2 calculated with the same set of UK biobank data. Each data point indicates a real pathway in MSigDB. The x-axis shows MAGMA -$\log_{10}$(P-value) and the y-axis shows Competitive R^2. Pearson Correlation Coefficient between the two statistics of height, BMI, FVC and fluid intelligence (measured by Gf score) were calculated. The blue reference lines are the regression line instead of y=x because the two statistics are of different scale.

Investigating the distribution of $\boldsymbol{Competitive\ R^2}$ with simulated data

A simulation is performed to test whether the expectation of null $Competitive\ R^2$ is $Competitive\ R^2$of the "background" gene set containing all the genic region, R^2_{genic}. Random gene set containing different number of SNPs were generated and the $Competitive\ R^2$ of these random gene set was compared with R^2_{genic}. The whole process of simulating the data and calculating $Competitive\ R^2$was repeated for 1000 times. The mean value of R^2_{genic} and the mean value of $Competitive\ R^2$ of gene sets were almost identical except small gene sets of

10 SNPs (Figure 50). Empirically, in the simulated scenario, the mean $Competitive\ R^2$ of “backgroupd” gene set can be used as the expectation of $Competitive\ R^2$ of different sizes.

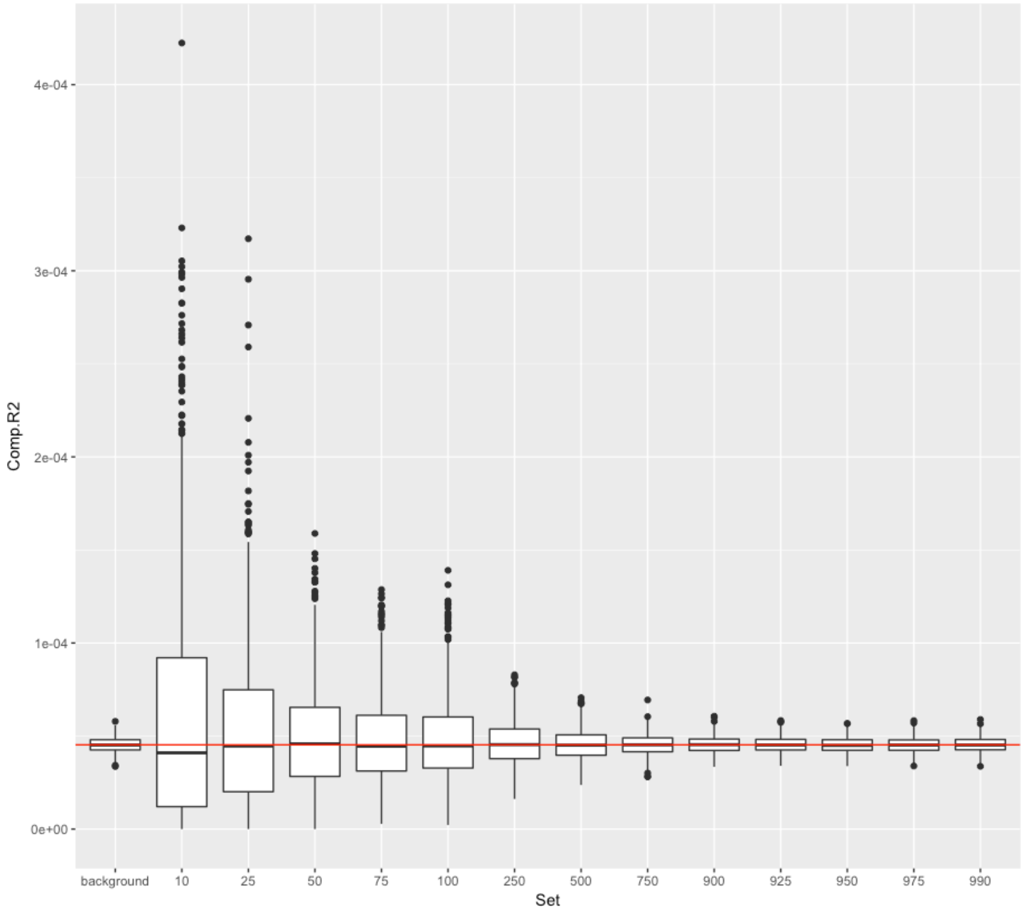


Figure 50 Comparing the "background" R^2_{genic} and Competitive R^2 of null gene sets of different sizes in simulated genotype and phenotype data. The x-axis shows the number of independent SNPs in the gene sets. The “background” gene set contains all the 1000 independent SNPs. The plot is based on 1000 repetitions of simulating data and analysis. The red line is the mean R^2_{genic}. The plot shows that the median Competitive R^2 of null gene sets of different sizes converged towards of the mean R^2_{genic}.

This simulation also provides information about standard deviation of $Competitive\ R^2$. In each round of simulation, the standard deviation of $Competitive\ R^2$ was calculated and the process was repeated for 1000 times. The standard deviation decreased as the gene set size increased (Figure 51). Yet the exact analytical relationship between the standard deviation and the gene set size was to be further investigated.

