Spring 2022

Statistical Methods for Genetic Prediction of Complex Traits in Single and Multiple Populations

Geyu Zhou

Yale University Graduate School of Arts and Sciences, eldronzhou@gmail.com

Abstract

Statistical Methods for Genetic Prediction of Complex Traits in Single and Multiple Populations

Geyu Zhou

2022

Genetic prediction of complex traits, also known as polygenic risk score (PRS), is constructed by combining the estimated effect sizes of genetic markers across the genome for an individual. PRS has shown great promise in biomedical and clinical research for disease prevention, monitoring and treatment. However, the development of accurate prediction models is challenging due to the wide diversity of genetic architecture, limited access to individual level data, and the demand for computational resources. The broader application of PRS to the general population is further hindered by the poor transferability of PRS developed in Europeans to non-European populations.

In this thesis, we develop two statistical methods to help address these limitations. Chapter 1 includes a review of PRS from a statistical perspective. In Chapter 2, we present a summary statistics-based nonparametric method SDPR that is adaptive to different genetic architectures, statistically robust, and computationally efficient. The material is drawn from the manuscript “A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics” with minor modification. In Chapter 3, we develop a statistical method called SDPRX that can effectively integrate genome wide association study summary statistics from different

populations to improve the prediction accuracy in non-European populations. The material is drawn from the manuscript “SDPRX: A statistical method for cross-population prediction of complex traits” in preparation.

Statistical Methods for Genetic Prediction of Complex Traits in Single and Multiple Populations

A Dissertation

Presented to the Faculty of the Graduate School

Of

Yale University

In Candidacy for the Degree of

Doctor of Philosophy

By

Geyu Zhou

Dissertation Director: Hongyu Zhao

May 2022

Contents

List of Figures

List of Tables

Acknowledgements

I have spent about 5 years working on the topic of this thesis. I learnt a lot through this experience and find this journey rewarding and memorable. There are many people I wish to thank, for their contribution, mentorship, collaboration, personal experience and my time in New Haven.

First and foremost, I wish to thank my PhD advisor Hongyu Zhao. I came to Yale with no experience about how to develop statistical methods. Hongyu gave me the freedom to explore my interested topics, supervised me on all my work and taught me how to conduct research independently. I would not complete this work without Hongyu's patience, kindness and encouragement. I also appreciate Zhou and Yong-hui for volunteering their time to serve as my committee members. Zhou's expertise in statistics and Yong-hui's expertise is clinical genetics helps me improve the quality of this thesis.

I was privileged to mentor several students during my PhD. Special thanks to Zhu for being my first mentee, and Tianqi for helping me on part of simulation work in Chapter 3. Thanks to many friends at Zhao lab and Yale— Yuhan, Jianlei, Jiawei and Wei for discussion and collaboration; Wei and Yixuan for being excellent classmates, lab mates, and friends; Jiahao, Tianxiao, Rihao, Feng, Xiang, Zenan, Haoyang, Miao and Qi for the joy of food and games.

Finally, I am grateful to my parents, for their unconditional support and love on whatever I am doing.

In dedication to my grandfather.

Chapter 1

1.1 Introduction

How genetic factors contribute to the variation of phenotypes (traits) is a central question of genetic research. Mendelian traits, defined as traits following Mendelian inheritance laws (recessive or dominant), are usually controlled by a single gene or locus. In humans, single gene disorders are relatively well-studied as they have clear inheritance pattern with few contributing genetic factors. For example, mutations of the CFTR gene cause cystic fibrosis and excessive CAG repeats of the HTT gene cause Huntington disease [1, 2]. In contrast, most traits, regardless of discrete or continuous, are complex as they do not follow a simple Mendelian inheritance pattern and are influenced by many genetic factors. Complex traits are also affected by environmental factors. For example, diet composition and exercise frequency affect the lipid level and the risk of heart disease.

The study of complex traits dates back to the late 1800s. Francis Galton found that heights of children and their parents were correlated [3]. At that time, the cause of correlation was hard to explain as there was no direct link between Mendelian and complex traits. In 1918, R.A. Fisher proposed the famous “infinitesimal model” and showed that a trait value can be broken down as the sum of a genetic and non-genetic (environmental) component, with the genetic component being a large number of Mendelian factors (alleles of genes) with additive effects [4, 5]. Random sampling of a large number of Mendelian factors, each with small effect sizes on average, produces a normally distributed trait in the population. A toy example to illustrate this point would be assuming that each allele identically and independently follows a Bernoulli distribution with a unit effect size, then the sum of alleles tends to the normal distribution when the number of alleles is large. The infinitesimal model has been quite successful in modeling quantitative traits in plants and animal breeding, and is still useful in the modern genomics era.

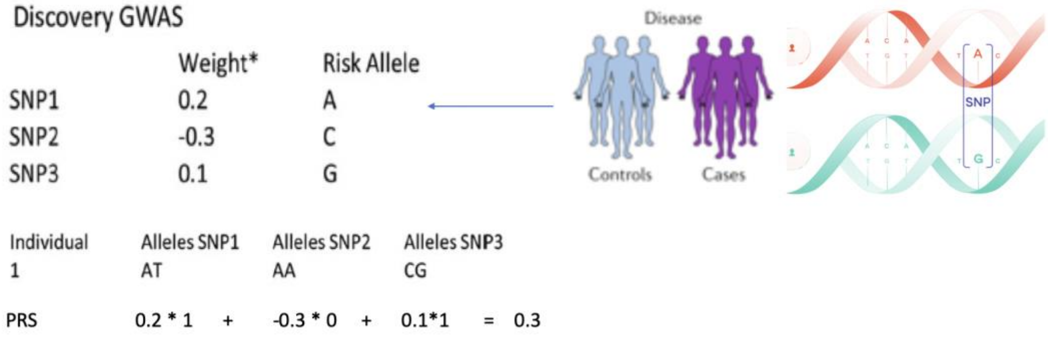
1.2 Genome-wide association studies

The completions of the Human Genome Project and the HapMap Project provided a reference map to study common genetic variations in human populations. It allows the design of high-throughput single-nucleotide polymorphism (SNP) array to measure the genotypes of an individual in a cost-effective way. A typical SNP array is able to simultaneously measure about 1 million markers of an individual across the genome, and imputation can further increase the number of markers up to around 100 millions. Consequently, it allows the experimental design of genome-wide association studies (GWAS), which helps dissect the genetic basis of complex traits.

GWAS starts by recruiting samples to form a cohort and collect their genotypes and phenotypes. A statistical test is then performed to assess the association of each SNP (a count of 0, 1, 2 of one of the two alleles) with the trait. The result is often recorded and made publicly available in the summary statistics format (Figure 1.1). Over the past decades, GWAS have been successful to identify a large number of SNPs associated with complex traits. The discovery power of GWAS increases with the sample size. As the sample size of GWAS increases, it reveals that the genetic architecture of most complex traits is polygenic, in the sense that they are associated with thousands of SNPs with relatively small effect sizes. For example, it is estimated that most 100Kb windows in the genome include variants that affect height and a randomly chosen 1Mb window contains variants that contribute to schizophrenia [6, 7].

1.3 Genetic prediction of complex traits

The data from GWAS can also be used to predict complex traits. Because most genotypes of an individual do not change across the lifetime, accurate genetic prediction can facilitate disease screening and prevention at an early stage or even long before the onset. Genetic prediction of complex traits is commonly referred as polygenic risk score (PRS) when the predicted trait is disease [8]. However, PRS can also be used to describe the prediction of continuous traits. The simplest way to construct PRS is a linear weighted sum of the dosage of genotypes by the estimated effect sizes (Figure 1.1) [8].



Discovery GWAS

	Weight*	Risk Allele
SNP1	0.2	A
SNP2	-0.3	C
SNP3	0.1	G

Individual	Alleles SNP1	Alleles SNP2	Alleles SNP3
1	AT	AA	CG

PRS 0.2 * 1 + -0.3 * 0 + 0.1*1 = 0.3

Figure 1.1. A simple illustration of GWAS and PRS. After collecting samples and performing the statistical test (linear regression for continuous traits and logistic regression for binary traits), the result is recorded in the summary statistics, which contain the identifier, tested allele and estimated effect size of each SNP. PRS of an individual is calculated as the weighted sum of the counts of alleles with weights being the estimated effect sizes.

The accuracy of PRS generally depends on two factors: heritability and the accuracy of estimated effect sizes. Heritability is the proportion of phenotypic variance that can be explained by genetic factors, which also serves as the theoretical upper bound of the prediction accuracy (proportion of phenotypic variance explained by PRS) [9]. The theoretical upper bound can only be achieved if all genetic markers affected the trait are known and their effect sizes are estimated without error [9]. These two conditions are neither satisfied based on current research progress.

On the one hand, SNPs on the high-throughput GWAS array are typically not the causal variants. Their associations are more likely because they are in linkage disequilibrium (LD) and thus tagged with the causal variants. Rare variants, presumably having large effect sizes, are usually not included or well-tagged in GWAS. Therefore, heritability estimated based on GWAS SNPs is smaller than the heritability of a trait. For example, heritability of height is believed to be 0.7-0.8 based on twin studies but the heritability estimated from GWAS SNPs is only around 0.5 [9]. Heritability estimated from GWAS SNPs thus serves as the upper bound of the prediction accuracy of PRS.

On the other hand, the effect sizes of genetic markers cannot be estimated without error as the sample size of GWAS is always finite. Currently, the number of GWAS SNPs is typically 1-10 million and the sample size of GWAS is 50-500K. Most complex traits are affected by a large number of SNPs with relatively small effect sizes. The variance of the estimated effect size through marginal linear regression is approximately reciprocal to the GWAS sample size. Hence, if the true effect size of one SNP is 10^{-3}, the accuracy of the estimated effect size is low even when the sample size of GWAS is 100K. The estimation error further propagates as effect sizes of many SNPs are aggregated to calculate PRS.

How to improve the performance of PRS is a key issue of PRS research. It can be foreseen that the prediction accuracy would increase as more GWAS with a large sample size are being conducted. Meanwhile, developing statistical methods that can achieve the best performance is also important.

1.4 Statistical methods

1.4.1 Multiple linear regression

Almost every statistical method of PRS can be understood in a multiple linear regression framework. We begin by defining y as an $N \times 1$ vector of the phenotypes measured in n individuals, and X as an $N \times M$ matrix with each column representing the genotypes of M SNPs in each individual. Genotypes are coded as the number of alleles (i.e. 0, 1, 2). We further assume that y and each column of X are normalized to have mean 0 and variance 1. The normalized genotypes no longer are coded as 0 ,1, 2. We note that the normalization usually does not significantly affect the prediction performance in the real data analysis. The following linear model connects y with X:

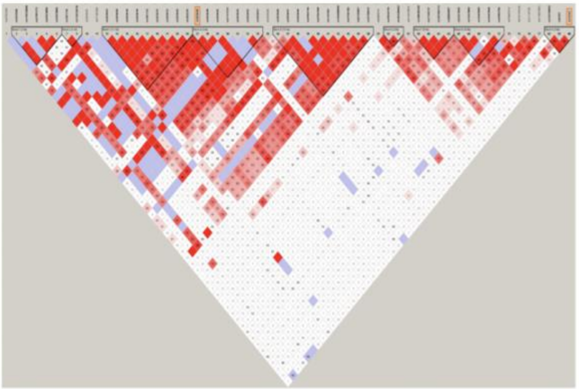
$$y = X\beta + \epsilon \tag{1.1}$$

where β is a $p \times 1$ vector of SNP effect sizes and ϵ is an $n \times 1$ vector of environmental effects. We assume that y is continuous for simplicity, though the discussion still applies to binary traits if we view y as the liability [10]. Because $N \ll M$ in the real GWAS setting, conventional maximal likelihood estimator (MLE) cannot be used to estimate β. Instead, most PRS methods make assumptions about β and fit the model using regularization or in a fully Bayesian way. We note that SNPs located closely to each other on the chromosome are often highly correlated due to linkage disequilibrium (LD). The estimator of β automatically adjusts for LD if the model is fitted in a joint way.

1.4.2 Summary statistics-based methods

Methods requiring the full knowledge of y and X are often referred as individual-level data-based, which were largely developed in the early GWAS era when the sample size was moderate. Individual-level data-based methods have two drawbacks, which limit their use in the current GWAS era. First, most studies would not make y and X publicly available due to privacy concerns. Second, the computational burden is intensive as typically N is 50-500K and M is 1-10 million for current GWAS.

Recent method development has shifted towards the use of summary statistics as most studies make their GWAS summary statistics publicly available (Figure 1.1). The discussion in the sections 1.2 and 1.3 are also based on summary statistics. For each SNP j, the weight $\hat{\beta}_j$ in the summary statistics are usually obtained by performing the marginal linear regression $\hat{\beta}_j = X_j^T y/N$ assuming normalized genotypes and phenotypes. Direct use of weights in the summary statistics (Figure 1.1) to construct PRS is problematic because the marginal regression does not adjust for LD and thus weights for SNPs in LD can be highly correlated or even the same (Figure 1.2).



$$\hat{\beta} = \frac{X^T y}{N}$$

X_1: 1 0 1 0 1 2

X_2: 1 0 1 0 1 2

y: 0 1 0 0 1 1

$$\widehat{\beta_1} = \widehat{\beta_2} = \frac{3}{6} = 0.5$$

Figure 1.2. The plot of linkage disequilibrium (LD) matrix. As shown in the red dots, SNPs that are located close to each other on the chromosome are highly correlated. If two SNPs X_1 and X_2 are perfectly correlated, then the marginal effect sizes $\hat{\beta}_1$ and $\hat{\beta}_2$ would be the same. This would cause the overcounting issue if no adjustment is applied.

Pruning and thresholding (P+T) is the simplest method to adjust for LD, which randomly selecting one of highly correlated SNPs based on LD and p value for calculation of PRS [8, 11]. P+T is currently the most popular PRS method for its simplicity and computational efficiency, though its prediction accuracy can often be improved by advanced statistical methods. Among these statistical methods, the most common way to adjust for LD is to introduce a LD matrix $R = X^T X/N$ and link the marginal effect sizes $\hat{\beta}$ with true effect sizes β through the following multivariate normal distribution.

$$\hat{\beta}|\beta \sim N(R\beta, R/N) \tag{1.2}$$

Similar to individual-level data-based methods, summary statistics-based methods also make assumptions about β for estimation, which will be discussed in the next section. We conclude this section with three remarks. First, because LD pattern is similar in a homogenous population, LD matrix R can be approximately estimated using a public reference panel without requiring the knowledge about X. Second, equation (1.2) may be violated if the approximation of the reference panel is not well or summary statistics are obtained via meta-analysis of heterogenous cohorts. Such violation often leads to the loss of prediction accuracy or even completely failure of the algorithm.

Third, individual-level data-based methods usually perform slightly better than the summary statistics-based methods for the same GWAS data. However, summary statistics methods applied to large scale GWAS often outperform individual-level data-based methods applied to small scale GWAS.

1.4.3 Assumptions

Most PRS methods differ in the assumptions made on the effect sizes. Here we review the assumptions for most commonly used methods and their connection with statistical literature. The list is by no means complete and we refer to the article by Ying and Xiang for a more compressive review [12].

One of the most successful statistical models applied in GWAS is the linear mixed model (LMM), which assumes that all SNPs are causal (non-zero) and their effect sizes follow a single normal distribution: $\beta_j \sim N(0, \sigma^2)$. Paired with equation (1.1), LMM was first introduced to estimate the SNP heritability h^2 using the restricted maximal likelihood (REML) approach [13]. Subsequently, the estimated heritability can be used to calculate the posterior effect sizes for prediction. Effect sizes estimated by LMM is equivalent as ridge regression (L2 regularization):

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \qquad (1.3)$$

where $\lambda = M(\frac{1}{h^2} - 1)$ and M is the number of SNPs. LMM can be easily extended to the use of summary statistics by replacing $X^T X$ and $X^T y$ with the corresponding terms of reference LD matrix R and marginal effect sizes β. LMM is implemented in the GCTA software and the summary statistics version is implemented in SBLUP and LDpred-inf [14-16]. In reality, not all SNPs have non-zero effect sizes, which violates the assumption of LMM. Although such violation does not affect the consistency of the estimation of heritability, the prediction accuracy can often be empirically improved with a sparse model assumption [17, 18]. Lasso is a classic frequentist method to perform variable selection and improve the prediction by imposing an L1 penalty on the effect sizes [19]. Lasso is implemented in the package snpnet, which is able to handle the computational challenge of individual-level data [20]. A summary statistics version is implemented in the lassosum package [21].

Another natural extension of LMM to incorporate sparsity is the point normal mixture model: $\beta_j \sim \pi N(0, \sigma^2) + (1 - \pi)\delta_0$. It assumes that only a portion (with probability π) of SNPs have non-zero effect sizes and these effect sizes follow a single normal distribution $N(0, \sigma^2)$. The remaining (with probability 1-π) SNPs have exactly zero effect sizes. Point normal mixture model is also known as Bayesian variable selection in the statistical literature [22]. The fitting of point normal mixture model can be computationally challenging due to the slow mixing of the Markov chain. Point normal mixture model is implemented in the software GEMMA for individual-level data and RSS/LDpred/LDpred2 for summary statistics [16, 18, 23, 24].

A limitation of the normal distribution is that its probability in the tail region quickly approaches zero and thus may over-shrink the large effect sizes. There are generally two ways to address this limitation. First, one may consider using a mixture of multiple normal distributions instead of a single one and thus allow multiple shrinkage for estimation. For example, BSLMM and DBSLMM assumes that effect sizes come from a mixture of two normal distributions $\beta_j \sim \pi N(0, \sigma_1^2) + (1-\pi) N(0, \sigma_1^2 + \sigma_2^2)$, while BayesR and SBayesR assume that effect sizes follow a mixture of point mass and three normal distributions $\beta_j \sim (1 - \pi_1 - \pi_2 - \pi_3)\delta_0 + \pi_1 N(0, 0.01\sigma^2) + \pi_2 N(0,\ 0.1\sigma^2) + \pi_3 N(0, \sigma^2)$ [25-27]. Second, another level of prior can be placed on the variance of normal distribution to induce a marginally heavy-tailed distribution. For example, BayesB places an inverse-gamma prior on the variance of normal distribution to induce a point t mixture distribution [28]. Motivated by the recent development of the horseshoe estimator for sparse signals, PRS-CS decomposes the variance of normal distribution as the product of a global scaling parameter and a Strawderman-Berger prior [29, 30].

1.5 Application

We conclude this chapter by a brief comment about the application of PRS. The arguably most important application of PRS is to predict the disease risk of individuals in a population. Although the prediction accuracy overall is low to moderate for most of diseases and quantitative traits, PRS is of great clinical interest due to its ability to identify individuals with high disease risk [31]. For example, Khera et al. demonstrated that PRS combined with sex and age was able to identify 1.5-8% of individuals in UK biobank with 3-fold increase risk of coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease and breast cancer [32]. Mavaddat et al. showed that the lifetime risk of overall breast cancer in the top centile of the PRS was 32.6% [33]. Another application of PRS is to serve as the instrumental variable to investigate the causal relationship between genetic predisposition of disease and traits. If hypertension is caused by obesity, then genetic variants linked to obesity will also affect hypertension, and thus individuals with a high PRS for obesity will on average have a higher blood pressure than those with a low PRS [34]. With the increase of GWAS sample size and development of PRS methods, it can be foreseen that there will be more applications of PRS with improved prediction accuracy.