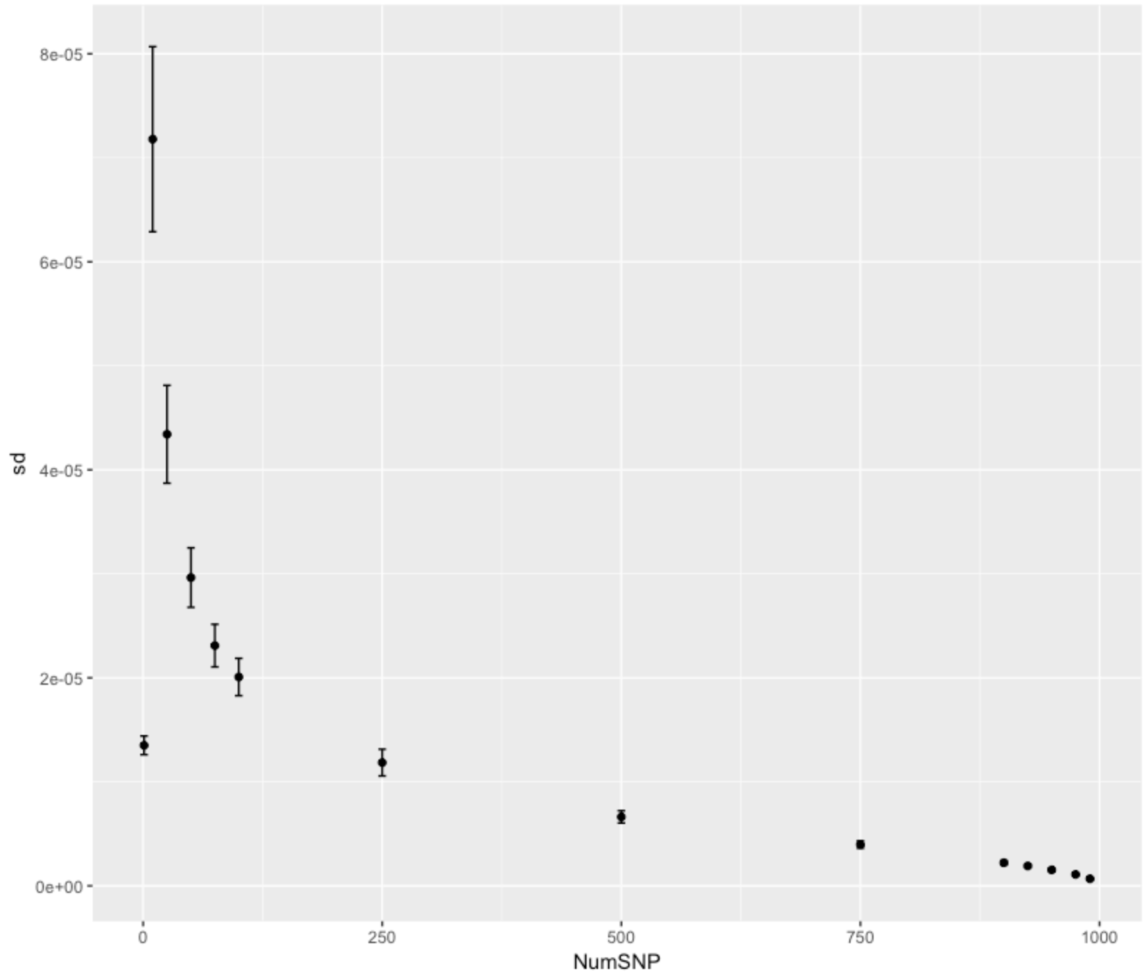


Figure 51 the standard deviation of Competitive R^2 of null gene set of different sizes in simulated genotype and phenotype data. The x-axis shows the size of gene set measured by the number of SNPs; the y-axis shows the mean $Competitive\ R^2$ sd of gene sets of the same size in one round of simulation. The mean and error bars are based on the results of repeating the simulation and calculation of gene set Competitive R^2 sd for 1000 times. The PRS R^2 of single SNP are also shown on the plot at the data point where the number of SNP is 1.

Investigating the distribution of $\boldsymbol{Competitive\ R^2}$ with real data

In the simulated situations, the factors such as population structure and LD are not included. In order to test whether the hypothesis of the expectation of null $Competitive\ R^2$ is true for real traits, gene sets containing different number of random real genes were generated. The $Competitive\ R^2$ of these random gene sets were calculated with real trait data. Theoretically, these gene sets do not have any biological meaning. They should a null distribution, which is probably a normal distribution, and a fixed expectation as in simulated data

However, the distribution of these null distribution showed that the mean value of $Competitive\ R^2$ varied with the gene set sizes. The mean $Competitive\ R^2$ varied with the number of genes in the gene set and the relationship was not linear (Figure 52). The $Competitive\ R^2$ was negatively correlated with the number of clumped SNPs and the slope varied from trait to trait (Figure 53).

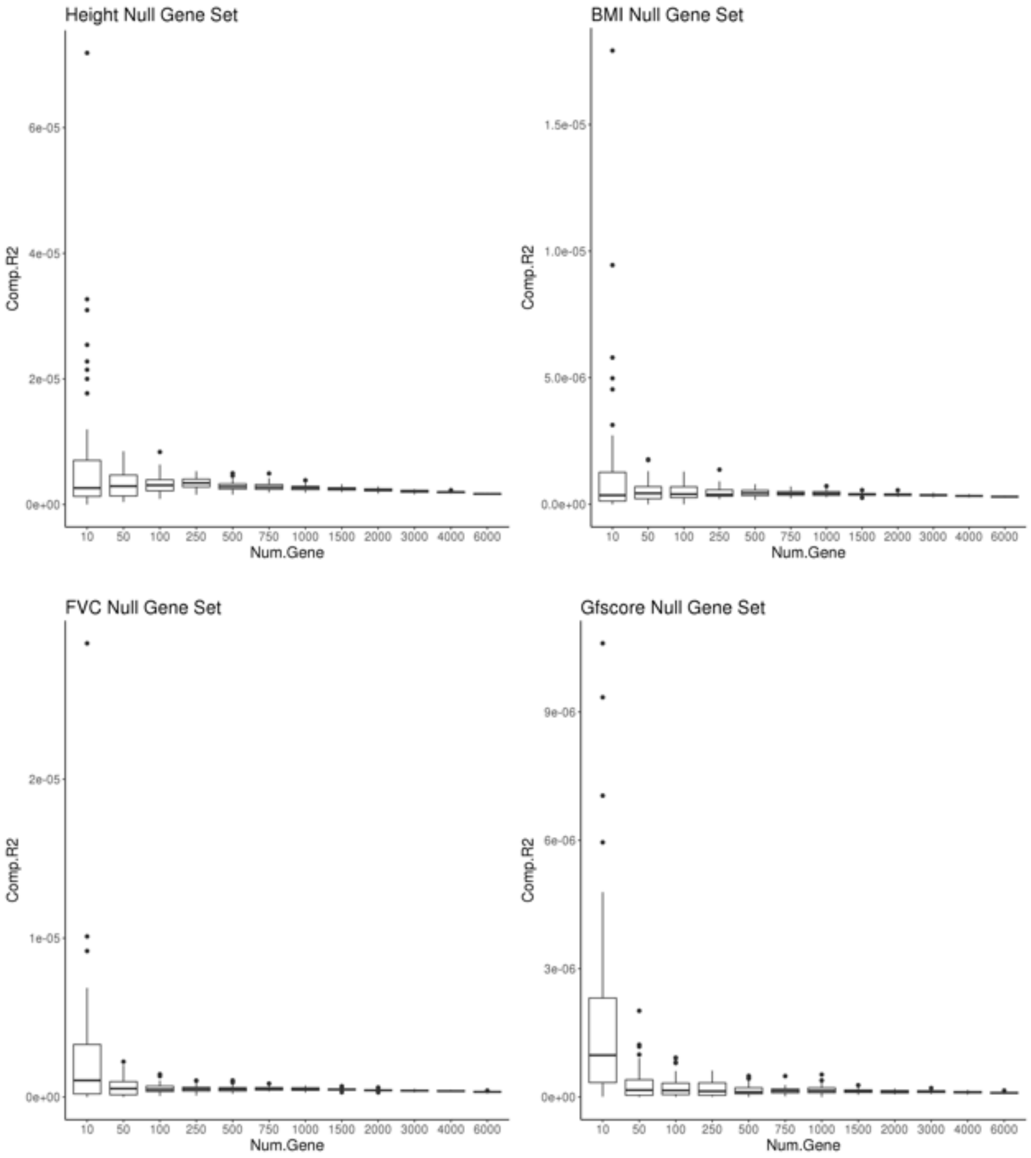


Figure 52 The distribution of Competitive R^2 of null gene sets of different sizes in UK biobank data: Height, BMI, FVC and fluid intelligence (Gf score). The gene sets were simulated so that they contained randomly chosen genes. The x-axis shows the gene set sizes measured in the number of genes in the gene set. The plots show that null Competitive R^2 of different sizes in the real data do not have the same median value and the relation between the gene size and Competitive R^2 is non-linear.