Chapter 2

2.1 Introduction

Results from large-scale genome-wide association studies (GWAS) offer valuable information to predict personal traits based on genetic markers through polygenic risk scores (PRS) calculated from different methods. For one individual, PRS is typically calculated as the linear sum of the number of the risk alleles weighted by the effect size for each marker, such as single nucleotide polymorphism (SNP) [8]. PRS has gained great interest recently due to its demonstrated ability to identify individuals with higher disease risk for more effective prevention and monitoring [32].

Appropriate construction of PRS requires the development of statistical methods to jointly estimate the effect sizes of all genetic markers in an accurate and efficient way. Statistical challenges associated with the design of PRS methods largely reside in how to account for linkage disequilibrium (LD) among the markers and how to capture the genetic architecture of traits. Meanwhile, practical issues to be addressed include making use of summary statistics as input, as well as reducing the computational burden.

One simple method to compute PRS is to use a subset of SNPs in GWAS summary statistics formed by pruning out SNPs in LD and selecting those below a p value threshold (P+T) [8]. P+T is computationally efficient, though the prediction accuracy can usually be improved by using more sophisticated methods [16]. At present, most of the existing methods that allow the use of summary statistics as input assume a prior distribution on the effect sizes of the SNPs in the genome and fit the model under the Bayesian framework. Methods differ in the choice of the prior distribution. For example, LDpred and LDpred2 assume a point-normal mixture distribution or a single normal distribution [16, 18]. SBayesR assumes a mixture of three normal distributions with a point mass at zero [27]. PRS-CS proposes a conceptually different class of continuous shrinkage priors [30]. In reality, there is wide diversity in the distribution of effect sizes for complex traits [35]. Therefore, there may be model specification for choosing a specific parametric prior if the true genetic architecture cannot be captured by the assumed parametric distribution. A natural solution is to consider a generalizable nonparametric prior, such as the Dirichlet process [36]. Dirichlet process regression (DPR) was shown to be adaptive to different parametric assumptions and could achieve robust performance when applied to different traits [37]. However, DPR requires access to individual-level genotype and phenotype data and has expensive computational cost when applied to large-scale GWAS data.

In this work, we derive a summary statistics-based method, called SDPR, which does not rely on specific parametric assumptions on the effect size distribution. SDPR connects the marginal coefficients in summary statistics with true effect sizes through Bayesian multiple Dirichlet process regression. We utilize the concept of approximately independent LD blocks and overparameterization to develop a parallel and fast-mixing Markov Chain Monte Carlo (MCMC) algorithm [38, 39]. Through simulations and real data applications, we demonstrate the advantages of our methods in terms of improved computational efficiency and more robust performance in prediction without the need of using a validation dataset to select tuning parameters.

2.2 Methods

2.2.1 Robust design of the likelihood function

Suppose GWAS summary statistics are derived based on N individuals and p genetic markers, the phenotypes and genotypes can be related through a multivariate linear model,

$$y = X\beta + \epsilon \tag{2.1}$$

where y is an $N \times 1$ vector of phenotypes, X is an $N \times p$ matrix of genotypes, and β is an $p \times 1$ vector of effect sizes. We further assume, without loss of generality, that both y and columns of X have been standardized. GWAS summary statistics usually contain the per SNP effect size $\hat{\beta}$ directly obtained or well approximated through the marginal regression $\hat{\beta} = \frac{X^T y}{N}$. From this approximation, one can derive the commonly used likelihood function,

$$\hat{\beta}|\beta \sim N(R\beta, \frac{R}{N}) \tag{2.2}$$

where $R = \frac{X^T X}{N}$ is the reference LD matrix.

Unlike individual-level data based methods, summary statistics based methods typically rely on external reference panel to estimate the LD matrix R. Ideally, the same set of individuals in the reference panel should be used to generate the summary statistics. However, due to the limited access to the individual level data of original GWAS studies, an external database with matched ancestry like the 1000 Genomes Project [40] or UK Biobank [41] is usually used instead to compute the reference LD matrix. It is possible that effect sizes of SNPs in summary statistics deviate from what are expected given the likelihood function and reference LD matrix, especially for SNPs in strong LD that are genotyped on different individuals (Table 2.1). This issue was also noted in the section 5.5 of the RSS paper [24]. Failure to account for such discrepancy can cause severe model misspecification problems for SDPR and possibly other methods.

SNP	A1	A2	beta	se	p	N	r2
rs1206549	A	G	0.0093	0.0065	0.15	197888	0.99
rs712951	A	G	-0.0037	0.0041	0.41	252571	

Table 2.1. Illustration of the model misspecification issue using two SNPs in height GWAS summary statistics. In 1000G EUR samples, rs1206549 and rs712951 are in strong LD. However, their effect sizes in GWAS summary statistics were in the opposite direction, whereas the likelihood function $\hat{\beta}|\beta \sim N(R\beta, \frac{R}{N})$ would expect these two SNPs to have similar effect sizes. The discrepancy observed here may be explained by the fact that the imputed sample sizes of the two SNPs were different.

One can derive that, if SNPs are genotyped on different individuals, then the likelihood function (2.2) should be modified as

$$\hat{\beta}|\beta \sim N(R\beta, \mathrm{R} \circ H) \tag{2.3}$$

where ° is the Hadamard product, $H_{ii} = \frac{1}{N_i}$, $H_{ij} = \frac{N_{s,ij}}{N_i N_j}$ $(i \neq j)$, N_i is the sample size of SNP i, N_j is the sample size of SNP j, and $N_{s,ij}$ is the number of shared individuals genotyped for SNPs i and j (Appendix A). Evaluation of the likelihood function (2.3) requires the knowledge about the sample size and inclusion of each study for each SNP. For example, SNPs of GWAS summary statistics of lipid traits were genotyped on two arrays in two separate cohorts (GWAS chip: $N_1 \approx 95{,}000$; Metabochip: $N_2 \approx 94{,}000$) [43]. Based on this information, $N_{s,ij}$ is set to 0 if SNPs i and j were genotyped on different arrays, N_1 if SNP i was genotyped on GWAS chip and SNP j was genotyped on both arrays, and N_2 if SNP i was genotyped on Metabochip and SNP j was genotyped on both arrays.

In reality, GWAS summary statistics are often obtained through meta-analysis, and information above is generally not available. Besides, double genomic control is applied to many summary statistics, which may lead to deflation of effect sizes [44, 45]. Therefore, we consider evaluating the likelihood function from the following distribution.

$$\frac{\widehat{\beta}}{c}|\beta \sim N\left(R\beta, \frac{R + NaI}{N}\right) \tag{2.4}$$

More specifically, the input is divided by a constant provided by SumHer if application of double genomic control significantly deflates the effect sizes [44]. Compared with equation (2.3), the correlation between two SNPs is $\frac{R_{ij}}{1+Na}$ rather than $\frac{R_{ij} N_{s,ij}}{\sqrt{N_i N_j}}$. For simulated data, c was set to 1 and a was set to 0 for Scenarios 1A-1C, 4 and 5, since there was no above-mentioned discrepancy in these scenarios. In real data application, Na was set to 0.1 except for lipid traits, and c was set to 1 except for BMI (BMI $c = 0.74$ given by SumHer).

2.2.2 Dirichlet process prior

Like many Bayesian methods, we assume that the effect size of ith SNP β_i, follows a normal distribution with mean 0 and variance σ_β^2. In contrast to methods assuming one particular parametric distribution, we consider placing a Dirichlet process prior on σ_β^2, i.e.

$$\beta_i \sim N\left(0, \sigma_\beta^2\right), \sigma_\beta^2 \sim DP\left(H_\beta, \alpha\right) \qquad (2.5)$$

where H_β is the base distribution and α is the concentration parameter controlling the shrinkage of the distribution on σ^2 toward H. To improve the mixing of MCMC and avoid the informativeness issue of inverse gamma distribution, we expand the parameter $\beta_i = \eta\gamma_i$ and assign the following prior [46]:

$$\begin{gathered} \eta \sim N(0, A), \\ \gamma_i \sim N(0, \sigma^2), \\ \sigma^2 \sim DP(H, \alpha) \\ H = IG(a_{0k}, b_{0k}) \end{gathered} \qquad (2.6)$$

We note that the product $\eta\gamma_i$ in (2.6) corresponds to β_i in (2.5), and $|\eta|\sigma$ in (2.6) corresponds to σ_β in (2.5). We set $A = 10^6, a_{0k} = 0.5, b_{0k} = 0.5$ so that marginally the base distribution H_β in equation (2.5) is approximately the square of uniform distribution on $[0, +\infty)$. (If $\eta \sim N(0, A), \sigma^2 \sim IG(0.5, 0.5)$, then $\frac{\eta}{\sqrt{A}}\sigma \sim Cauchy(0,1)$ and $p(|\eta|\sigma) \propto 1$ as $A \to \infty$).

Under the truncated stick-breaking representation of Dirichlet process, the full model can be rewritten as:

$$\hat{\beta}|\gamma, \eta \sim N\left(R\eta\gamma, \frac{R + aNI}{N}\right)$$

$$\eta \sim N(0, 10^6)$$

$$\gamma_j|\sigma_k^2, p_k \sim \sum_{k=1}^{M} p_k N(0, \sigma_k^2)\, , k = 1, \dots, M, j = 1, \dots, p$$

$$p_k = V_k \prod_{m=1}^{k-1} (1 - V_m), \;\; V_k|\alpha \sim Beta(1, \alpha),$$

$$\sigma_k^2 \sim IG(0{,}5, 0.5), \alpha \sim Gamma(0.1{,}0.1) \tag{2.7}$$

We set M to 1000 as default for our methods. To assess whether our choice of M was a good approximation to the infinite stick-breaking process model, we counted the number of variance components to which SNPs were assigned. It turned out that SNPs were assigned to only 600 to 800 components in simulations and real data applications. After all, if the number of variance components is infinite, then some variance components will have no assignments of SNPs. Therefore, we believe that our choice of M was sufficient to approximate the Dirichlet process.

2.2.3 construction and partition of the reference LD matrix

We used an empirical Bayes shrinkage estimator to construct the LD matrix since the external reference panel like 1000G contains a limited number of individuals [47]. LD matrix can be divided into small "independent" blocks to allow for efficient update of posterior effect sizes using the blocked Gibbs sampler [30]. At present, ldetect is widely used for performing such tasks [38]. Ldetect works by computing the antidiagonal sum of the covariance matrix and applying the signal filtering approach to find the local minima in order to set the breakpoint. Originally developed to facilitate the interpretation of GWAS association signals, ldetect is not optimized for providing a precise partition for computation of PRS.

We used simulation data (Scenario 4) to assess the accuracy of LD blocks provided by ldetect. We generated marginal effect sizes and plotted them against the theoretical ones assuming the likelihood function $\hat{\beta}|\beta \sim N\left(R\beta, \frac{R}{N}\right)$. Unexpectedly, we found that the marginal effect sizes of some SNPs did not agree with the likelihood function (Figure 2.1). These SNPs were from a region (Chr10: 33 Mb) where ldetect cut the entire block into two independent ones (bottom left and upper right as separated by the cross). The cut was incorrect as SNPs in the second block had significant amount of correlation with SNPs in the first block. Therefore, some SNPs in the second blocks would have non-zero marginal effect sizes if they were in LD with the causal SNP in the first block, whereas theoretically they would have zero effect sizes assuming two blocks were independent.

To solve this issue, we designed a simple algorithm to ensure that each SNP in one LD block does not have nonignorable correlation ($r^2 > 0.1$) with SNPs in other blocks. Assuming SNPs are sorted based on their physical locations on the chromosome, for each SNP we recorded the index of the rightmost SNP with the nonignorable correlation using a sliding window. We then computed the cumulative maximum of the index along the list. We set the breakpoint at the SNP whose cumulative maximum index equals its original index. When applied to the example mentioned above, our algorithm did not cut the block and the marginal effect sizes were consistent with the theoretical ones (Figure 2.1). Compared with ldetect, overall our algorithm produced similar number of blocks. However, the number of the blocks containing more than 1000 SNPs was larger for our algorithm.

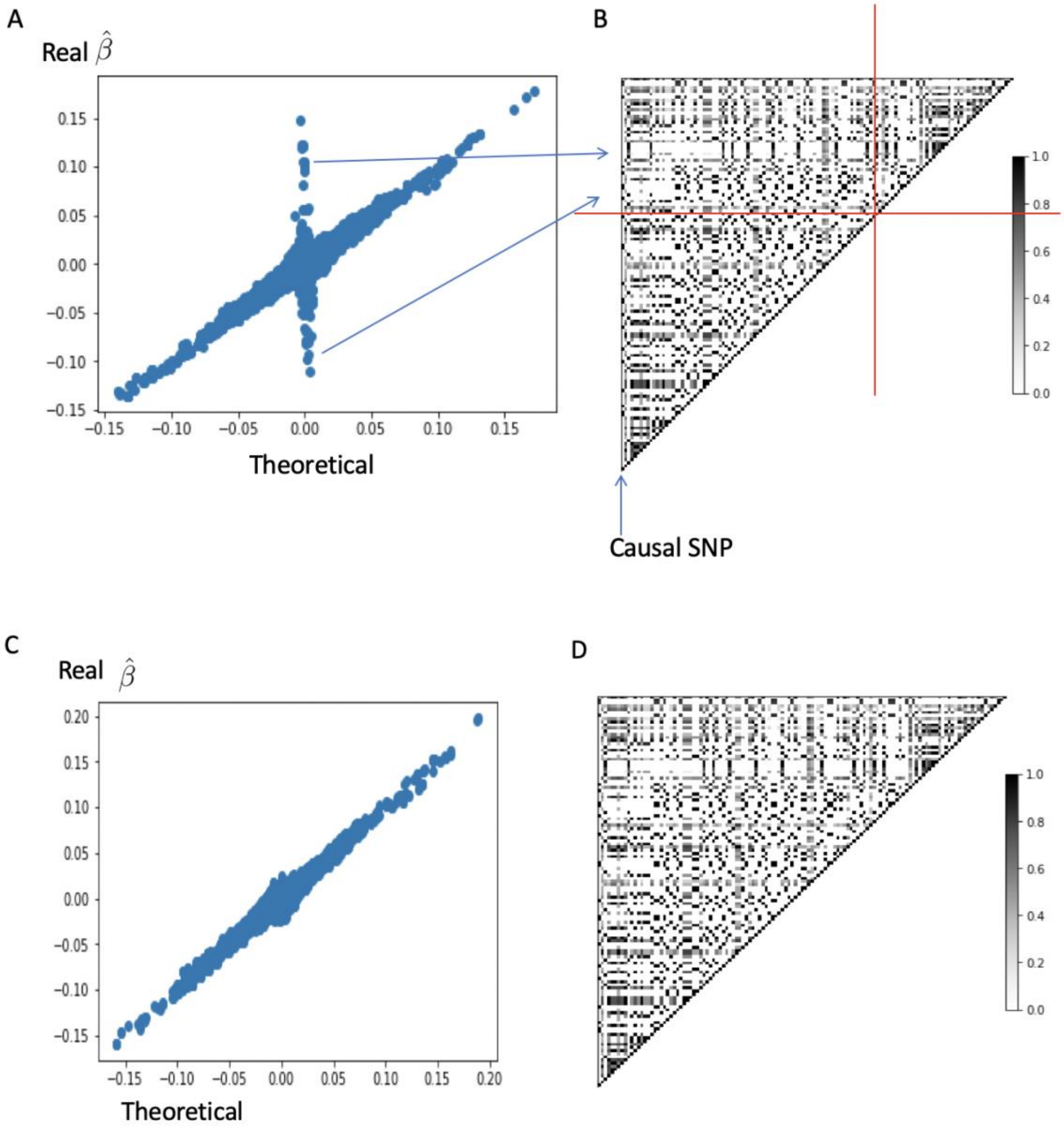


Figure 2.1. Comparison of Independent LD blocks defined by ldetect and SDPR. (A) Comparison of theoretical (using the partition by ldetect) and marginal effect sizes in GWAS summary statistics. (B) Correlation matrix of SNPs in the Chr10:33 Mb region of UK Biobank genotype data. Ldetect divided these SNPs into two independent blocks as separated by the red cross. The upper-left dots indicated that SNPs in two blocks had nonignorable correlation and some SNPs in block 2 were in LD with the marked causal

SNP. (C) Comparison of theoretical (using the partition by our method is correct) and marginal effect sizes in GWAS summary statistics. (D) Correlation matrix of SNPs in the Chr10:33 Mb region of UKB genotype data. Our method did not divide SNPs into two blocks.

2.2.4 MCMC algorithm

Here we describe our MCMC algorithm to obtain the posterior samples to estimate the effect sizes. We introduce a vector z indicating the assignment of variance component for each SNP.

Compute A, B: $(R/N + aI)A = R$ or $(R \circ H)A = R$, $B = RA$. If $R \circ H$ is not positive definite, we add $-1.1 \times$ its minimum eigenvalue to the diagonal to make it positive definite.

Sampling z_j: For each LD block, we first integrate out the effect size γ to derive the full conditional likelihood of $P\left(z_j = k \,\middle|\, .\,\right)$:

$$P\left(z_j = k \middle| .\right) \propto \int p\left(\hat{\beta} \middle| \gamma, \eta\right) p\left(\gamma_j \middle| z_j = k\right) d\gamma_j \;\times\; P\left(z_j = k\right)$$

$$\propto \int \exp\left\{-\frac{1}{2}\left(\hat{\beta} - \eta R\gamma\right)^T \left(\frac{R}{N} + aI\right)^{-1} \left(\hat{\beta} - \eta R\gamma\right)\right\} \frac{1}{\sigma_k} \exp\left\{-\frac{\gamma_j^2}{2\sigma_k^2}\right\} d\gamma_j \times p_k$$

$$\propto \int \exp\left\{-\frac{1}{2}\eta^2 \gamma^T B \gamma + \eta \hat{\beta}^T A \gamma\right\} \frac{1}{\sigma_k} \exp\left\{-\frac{\gamma_j^2}{2\sigma_k^2}\right\} d\gamma_j \times p_k$$

$$\propto \int \exp\left\{-\frac{1}{2}\eta^2 B_{jj}\gamma_j^2 - \eta^2 \sum_{i \neq j} B_{ij}\gamma_i\gamma_j + \eta \sum_i A_{ij}\hat{\beta}_i\gamma_j \right\} \frac{1}{\sigma_k} \exp\left\{-\frac{\gamma_j^2}{2\sigma_k^2}\right\} d\gamma_j \times p_k$$

$$\propto \int \exp\left\{-\frac{1}{2}\left(\eta^2 B_{jj} + \frac{1}{\sigma_k^2}\right)\gamma_j^2 + b_j\gamma_j\right\} d\gamma_j \;\times \frac{p_k}{\sigma_k}$$

$$\propto \frac{1}{\sqrt{\eta^2 B_{jj}\sigma_k^2 + 1}} \exp\left\{\frac{b_j^2}{2\left(\eta^2 B_{jj} + \sigma_k^{-2}\right)}\right\} p_k$$

$$(2.8)$$

where $b_j = \eta \sum_i A_{ij}\hat{\beta}_i - \eta^2 \sum_{i \neq j} B_{ij}\gamma_i$. We set the first variance component to 0 in analogous to Bayesian variable selection, and we have $P\left(z_j = 1\middle|.\right) \propto p_1$ as the integration equals 1 when γ_j is degenerated at 0. We use log-exp-sum trick to avoid numerical overflow. Note that because SNPs in different LD blocks are approximately independent, we can sample their assignments in parallel.