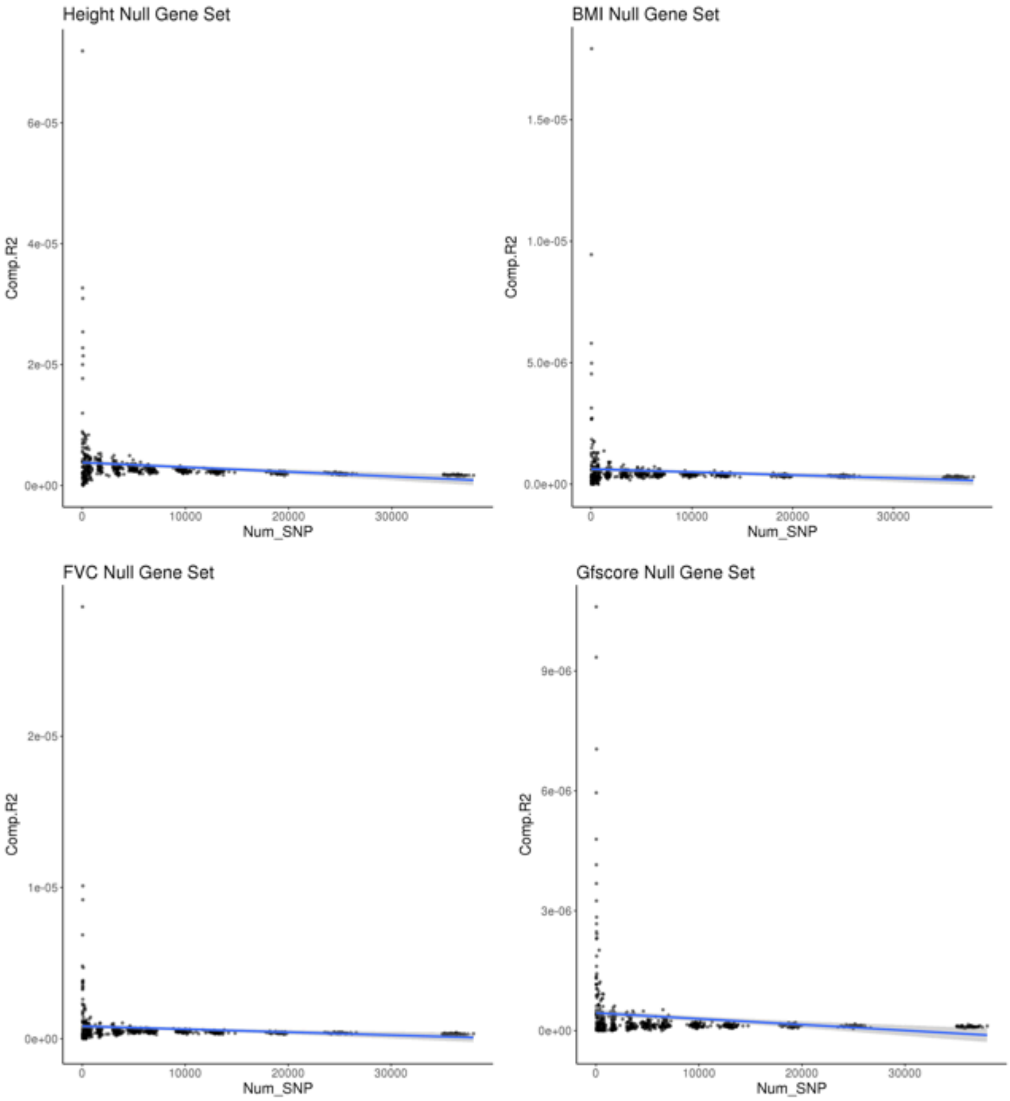


Figure 53 The distribution of Competitive R^2 of null gene set of different sizes in UK biobank data: Height, BMI, FVC and fluid intelligence (Gf score). The gene sets were simulated so that they contained randomly chose genes. The x-axis shows the gene set sizes measured in the number of clumped SNPs in the gene set. Similar with Figure 52, the plot shows that the relation between the gene size and Competitive R^2 are non-linear.

$Competitive\ R^2$ can be used as a statistic to represent the competitive significant of the gene sets. Yet, its distribution is too complicated to be described in an analytical way. The expectation of null $Competitive\, R^2$ in simulated data was $Competitive\, R^2$ of the "background" gene set containing all the genic region while the analytical description of the standard deviation is yet to be investigate. The null $Competitive\ R^2$ in real trait had a more complicated distribution probably due to the LD structure or the population structure that even the expectation of null $Competitive\, R^2$ was influenced by the gene set size and the relationship between $Competitive\, R^2$ and gene set was difficult to be described in an analytical way. Therefore, it is difficult to analytically describe $Competitive\ R^2$ distribution and derived the competitive P-value.

4.4.Discussion

In this chapter, PRSet, a group of GSA methods based on PRS, were developed. PRS is a summary of genetic signals across the genome or gene sets. In previous studies, PRS was used to test the polygenic basis of traits or the correlations between different traits. In PRSet, the applications of PRS was extended to gene set analysis.

The tests in this chapter showed it had similar power compared with the current leading GSA method MAGMA in both simulated and real trait data test.

Two competitive PRS-based GSA were developed: PRSet-MAGMA.like and PRSet-perm. PRSet-MAGAM borrows the framework of MAGMA-geno and also inherits the advantage of being much less computational expensive than permutation methods[65]. However, one of the aims of developing PRS-based GSA methods is to have an individual-level representative of genetic burden enriched in a gene set while PRSet-MAGMA.like method can only estimate the association between gene set and the phenotype at population level. Therefore, permutation test was developed because it calculates the individual-level gene set PRS.

In PRSet, a new way of controlling for LD structure was introduced. Controlling for LD is an important part of competitive GSA tests: If the approach is to compare the genes inside and outside the gene set, like in MAGMA, the correlation among the genes due to the LD structure should be considered and corrected for; if the approach is to generate a null distribution, the LD structure of the observed gene set and the null gene sets should be considered. In this chapter, a permutation approach using clumped SNPs was introduced to control for the LD in the observed gene set: the gene set is viewed as a group of clumped SNPs. The null distribution is therefore constructed with random clumped SNPs from genic regions.

In this chapter, LD was found to influence the GSA method in many aspects. LD may cause inflation and make the results calculated by different methods inconsistent with each other. Besides, LD structure may it difficult to identify the causal SNPs: A statistically significant SNPs in a gene does not necessarily mean that the gene contribute to the phenotype; this SNP may become significant because of other SNPs or other genes that in LD with it. In the future when causal SNPs and causal genes can be better identified thanks to the progress of fine-mapping[92], the input for GSA can be more accurate. GSA can be more powerful and less influenced by the bias caused by LD structure.

Chapter 5. Conclusion

Large-scale GWAS have reached the power to identify many genetic variants of modest or small effect contributing to complex traits with genome-wide significance. These findings indicate that most complex traits are polygenic, which means that hundreds or even thousands of SNPs of modest or small effect collectively contribute to human complex traits. Polygenic risk scores have been used to represent the contribution of common SNPs across the genome to polygenic traits. PRS has many successful applications, such as demonstrating the existence of the polygenic genetic basis of psychiatric diseases[82], detecting the association between different traits and discriminating subgroups of a complex disease[44,45]. In this project, I developed methods that can increase the power of PRS and expand its applications.

GWAS is the foundation of PRS and many other prediction methods. The validity of GWAS effect size estimates greatly influences the prediction power of PRS. In chapter 2 and chapter 3, I developed shrinkage methods Permutation Shrinkage (PS) and Order Statistics Shrinkage (OSS) to improve GWAS effect size estimates. The main principle of these two shrinkage methods is estimating the null effect sizes, or the contribution of 'error' to effect size estimates, and removing the null distribution from the observed value. The new methods assume that 1) the true effect sizes of variables are almost zero and only with a small proportion of exceptions 2) the highest observed values are the most inflated (Winner's Curse). Previous research showed that the proportion of susceptibility SNPs was less than 5%[71], which is in accord with the first assumption. Unlike other shrinkage methods, such as LASSO[37], ridge[53], which aim to minimize the residual error (/squared) with a penalty term or a constraint of the sum of effect sizes, these two methods estimated the error quantity and remove the error quantity from the observed estimates.