Sampling β: We jointly sample the effect size of causal SNPs γ_θ in one independent LD block. The full conditional likelihood of γ_θ is

$$p\left(\gamma_\theta \middle| z_j \neq 1,\ .\right) \propto \exp\left\{-\frac{1}{2}\eta^2\gamma^T B\gamma + \eta\hat{\beta}^T A\gamma\right\} \exp\left\{-\frac{1}{2}\gamma_\theta^T \Sigma_0^{-1}\gamma_\theta\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\eta^2\gamma_\theta^T B_\theta\gamma_\theta + \eta\hat{\beta}^T A_\theta\gamma_\theta\right\} \exp\left\{-\frac{1}{2}\gamma_\theta^T \Sigma_0^{-1}\gamma_\theta\right\}$$

$$= MVN(\eta\Sigma A_\theta^T\hat{\beta},\ \Sigma)$$

$$(2.9)$$

where $\Sigma = (\ \eta^2 B_\theta + \Sigma_0^{-1})^{-1}, \Sigma_0 = diag\left(\sigma_{z_1}^2, \dots, \sigma_{z_p}^2\right)$ for causal SNPs ($z_j \neq 1$). A_θ, B_θ are the submatrices by selecting columns corresponding to SNPs with non-zero effect sizes from matrices A, B. For SNPs whose variance components are 0 ($z_j = 1$), we simply set the posterior effect sizes to 0 as $p\left(\gamma_j \middle| z_j = 1, .\right) = 0$.

Sampling η: The full conditional likelihood is

$$p(\eta|.) \propto \exp\left\{-\frac{1}{2}\eta^2\sum\gamma_\theta^T B_\theta\gamma_\theta + \eta\sum\hat{\beta}^T A_\theta\gamma_\theta\right\}\exp\left\{-\frac{\eta^2}{2\times 10^{-6}}\right\}$$

$$= N\left(\frac{\sum\hat{\beta}^T A_\theta\gamma_\theta}{\sum\gamma_\theta^T B_\theta\gamma_\theta + 10^{-6}}, \frac{1}{\sum\gamma_\theta^T B_\theta\gamma_\theta + 10^{-6}}\right)$$

$$(2.10)$$

Sampling σ_k^2: The first variance component is always 0. The full conditional likelihood is

$$p(\sigma_k^2|.) \propto \prod_{j:z_j=k}\frac{1}{\sigma_k}\exp\left\{-\frac{\gamma_j^2}{2\sigma_k^2}\right\}\sigma_k^{-2(a_{0k}-1)}\exp\left\{-\frac{b_{0k}}{\sigma_k^2}\right\}$$

$$= IG(\frac{M_k}{2} + .5, \frac{\sum_{j:z_j=k}\gamma_j^2}{2} + .5)$$

$$(2.11)$$

where $M_k = \sum_j I(z_j = k)$ and I is the indicator function.

Sampling V_k: The full conditional likelihood is

$$p(V_k|.) \propto p_k^{M_k} \dots p_{M-1}^{M_{M-1}}\, p_M^{M_M} V_k^{1-1}(1 - V_k)^{\alpha-1}$$

$$\propto V_k^{M_k}(1 - V_k)^{M_{k+1}+\cdots+M_M+\alpha-1}$$

$$= Beta(1 + M_k, \alpha + \sum_{l=k+1}^{M} M_l\,)$$

$$(2.12)$$

for k = 1, …, M – 1. V_M equals 1 according to the definition of the truncated stick-breaking process.

Computing p_k: The prior probability can be computed as

$$p_1 = V_1$$

$$p_k = \prod_{l=1}^{k-1} (1 - V_l)\, V_k \;\; (k \geq 2)$$

Sampling α: The full conditional probability is

$$p(\alpha|.) \propto \prod_{l=1}^{M-1} \alpha(1 - V_l)^{\alpha-1} \alpha^{.1-1} \exp\{-.1 \times \alpha\}$$

$$= Gamma(0.1 + M - 1, 0.1 - \sum_{k=1}^{M-1} \log{(1 - V_k)})$$

(2.13)

We record the effect size $\beta = \eta\gamma$ and heritability $h^2 = \beta^T R \beta$ for each iteration and compute the average of all posterior samples as the final estimator.

2.2.5 Other methods

We compared the performance of SDPR with seven other methods: (1) PRS-CS as implemented in the PRS-CS software; (2) SBayesR as implemented in the GCTB software (version 2.02); (3) LDpred as implemented in the LDpred software (version 1.0.6); (4) P+T as implemented in the PLINK software (version 1.90) [48]; (5) LDpred2 as implemented in the bigsnpr package (version 1.6.1); (6) Lassosum as implemented in the lassosum package (version 0.4.5) [21]; and (7) DBSLMM as implemented in the DBSLMM package (version 0.21) [49]. We used the default parameter setting for all methods. For PRS-CS, the global shrinkage parameter was specified as {1e-6, 1e-4, 1e-2, 1, auto}. For SBayesR, gamma was specified as {0, 0.01, 0.1, 1} and pi was specified as {0.95, 0.02, 0.02, 0.01}. For LDpred, the polygenicity parameter was specified as {1e-5, 3e-5, 1e-4, 3e-4, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3 ,1, LDpred-Inf}. For P+T, SNPs in GWAS summary statistics were clumped for r^2 iterated over {0.2, 0.4, 0.6, 0.8}, and for p value threshold iterated over {5e-8, 5e-6, 1e-5, 1e-4, 5e-4, 1e-3, 1e-2, 0.04, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}. For LDpred2, we ran LDpred2-inf, LDpred2-auto and LDpred2-grid, and reported the best performance of three options. The grid of hyperparameters was set as non-sparse, p in a sequence of 21 values from 10^{-5} to 1 on a log-scale, and h^2 within {0.7, 1, 1.4} of h^2_{LDSC}. For lassosum, lambda was set in a sequence of 20 values from 0.001 to 0.1 on a log-scale, and s within {0.2, 0.5, 0.9, 1}. For DBSLMM, p value threshold was iterated within {10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}}, r2 was iterated within {0.05, 0.1, 0.15, 0.2, 0.25}, and h^2 was set as h^2_{LDSC}. We tuned the parameters for PRS-CS, LDpred, P+T, LDpred2, lassosum, and DBSLMM using the validation dataset.

2.2.6 Genome-wide simulations

We used genotypes from UK Biobank to perform simulations. UK Biobank's database contains extensive phenotypic and genotypic data of over 500,000 individuals in the United Kingdom [41]. We selected 276,732 unrelated individuals of European ancestry based on data field 22021 and 22006. A subset of these individuals was randomly selected to form the training, validation and test datasets. Training datasets contained 10,000, 50,000, and 100,000 individuals, while validation and test datasets contained 10,000 individuals. We applied quality control (MAF > 0.05, genotype missing rate < 0.01, INFO > 0.3, pHWE > 1e-5) to select 4,458,556 SNPs from the original ~96 million SNPs. We then intersected these SNPs with 1000G HM3 SNPs (MAF > 0.05) and removed those in the MHC region (Chr6: 28-34 Mb) to form a set of 681,828 SNPs for simulation.

To cover a range of genetic architectures, we simulated effect sizes of SNPs under four scenarios: (1)-(3) $\beta_j \sim \pi N\left(0, \frac{h^2}{M\pi}\right) + (1-\pi)\delta_0$, where $h^2 = 0.5, M = 681828, \pi$ equaled 10^{-4} (scenario 1A), 10^{-3} (scenario 1B) and 10^{-2} (scenario 1C); (4) $\beta_j \sim \sum_{i=1}^{3} \pi_i N(0, c_i\sigma^2) + (1 - \sum_{i=1}^{3} \pi_i)\delta_0$ where $c = (1, 0.1, 0.01),\ \pi = (10^{-4}, 10^{-4}, 10^{-2})$ with σ^2 calculated so that the total heritability equaled 0.5; (5) $\beta_j \sim N(0, \frac{h^2}{M})$. Importantly, scenario 1A-1C satisfied the assumption of LDpred/LDpred2, scenario 5 satisfied the assumption of LDpred-inf/LDpred2-inf, whereas scenario 4 satisfied the assumption of SBayesR. Phenotypes were generated from simulated effect sizes using GCTA-sim, and marginal linear regression was performed on the training data to obtain summary statistics using PLINK2 [14, 50]. In each scenario, we performed 10 simulation replicates.

We applied different methods on the training data, and used the 10,000 individuals in the validation dataset to estimate the LD matrix. Parameters for LDpred, P+T, PRS-CS, LDpred2, lassosum, and DBSLMM were also tuned using the validation data. We then evaluated the prediction performance on the test data by computing the square of Pearson correlation of PRS with simulated phenotypes.

2.2.7 Real data application using public summary statistics and UK biobank data

We obtained public GWAS summary statistics for 12 traits and evaluated the prediction performance of each method using the UK Biobank data. Individuals in GWAS do not overlap with individuals in UK Biobank. For this reason, we did not use the latest summary statistics of height and BMI [51]. To standardize the input summary statistics, we generally followed the guideline of LDHub to perform quality control on the GWAS summary statistics [52]. We removed strand ambiguous (A/T and G/C) SNPs, insertions and deletions (INDELs), SNPs with an effective sample size less than 0.67 times the 90th percentile of sample size. SNPs within the MHC region were removed except for IBD, since MHC region plays an important role in autoimmune diseases. The remaining SNPs were then intersected with 1000G HM3 SNPs provided in the PRS-CS reference panel.