As to the implementation of these methods, the SNPs are first divided in to MAF bins so that in each bin the SNPs are more homogenous than all the SNPs across the genome and therefore the null distribution of the effect sizes is easier to estimate. In each bin, the null distribution is estimated by either permuting the outcome or inferring from order statistics. When the individual-level data are available, the null distribution can be generated by running GWAS on the permuted phenotype; when only summary statistics data are available, then the null effect size can be inferred from the MAF of the base dataset or other alternative summary statistics and an assumption that the null *P*-value follows a uniform distribution U(0,1). Depending on how the null distribution is generated, the methods are called Permutation Shrinkage (PS) or Order Statistics Shrinkage (OSS). These methods were tested with quantitative traits in UK Biobank and significantly increased the PRS prediction. In our tests, both shrinkage methods achieved an average of 35% relative increase in prediction R^2.

Nevertheless, my work on these two methods has limitations. First, I have not tested whether the PS or OSS can work efficiently on binary data; second, OSS assumes that the SNPs in the MAF bins can be viewed as independent. It may be true for sparse data like UK Biobank but in order to analyse denser data such as imputed GWAS data, OSS needs to be further tested and optimised; third, I only tested the methods with UK Biobank data, which is very large, while in many cohort studies the sample size may not be as large and powerful.

Gene set analysis is another future application of PRS analysis. PRS can be used to investigate the genome-wide polygenic genetic basis and genetic correlation between two polygenic traits. However, more specific questions on the mechanism are yet to be answered. It makes intuitive sense that different gene sets or pathways contribute differently to the outcome. For example, different subtypes of the same disease should share some common causal pathways that give the subtypes the similar features; they should also have their specific causal pathways that distinguish them from each other. Critical information of the subtype-specific feature may be lost if only the genome-wide genetic basis or genetic correlation of the polygenic traits are investigated. Gene set PRS may capture the more detailed and specific features relevant to the trait aetiology. Besides, gene set PRS can have many applications, such as constructing complex models that investigate the mechanism and causality of gene set-specific genetic basis. In chapter 4, I developed a set of PRS-based gene set analysis methods. These methods were implanted into PRSice[77], the software developed in our group, and are collectively called "PRSet".

Usually the genome-wide PRS is the sum of clumped risk alleles weighted by SNP effect size estimates. In the standard approach, PRS are optimized by iterating through a series of *P*-value thresholds and choosing the one that produces the most significant result. In this project, the gene set PRS was defined as the sum of all the clumped risk alleles that fall into the range of the genes in that gene set, weighted by the SNP effect size estimates from previous GWAS.

The flanking 5' and 3' regions of the gene can be counted as part of the gene according to the requirement of the research. The size of the flanking regions can be adjusted according to the specific hypothesis. Besides, the gene set PRS is not to be optimized by choosing the best *P*-value threshold. The gene PRS can be calculated in the same way.

To develop PRS-based GSA method, I started with a fundamental application of identifying the gene sets associated with the traits. The association test can be either self-contained or competitive. The null hypothesis for self-contained analysis is that gene set is not associated with the trait; the null hypothesis for competitive analysis is that the gene set is no more associated with the trait than other gene sets. The self-contained test is implemented by regressing the trait over the gene set PRS; the competitive test can be implemented in two ways: one is borrowed from an existing method MAGMA[65], which calculates the association of each gene with the trait and compares the gene inside the gene set and outside the gene set; The other uses a permutation approach. It calculates the association between the observed gene set and the outcome and the association of corresponding null gene sets and the outcome, where the null gene sets consist of the same number of clumped SNPs randomly drawn from the genic regions. The empirical *P*-value comparing the observed gene set association against the null distribution of gene set association is the competitive *P*-value.

In our tests with the UK Biobank data, the PRS methods had similar performance with MAGMA, the current leading method for calculating gene set association. The competitive *P*-value calculated by permutation approach is limited by the permutation number and it is very computational expensive. However, individual-level gene set PRS can not only be used for calculating the gene set association in the permutation approach but also many other applications, for example, complex graphical models that investigate the possible interaction of gene set, endophenotype, phenotype and environment exposures, stratifying individuals with diseases into more homogenous subsets, and identifying the causes of differences in treatment response.

Nevertheless, many technical details of the PRSet methods can be further optimised: for example, the method for calculating gene set PRS can be optimised so that it can be both unbiased and more powerful than using all the clumped SNP. Currently, PRSet uses clumping to account for the LD structure but there seems to be some effect of residual LD that caused slight inflation of PRSet permutation method. Furthermore, gene sets could be defined according to information on expression of genes, such as via eQTL information, instead of by physical location of genes. Also, conditional analyses could be incorporated into PRSet analyses so that the signal of each gene set is conditioned on that of overlapping gene sets, to identify the gene sets driving the signal; this is possible with the PRSet approach since PRS across each gene set is available for every individual. Despite the current limitations of the PRSet approach, chapter 4 builds a scaffold for PRS-based gene set analysis.

Chapter 6. Future Prospects

6.1.Systematic PRS power test with UK Biobank data

Many research studies use the data collected especially for a single project. The collected sample size is limited by funding, time and the scarcity of the samples and can be as small as several hundreds. In previous research, Dudbridge *et al* [52] studied aspects that influenced the power of PRS, such as sample size, heritability and *P*-value thresholding, and proposed a set of tools to analytically calculate the base and target sample size needed to achieve certain power in a typical PRS analysis[93–95]. These research were based on simulated genotype data, which made a range of assumptions, such as all the SNPs being in Hardy-Weinberg equilibrium. Therefore, it is worthwhile verifying the predictions made by Dudbridge to systematically test the power of PRS using data simulated directly from real genotype data.