For UK Biobank, we first selected unrelated European individuals as we did in simulations. We then applied quality control (MAF > 0.01, genotype missing rate < 0.05, INFO > 0.8, pHWE > 1e-10) to obtain a total of 1,114,176 HM3 SNPs. UK Biobank participants with six quantitative traits-height, body mass index (BMI), high-density lipoproteins (HDL), low-density lipoproteins (LDL), total cholesterol, and triglycerides-were selected based on relevant data fields (50 for height, 21001 for BMI, 30780 for LDL, 20760 for HDL, 20690 for total cholesterol, and 30870 for triglyceride). Selected participants were randomly assigned to form validation and test datasets, each composing half of the individuals. We used the first instance if multiple measurements were available.

For six diseases, cases were selected based on ICD code in the EHR and self-reported questionnaire (data field 20002). For coronary artery disease, cases were selected based on ICD-9 codes of 410.X, 411.0, 412.X, or 429.79 or ICD-10 codes of I21.X, I22.X, I23.X, I25.2, or self-reported myocardial infarction [32]. For breast cancer, cases were selected among female participants based on ICD-9 codes 174 or 174.9, or ICD-10 codes C50.X, or self-report history of breast cancer. For inflammatory bowel disease, cases were selected based on ICD-10 codes of K50.X, or ICD-9 codes of 555.X, or self-reported history of Crohn's disease, ulcerative colitis, and inflammatory bowel disease. Participants with self-reported history of immunological/system disorders were excluded from controls. For type 2 diabetes, cases were selected based on ICD-10 codes of E11.X, or ICD-9 codes of K51.X, or self-reported history of type 2 diabetes. Participants with self-reported history of diabetes were excluded from controls. For schizophrenia, cases were selected based on ICD-10 codes of F20.X, or ICD-9 codes of 295.X, or self-reported history of schizophrenia. Participants with self-reported history of neurobiology/eye/psychiatry disorders were excluded from controls. For bipolar, cases were selected based on ICD-10 codes of F31.X, or ICD-9 codes of 296.X, or self-reported history of type I and type II bipolar disorder. Participants with self-reported history of neurobiology/eye/psychiatry disorders were excluded from controls.

Validation dataset consisted of an equal number of cases and controls, the rest of which were assigned to the test dataset. Random assignments of individuals to validation and test datasets were repeated for 10 times.

For six quantitative traits, we reported the prediction R^2 of PRS (variance explained by PRS) defined as $R^2 = 1 - \frac{SS_1}{SS_0}$, where SS_0 is the sum of squares of the residuals of the restricted linear regression model with covariates (an intercept, age, sex, top 10 PCs of the genotype data), and SS_1 is the sum of squares of the residuals of the full linear regression model (covariates above and PRS). For six diseases, we reported the AUC of PRS only for better comparison of different methods.

2.2.8 Code availability

SDPR is available on https://github.com/eldronzhou/SDPR under the GPLv3 license. The scripts used for analysis in this paper are available on https://github.com/eldronzhou/SDPR_paper.

2.3 Results

2.3.1 Adaptiveness of Dirichlet Process prior

Theoretically, Dirichlet process as an infinite Gaussian mixture model is able to approximate any continuous parametric distribution, thus including other published parametric distributions as special cases [53]. For example, the density of Dirichlet process prior adapts well to the density of normal distribution (LDpred-inf), point normal mixture distribution (LDpred/LDpred2), and three-point normal mixture distribution (SBayesR) (Figure 2.2). Compared with SBayesR, Dirichlet process prior does not constrain the relationship between three non-zero normal variance components. We also explicitly incorporate Bayesian variable selection by setting the first variance component as 0, which is different from PRS-CS. The adaptiveness of Dirichlet process prior potentially makes it more robust to the distribution of effect sizes of real traits.

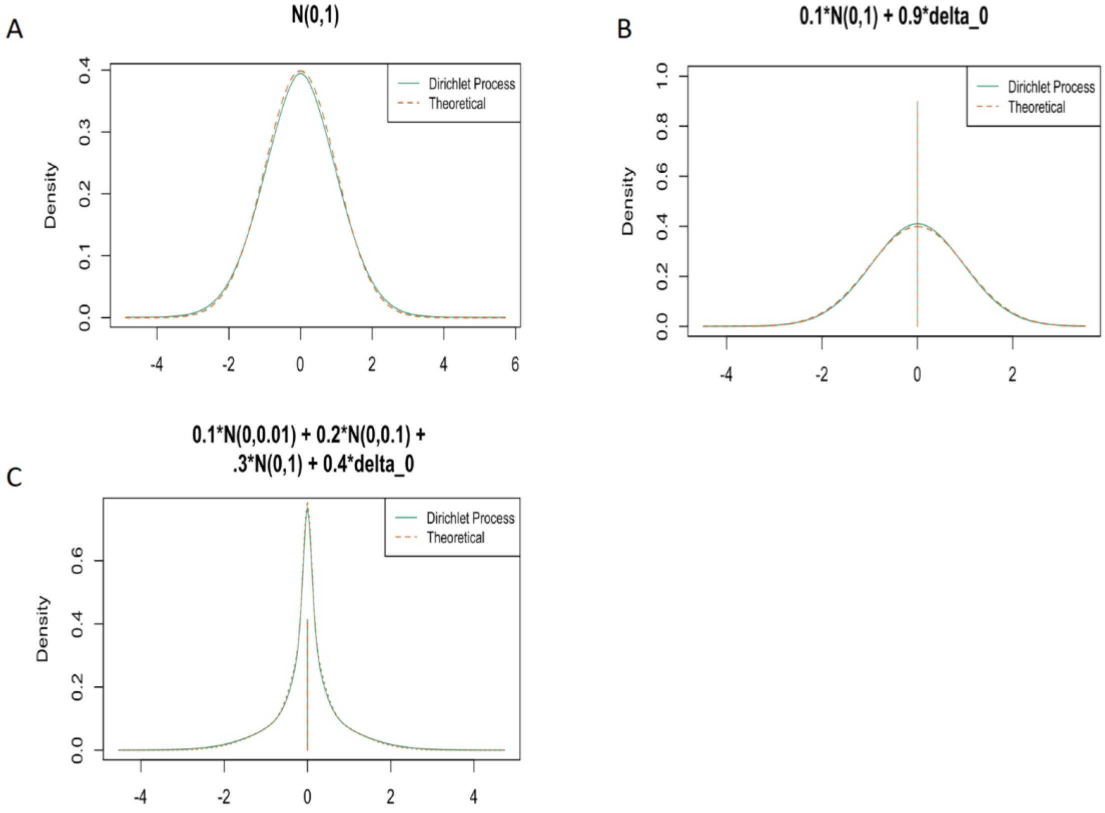


Figure 2.2. Adaptiveness of Dirichlet process prior to different parametric assumptions. A: 2000 data were simulated from $N(0,1)$ satisfying the assumption of LDpred-inf. B:2000 data were simulated from $0.1N(0,1) + 0.9\delta_0$ satisfying the assumption of LDpred. C: 2000 data were simulated from $0.1N(0, 0.01) + 0.2N(0, 0.1) + 0.3N(0,1) + 0.4\delta_0$ satisfying the assumption of SBayesR. Theoretical and Dirichlet process fitted density was shown in the plot.

2.3.2 Simulations

We first compared the performance and computational time of SDPR with DPR in a small-scale simulation setting using 10,000 individuals and 58,432 SNPs on chromosome

1. The effect sizes were generated under the mixture of Dirichlet delta and three normal distributions with total heritability fixed as 0.3. We fitted DPR model with four components and 5000 MCMC iterations, and SDPR model with the input of summary statistics. The average R^2 of DPR was 0.227, and the average R^2 of SDPR was 0.204 (Figure 2.3). DPR took about 3.5 hours and consumed 10.4 Gb of memory to finish MCMC, while SDPR took only 10 minutes and used 1.1 Gb of memory. This demonstrated the improved computational efficiency of SDPR over DPR without loss of much prediction accuracy.

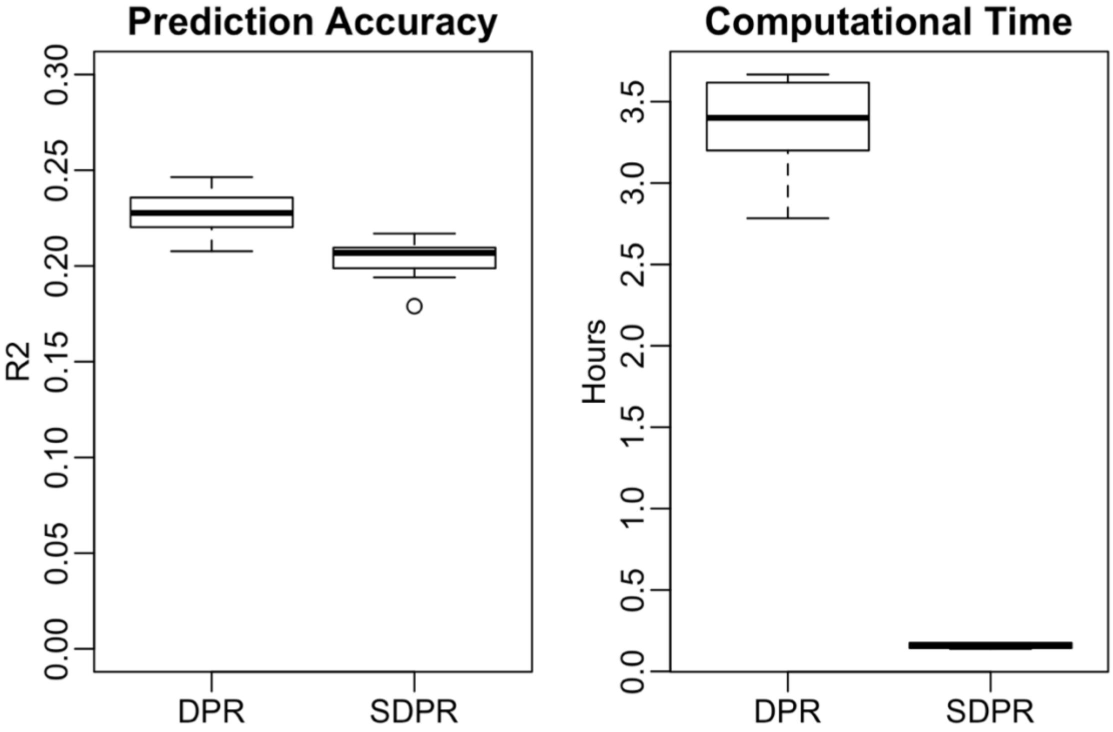


Figure 2.3. Performance and Computational time of SDPR with DPR under a small-scale simulation. Effect sizes were generated as $\beta_j \sim \sum_{i=1}^{3} \pi_i N(0, c_i\sigma^2) + (1 - \sum_{i=1}^{3} \pi_i)\delta_0$ where $c = (1, 0.1, 0.01),\ \pi = (10^{-4}, 10^{-4}, 10^{-2})$ with σ^2 calculated so that the total

heritability equaled 0.3. DPR was fit with 4 normal components, 2000 burnin and 4000 sampling iterations. SDPR was fit with 1000 maximum components and 1000 iterations. Simulation in each scenario was repeated for 10 times.

We then compared the performance of SDPR with several other summary statistics-based methods via genome-wide simulations across different genetic architectures and training sample sizes. Effect sizes of SNPs were simulated under a point-normal mixture model with increasing number of causal variants, a point-three-normal mixture model satisfying SBayesR's assumption, and a normal model satisfying LDpred-inf's assumption (details in methods). The heritability was fixed as 0.5 and 10 replicates were performed in each simulation setting. Tuning parameters of PRS-CS, LDpred, P+T, LDpred2, lassosum, and DBSLMM were selected using a validation dataset (N = 10,000). 10,000 individuals in the validation dataset were used to construct the LD matrix. We evaluated the prediction performance on the independent test data (N = 10,000) using the squared Pearson correlation coefficient (R^2).

The prediction accuracy of all methods generally increased along the sample size of training data (Figure 2.4; Table 2.2-2.6). Similarly, all methods performed better when the number of causal variants was small. Since the standard error of the regression coefficient estimator in GWAS summary statistics is roughly reciprocal to the square root of the sample size of the training cohort, the dominance of noise over signal poses significant challenges for accurate estimation of effect sizes when the training sample size or per SNP effect size is small.

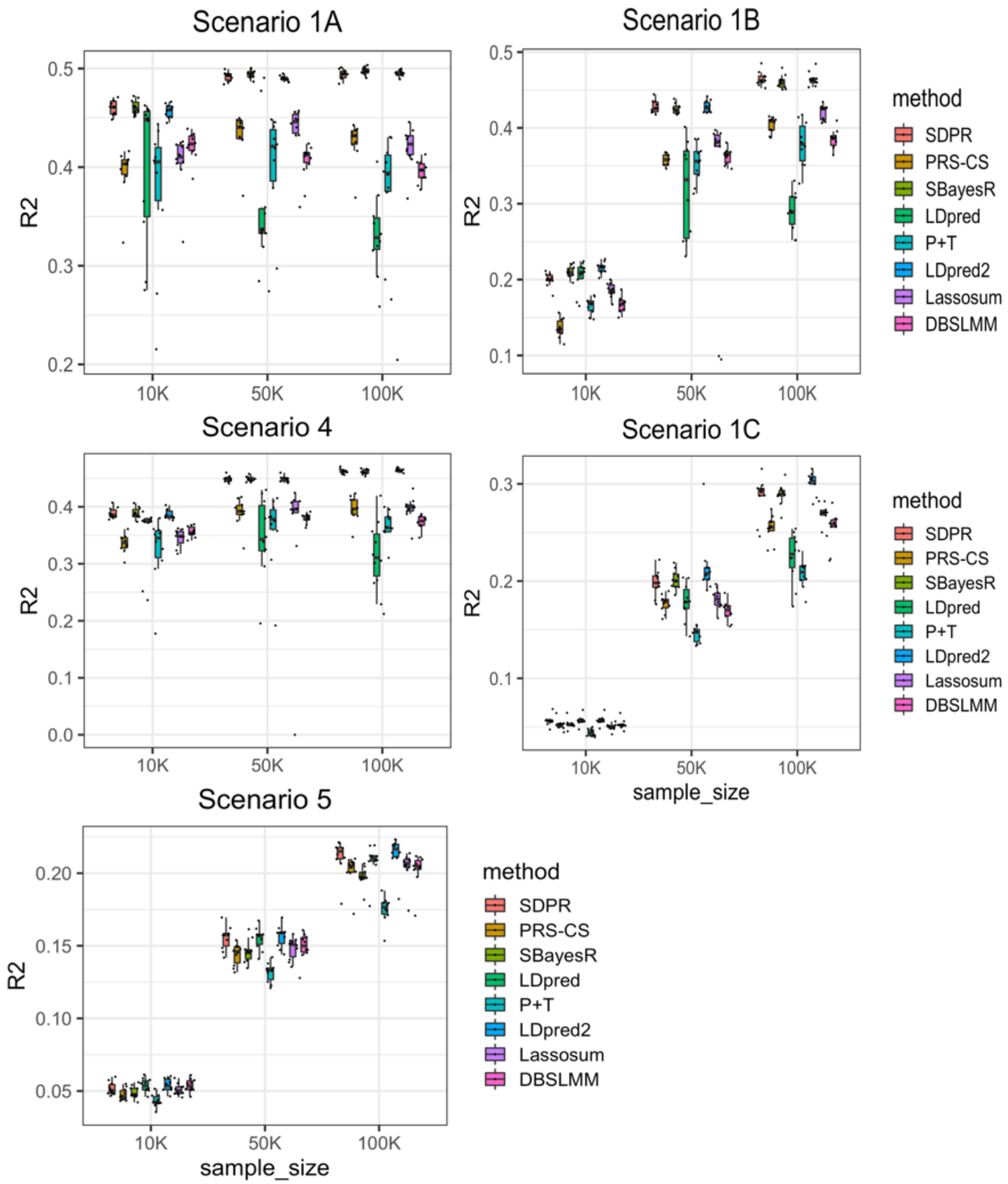


Figure 2.4. Prediction performance of different methods on simulated data with varying samples sizes of the training cohort. Scenarios 1A-1C: mixture of Dirichlet delta and

normal distribution (spike and slab) with number of causal SNPs increasing from 100, 1000 to 10000. Scenario 4: mixture of Dirichlet delta and three normal distributions. Scenario 5: single normal distribution. The total heritability in all scenarios was fixed to 0.5. Simulation in each scenario was repeated for 10 times. For each boxplot, the central mark is the median and the lower and upper edges represents the 25th and 75th percentiles. The median is recorded in the Table 2.2-2.6.

Sample size	SDPR	PRS-CS	SBayesR	LDpred	P+T	LDpred2	Lassosum	DBSLMM
10K	0.461	0.393	0.459	0.448	0.405	0.458	0.410	0.423
50K	0.493	0.428	0.495	0.337	0.421	0.489	0.446	0.412
100K	0.494	0.424	0.497	0.328	0.384	0.495	0.423	0.397

Table 2.2. The median of square of Pearson correlation across 10 simulations for Scenario 1A.

Sample size	SDPR	PRS-CS	SBayesR	LDpred	P+T	LDpred2	Lassosum	DBSLMM
10K	0.200	0.137	0.209	0.209	0.168	0.217	0.186	0.168
50K	0.426	0.354	0.424	0.332	0.358	0.426	0.381	0.364
100K	0.462	0.405	0.459	0.289	0.378	0.463	0.425	0.386

Table 2.3. The median of square of Pearson correlation across 10 simulations for Scenario 1B.

Sample size	SDPR	PRS-CS	SBayesR	LDpred	P+T	LDpred2	Lassosum	DBSLMM
10K	0.056	0.052	0.053	0.056	0.043	0.056	0.050	0.052

50K	0.198	0.179	0.2	0.179	0.147	0.208	0.181	0.170
100K	0.293	0.254	0.289	0.278	0.209	0.305	0.270	0.259

Table 2.4. The median of square of Pearson correlation across 10 simulations for Scenario 1C.

Sample size	SDPR	PRS-CS	SBayesR	LDpred	P+T	LDpred2	Lassosum	DBSLMM
10K	0.385	0.335	0.386	0.375	0.345	0.383	0.348	0.355
50K	0.449	0.386	0.447	0.343	0.376	0.449	0.396	0.382
100K	0.462	0.389	0.461	0.311	0.363	0.463	0.399	0.375

Table 2.5. The median of square of Pearson correlation across 10 simulations for Scenario 4.

Sample size	SDPR	PRS-CS	SBayesR	LDpred	P+T	LDpred2	Lassosum	DBSLMM
10K	0.050	0.046	0.048	0.054	0.042	0.054	0.050	0.053
50K	0.157	0.146	0.146	0.157	0.133	0.159	0.151	0.150
100K	0.215	0.204	0.197	0.210	0.177	0.216	0.207	0.205

Table 2.6. The median of square of Pearson correlation across 10 simulations for Scenario 5.

SDPR, LDpred2, and SBayesR performed better than other methods in the sparse setting (Figure 2.4 Scenarios 1A-1C, 4; Table 2.2-2.5). Consistent with others' findings, we observed that when the genetic architecture was sparse, the performance of LDpred decreased as the training sample size increased [30]. In contrast, LDpred2 performed significantly better than LDpred. Meanwhile, PRS-CS performed worse when the training sample size was small. In the polygenic setting, SDPR and LDpred-inf/LDpred2-inf performed better than other methods (Figure 2.4 Scenario 5; Table 2.6). Overall, SDPR and LDpred2 performed well across a range of simulated sparse and polygenic genetic architectures. LDpred2 is expected to perform well in Scenarios 1A-1C and 5 since it satisfied the assumption of LDpred2/LDpred2-inf. The robust performance of SDPR demonstrates the advantage of using Dirichlet process prior to model the genetic architecture.