The UK Biobank has a large sample size and makes it possible to run PRS power tests based on UK Biobank genotype data and simulated phenotypes with different genetic architecture, such as different fraction of causal SNPs and heritability, across different base and target sample sizes. Besides the power test, it is also worthwhile to systematically test the performance of the shrinkage methods developed in this project and other shrinkage methods such as LDpred[35] and lassosum[36], since usually when the raw data appear underpowered, researchers are more likely to try to increase the power by using shrinkage methods, but it is important to know which methods have most power.

6.2.Power test and development of optimization methods for cross-population PRS analysis

The power test of cross-population PRS analysis are highly needed because most of the large-scale GWAS were conducted on Caucasian sample [96,97]. The power may be reduced if we directly use Caucasian GWAS data to construct PRS model in other non-Caucasian populations because the LD structure, allele frequencies, and other factors are different[96].

Although the majority of available GWAS data are from Caucasian samples, large scale data sets e.g. GIANT consortium data, Psychiatric Genomics Consortium data, Wellcome Trust Case Control Consortium and UK Biobank also included non-Caucasian samples and non-Caucasian data sets, such as China Kadoorie Biobank data (http://www.ckbiobank.org/site/), are becoming more and more available. It is possible to systematically test PRS power and shrinkage method performance using cross-population base and target samples. The result of using the cross-population samples can be compared with the results using UK Biobank data only to estimate the influence of using cross-population samples on PRS predictive power.

It is possible that we may observe a decline of PRS power and shrinkage method performance when using cross-population samples. The shrinkage methods that were developed in this project were tested with UK Biobank data. UK Biobank is a large-scale homogenous data set and only Caucasian samples were included in the analysis. This is an over-simplified and over-optimized scenario compared with what we might encounter in smaller cohort studies where the samples can of smaller size or more heterogenous. However, the performance of the shrinkage methods in various scenarios will be a useful reference for other researchers in the field.

It is also possible that the power tests mentioned above provide clues for optimising PRS analysis for cross-population samples. Generalized PRS models were proposed as a solution for the problem of power reduction in cross-population analysis[96] but there may be other options For example, the shrinkage methods developed in this PhD project showed potential to increase PRS prediction in the same population. The mechanism of these shrinkage methods is to remove the noise from the observed effect size estimates. If removing the noise can increase the PRS power using samples in the same population, it is possible that the similar methods may work for cross-population samples.

6.3. Expanding and refining shrinkage methods

In my PhD project, shrinkage methods for quantitative data were developed and tested. Since most of the disease-related GWAS data are binary data, it is more demanding to develop shrinkage methods that work for the binary data. I proposed that the development of shrinkage methods for binary data should first start with permutation shrinkage methods: The null distribution of the odds ratio (OR) or log(OR) can be estimated by permutation. The corrected log(OR) distribution is the observed log(OR) with the null log(OR) subtracted. At the same time, the nature of the null distribution of the OR or log(OR) need to be investigated. If permutation shrinkage works for binary traits and the null distribution of the odds ratio can be transferred from order statistics, the order statistics shrinkage can be developed accordingly.

There are other specific technique issues that needs to be highlighted. First, we need a more rigorous and effective methods to correct data with dense SNPs. Our methods so far only assumed that the SNPs in each MAF bin are sparse enough to be viewed as independent. If not, the SNPs will be first clumped and then corrected. However, in our test, correcting the clumped data will lead to less power increase. Ideally, the effective number of independent tests should be properly calculated, and the null distribution need to be extrapolated with the consideration of the LD structure. Second, the available GWAS summary statistics were not necessarily calculated from standardised phenotype, while our order statistics method works for effect sizes calculated from standardised phenotype. It is necessary to expand the order statistics so that it can be applied to GWAS using unstandardized phenotype.

6.4.PRS-based gene set analysis.

In this project, a set of methods that use PRS to identify the associated gene sets were developed. PRS-based gene set analysis (GSA) can be further expanded to investigate the aetiology of complex traits in more detail.

6.4.1. Discriminate the gene set PRS profile of diseases or disease subgroups

Previous research have shown that patients of different diseases or different subtypes have different PRS profiles. For example, schizophrenia PRS of bipolar disorder (BD) patients are significantly different from controls, while no significant difference was found in non-psychiatric disease patients and controls[82]; schizoaffective BD patients have higher schizophrenia PRS than non-schizoaffective BD patients[44]. These findings show that different diseases / subtypes have different genetic basis. Yet specific mechanism about which gene sets or pathways contribute to the different diseases is still unclear. The gene set PRS profiles may help us answer the question whether different diseases/subgroups have shared associated gene sets and/or distinguishing associated genes sets. These findings can help us understand the aetiology of diseases. Hopefully the distinguishing gene sets can help to classify the clinical samples according to their gene set PRS profile.

6.4.2. Combining the gene set PRS with other functioning/endophenotype data.

A complex trait may have some simpler underlying phenotype. For example, a complex psychiatric disease may be associated with the change of working memory, prepulse inhibition, or activity or structure changes of the brain that are found by fNMR, etc. These underlying phenotypes are referred to as endophenotypes.

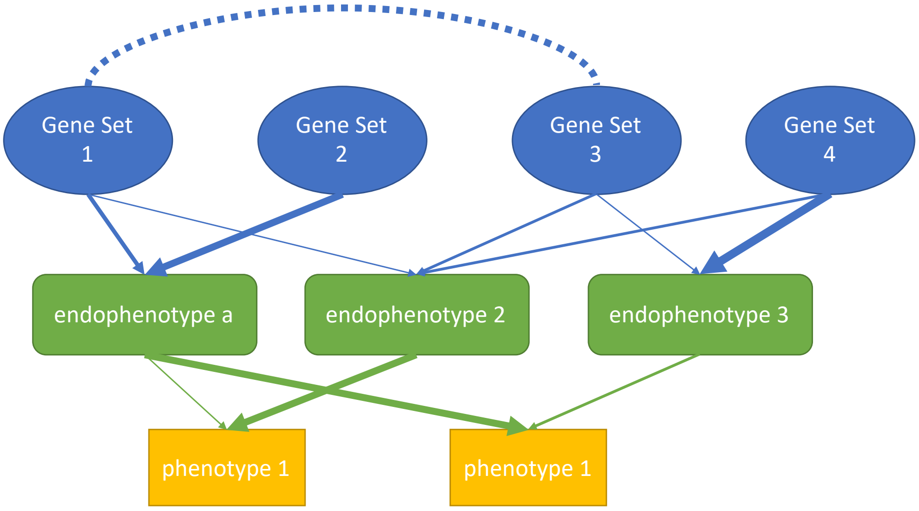


Figure 54 Schematic diagram of network model where gene set PRS, endophenotype and phenotype can be used. The dashed line between gene sets indicates the correlation between gene sets due to shared SNPs and LD structure. The correlation between the arrows indicate causal relationship; the weight of the line indicates the strength of the causal relationship.

Since endophenotypes are much more clearly defined and it is easier to study their molecular mechanism, it is easier to link endophenotype with certain gene sets or pathways than a complex trait. A network, as shown in the schematic diagram in Figure 54, can be constructed to model the underlying causal relationship if the endophenotypic data are available. One possible advance of using gene set PRS is that the correlation between gene sets simply due to shared SNPs and LD can be easily calculated and controlled for. It can make the interpretation of the gene set function analysis more rigorous. For example, the correlation of the downstream outcome that cannot be explained by the correlation of the shared SNPs and LD is more likely to be caused by the overlapping of their actual biological function.

Reference

1. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Brooker, R. J. *Genetics: Analysis and Principles*. (McGraw-Hill Higher Education, 2009).
3. Bonci, A., Lupica, C. R., Morales, M., Bellen, H. J. & Yamamoto, S. Morgan's Legacy: Fruit Flies and the Functional Annotation of Conserved Genes. *Cell* **163**, 12–14 (2015).
4. Pulst, S. M. Genetic linkage analysis. *Arch. Neurol.* **56**, 667–672 (1999).
5. Riley, J. *et al.* A novel, rapid method for the isolation of terminal sequences from yeast chromosomes clones. *Nucleic Acids Res.* **18**, 2887–2890 (1990).
6. Meng, J., Schroeder, C. M., O&#039;Connor, M., Peifer, M. & Bender, W. Construction of large DNA segments in Escherichia coli. *Science (80-.).* **244**, 1307 LP – 1312 (1989).
7. MacDonald, M. E. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
8. Buckler, A. J. *et al.* Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 4005–9 (1991).
9. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
10. Zhu, C., Gore, M., Buckler, E. S. & Yu, J. Status and Prospects of Association Mapping in Plants. *Plant Genome J.* **1**, 5 (2008).
11. Risch, N. *et al.* The future of genetic studies of complex human diseases. *Science* **273**, 1516–7 (1996).
12. Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. Genomewide Scans of Complex Human Diseases: True Linkage Is Hard to Find. *Am. J. Hum. Genet.* **69**, 936–950 (2001).
13. Consortium, I. H. G. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860 (2001).
14. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

15. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

16. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

17. Long, A. D. & Langley, C. H. The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits. 720–731 (1999). doi:10.1101/gr.9.8.720

18. Astle, W. & Balding, D. J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Stat. Sci.* **24**, 451–471 (2009).

19. The Wellcome Trust Case Control Consortium *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 (2007).

20. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **10**, 8192–8192 (2009).

21. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

22. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Gen* **42**, 565–569 (2010).

23. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).

24. Zaitlen, N. *et al.* Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS Genet.* **9**, (2013).

25. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135 (2012).

26. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

27. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).

28. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–9 (2013).

29. Choi, S. W., Shin, T., Mak, H. & Reilly, P. F. O. A guide to performing Polygenic Risk Score analyses. *bioRxiv* **5**, 11–13 (2018).

30. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **advance on**, 291–295 (2015).

31. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236–1241 (2015).

32. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, (2013).

33. Márquez-Luna, C., Loh, P. R. & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).

34. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

35. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

36. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).

37. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).

38. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).

39. Goddard, M. E., Wray, N. R., Verbyla, K. & Visscher, P. M. Estimating Effects and Making Predictions from Genome-Wide Marker Data. *Stat. Sci.* **24**, 517–529 (2010).

40. Robinson, M. R. *et al.* Genetic evidence of assortative mating in humans. *Nat. Hum. Behav.* **1**, 16 (2017).

41. Maier, R. M. *et al.* Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat. Commun.* **9**, 1–17 (2018).

42. Selzam, S. *et al.* Predicting educational achievement from DNA. *Mol. Psychiatry* **22**, 267–272 (2017).

43. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

44. Hamshere, M. L. *et al.* Polygenic dissection of the bipolar phenotype. *Br. J. Psychiatry* **198**, 284–288 (2011).

45. Aminoff, S. R. *et al.* Polygenic risk scores in bipolar disorder subgroups. *J. Affect. Disord.* **183**, 310–314 (2015).

46. Meyers, J. L. *et al.* Interaction between polygenic risk for cigarette use and

environmental exposures in the Detroit Neighborhood Health Study. *Transl. Psychiatry* **3**, e290–e290 (2013).

47. Peyrot, W. J. *et al.* Effect of polygenic risk scores on depression in childhood trauma. *Br. J. Psychiatry* **205**, 113–119 (2014).

48. Mullins, N. *et al.* Polygenic interactions with environmental adversity in the aetiology of major depressive disorder. *Psychol. Med.* **46**, 759–770 (2016).

49. Peyrot, W. J. *et al.* Does Childhood Trauma Moderate Polygenic Risk for Depression? A Meta-analysis of 5765 Subjects From the Psychiatric Genomics Consortium. *Biol. Psychiatry* **84**, 138–147 (2018).

50. Zhang, J.-P. *et al.* Schizophrenia Polygenic Risk Score as a Predictor of Antipsychotic Efficacy in First-Episode Psychosis. *Am. J. Psychiatry* appi.ajp.2018.17121363 (2018). doi:10.1176/appi.ajp.2018.17121363

51. García-González, J. *et al.* Pharmacogenetics of antidepressant response: A polygenic approach. *Prog. Neuro-Psychopharmacology Biol. Psychiatry* **75**, 128–134 (2017).

52. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, (2013).

53. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55–67 (1970).

54. James, W. & Stein, C. Estimation with Quadratic Loss. in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* 361–379 (University of California Press, 1961).

55. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **67**, 301–320

56. Meinshausen, N. Relaxed lasso. *Comput. Stat. Data Anal.* **52**, 374–393 (2007).

57. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

58. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27**, 29–34 (1999).

59. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

60. Wang, K., Li, M. & Bucan, M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am. J. Hum. Genet.* **81**, 1278–1283 (2007).

61. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of

gene-set analysis. *Nat. Rev. Genet.* **17**, 353–364 (2016).

62. De Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **17**, 353–364 (2016).

63. Wadi, L., Meyer, M., Weiser, J., Stein, L. D. & Reimand, J. Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* **13**, 705–706 (2016).

64. Pedroso, I. I., Barnes, M. R., Lourdusamy, A., Al-Chalabi, A. & Breen, G. FORGE: multivariate calculation of gene-wide p-values from Genome-Wide Association Studies Authors and Affiliations. *bioRxiv* (2015).

65. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol.* **11**, 1–19 (2015).

66. Pedroso, I. *et al.* Common genetic variants and gene-expression changes associated with bipolar disorder are over-represented in brain signaling pathway genes. *Biol. Psychiatry* **72**, 311–317 (2012).

67. Davis, L. K. *et al.* Partitioning the Heritability of Tourette Syndrome and Obsessive Compulsive Disorder Reveals Differences in Genetic Architecture. *PLoS Genet.* **9**, (2013).

68. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Doi.Org* **50**, 176834 (2017).

69. Nakaoka, H. & Inoue, I. Meta-analysis of genetic association studies: Methodologies, between-study heterogeneity and winner's curse. *J. Hum. Genet.* **54**, 615–623 (2009).

70. Forstmeier, W. & Schielzeth, H. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behav. Ecol. Sociobiol.* **65**, 47–55 (2011).

71. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, (2018).

72. Johnson, R. C. *et al.* Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* **11**, 724 (2010).

73. Goeman, J. J. & Solari, A. Multiple hypothesis testing in genomics. *Stat. Med.* **33**, 1946–1978 (2014).

74. Park, J.-H. *et al.* Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl. Acad. Sci.* **108**, 18026–18031 (2011).

75. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits

shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

76. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

77. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2014).

78. Wray, N. R. *et al.* Research Review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry Allied Discip.* **55**, 1068–1087 (2014).

79. Li, M. X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**, 747–756 (2012).

80. Davis, J. R. *et al.* An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. *Am. J. Hum. Genet.* **98**, 216–224 (2016).

81. Yang, J. *et al.* Ubiquitous Polygenicity of Human Complex Traits: Genome-Wide Analysis of 49 Traits in Koreans. *PLoS Genet.* **9**, (2013).

82. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **10**, 8192–8192 (2009).

83. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segrè AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Rand, H. J. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).

84. Eleftherohorinou, H. *et al.* Pathway Analysis of GWAS Provides New Insights into Genetic Susceptibility to 3 Inflammatory Diseases. *PLoS One* **4**, e8068 (2009).

85. Eleftherohorinou, H., Hoggart, C. J., Wright, V. J., Levin, M. & Coin, L. J. M. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. *Hum. Mol. Genet.* **20**, 3494–3506 (2011).

86. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

87. Shrestha, S. *et al.* A systematic review of microRNA expression profiling studies in human gastric cancer. *Cancer Med.* **3**, 878–888 (2014).

88. Schubert, K. O., Föcking, M. & Cotter, D. R. Proteomic pathway analysis of the hippocampus in schizophrenia and bipolar affective disorder implicates 14-3-3 signaling, aryl hydrocarbon receptor signaling, and glucose metabolism: Potential roles in GABAergic interneuron pathology. *Schizophr. Res.* **167**, 64–72 (2015).

89. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44**, D481–D487 (2016).

90. Mooney, M. A. & Wilmot, B. Gene set analysis: A step-by-step guide. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **168**, 517–527 (2015).

91. Baker, E. *et al.* POLARIS: Polygenic LD-adjusted risk score approach for set-based analysis of GWAS data. *Genet. Epidemiol.* **42**, 366–377 (2018).

92. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24**, R111–R119 (2015).

93. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, e1003348 (2013).

94. Palla, L. & Dudbridge, F. A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait. *Am. J. Hum. Genet.* **97**, 250–259 (2015).

95. Dudbridge, F., Pashayan, N. & Yang, J. Predictive accuracy of combined genetic and environmental risk scores. *Genet. Epidemiol.* **42**, 4–19 (2018).

96. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

97. GWAS to the people. *Nat. Med.* **24**, 1483 (2018).