It is important to note that while SBayesR and SDPR do not need a validation dataset to tune parameters, they may be more susceptible to heterogeneity and errors in the summary statistics. Therefore, we tested whether our modified likelihood function (2.4) makes SDPR more robust when dealing with discrepancies between summary statistics and reference panel. We generated summary statistics from 50,000 individuals under the same setting as scenario 1B. For half of the SNPs (340,914), linear regression was performed on 40,000 individuals to obtain the marginal effect sizes. According to equation (2.3), the correlation of effect sizes of these SNPs would be 80% of what was expected from the reference panel. Such discrepancy indeed caused the divergence of SBayesR, while SDPR with modified likelihood function (2.4) converged and performed well (N = 50,000, Na = 0.25, R^2 = 0.422).

2.3.3 Real data applications

We compared the performance of SDPR with other methods in real datasets to predict six quantitative traits (height, body mass index, high-density lipoproteins, low-density lipoproteins, total cholesterol, and triglycerides) and six diseases (coronary artery diseases, breast cancer, inflammatory bowel disease, type 2 diabetes, bipolar, and schizophrenia) in UK Biobank. We obtained public GWAS summary statistics of these traits and performed quality control to standardize the input (details in Methods; Table 2.7). A total of 503 1000G EUR individuals were used to construct the reference LD matrix for SDPR, PRS-CS, LDpred, P+T, LDpred2, lassosum, and DBSLMM. For SBayesR, we used 5000 EUR individuals in UK Biobank to create the LD matrix (shrunken and sparse) instead, as it was reported to have suboptimal prediction accuracy when using 1000G samples [27].

Trait	GWAS sample size	GWAS ref	1KG HM3 & UKB & GWAS SNPs	UKB validation Sample size	UKB testing sample size
Height	252,230	[42]	885,791	138,066	138,066
BMI	233,766	[54]	886,654	137,921	137,920
HDL	94,288	[43]	868,645	37,774	37,774
LDL	89,866	[43]	868,179	40,807	40,807
Total Cholesterol	94,571	[43]	868,167	40,898	40,898
Triglycerides	90,989	[43]	86,8243	40,858	40,857
Coronary artery disease	61,294 (22,233/64,762)	[55]	814,337	4475/4475	4475/258,345
Breast Cancer	227,688 (122,977/105,974)	[56]	927,706	4539/4539	4539/133,649
Inflammatory bowel disease	32,372 (12,882/21770)	[57]	918,369	1840/1840	1839/198,815
Type 2 diabetes	156,109 (26,676/132,532)	[58]	974,907	7240/7240	7239/182,292
Bipolar	41,606 (20,129/21,524)	[59]	928,032	832/832	832/176,069
Schizophrenia	65,955 (33,426/32541)	[59]	941,216	223/223	223/203,471

Table 2.7. Summary information about the sample size and SNPs in GWAS summary statistics and UK Biobank datasets. For binary traits, effective sample size was used ($\frac{4*N_{case}*N_{control}}{N_{case}+N_{control}}$) and the validation datasets consisted of equal numbers of cases and controls. If the summary statistics included sample sizes for individual SNPs, the median of all SNPs passing QC was reported. For binary traits, the number of cases and controls were reported in the parenthesis.

For six continuous traits, the prediction performance was measured by variance of phenotype explained by PRS (Figure 2.5; Table 2.8). Overall, SDPR, PRS-CS and LDpred2 performed better than other methods, and there was minimal difference of these three methods. In terms of ranking, SDPR and PRS-CS performed best for height.

SDPR and LDpred2 performed best for BMI. SDPR performed best for HDL, LDL and total cholesterol, while PRS-CS performed best for triglycerides. We observed convergence issues when running SBayesR on these traits, and followed its manual to filter SNPs based on GWAS P-values and LD R-squared (--p-value 0.4 --rsq 0.9). The filtering approach improved the prediction performance of SBayesR, but it still failed to achieve the top tier performance. We suspect that the convergence issue of SBayesR was also caused by the violation of the likelihood assumption, similar to what we observed in the simulation. To address this issue, our approach of modifying the likelihood function might be better than the simple filtering approach used in SBayesR and P+T as it retained all SNPs for prediction.

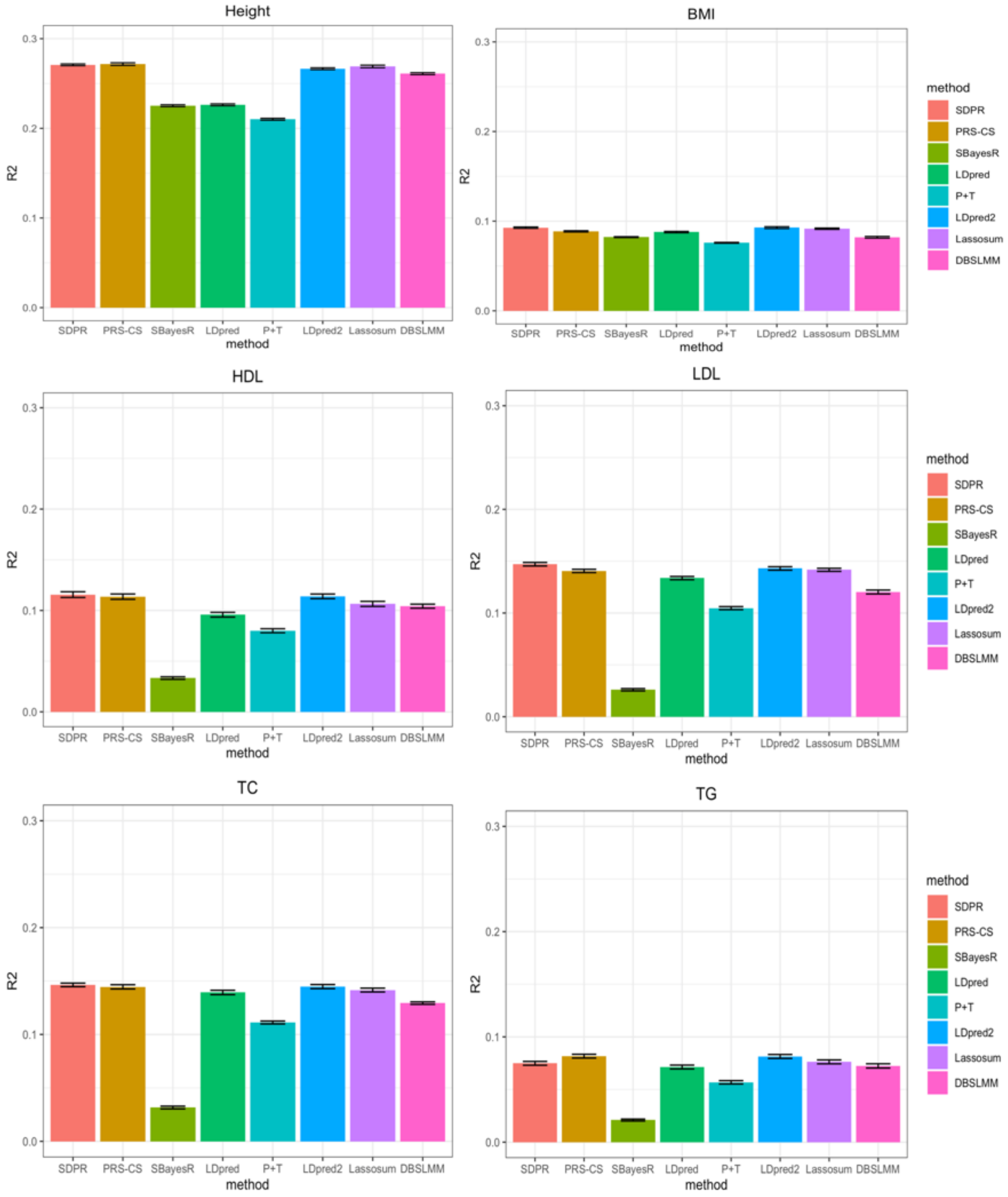


Figure 2.5. Prediction performance of different methods for six quantitative traits in the UK Biobank. Selected participants with corresponding phenotypes were randomly assigned to form validation and test dataset, each composing half of individuals. For PRS-CS, LDpred, P+T, LDpred2, lassosum, and DBSLMM, parameters were tuned based on the performance on the

validation dataset. We repeated the split and tuning process 10 times. The mean of variance of phenotypes explained by PRS across 10 random splits was reported in the Table 2.8.

Traits	SDPR	PRS-CS	SBayesR	LDpred	P+T	LDpred2	Lassosum	DBSLMM
Height	0.271	0.271	0.226	0.23	0.210	0.267	0.269	0.261
BMI	0.093	0.089	0.082	0.088	0.076	0.093	0.092	0.082
HDL	0.116	0.114	0.033	0.096	0.079	0.114	0.106	0.104
LDL	0.147	0.141	0.026	0.133	0.105	0.143	0.142	0.120
Total cholesterol	0.146	0.144	0.032	0.139	0.111	0.145	0.141	0.129
Triglycerides	0.075	0.082	0.021	0.071	0.057	0.081	0.076	0.072

Table 2.8. The mean of variance of phenotypes explained by PRS across 10 random splits for six quantitative traits.

For six disease traits, the prediction performance was measured by AUC of PRS only (Figure 2.6; Table 2.9). Overall, SDPR achieved top tier performance (within 0.003 difference of AUC of the best method) for five out of six diseases. In terms of ranking, LDpred and LDpred2 performed best for coronary artery disease. SDPR and PRS-CS performed best for breast cancer. LDpred2 performed best for IBD. For schizophrenia and type 2 diabetes, SBayesR performed best. LDpred, SDPR, LDpred2 and SBayesR performed best for bipolar.

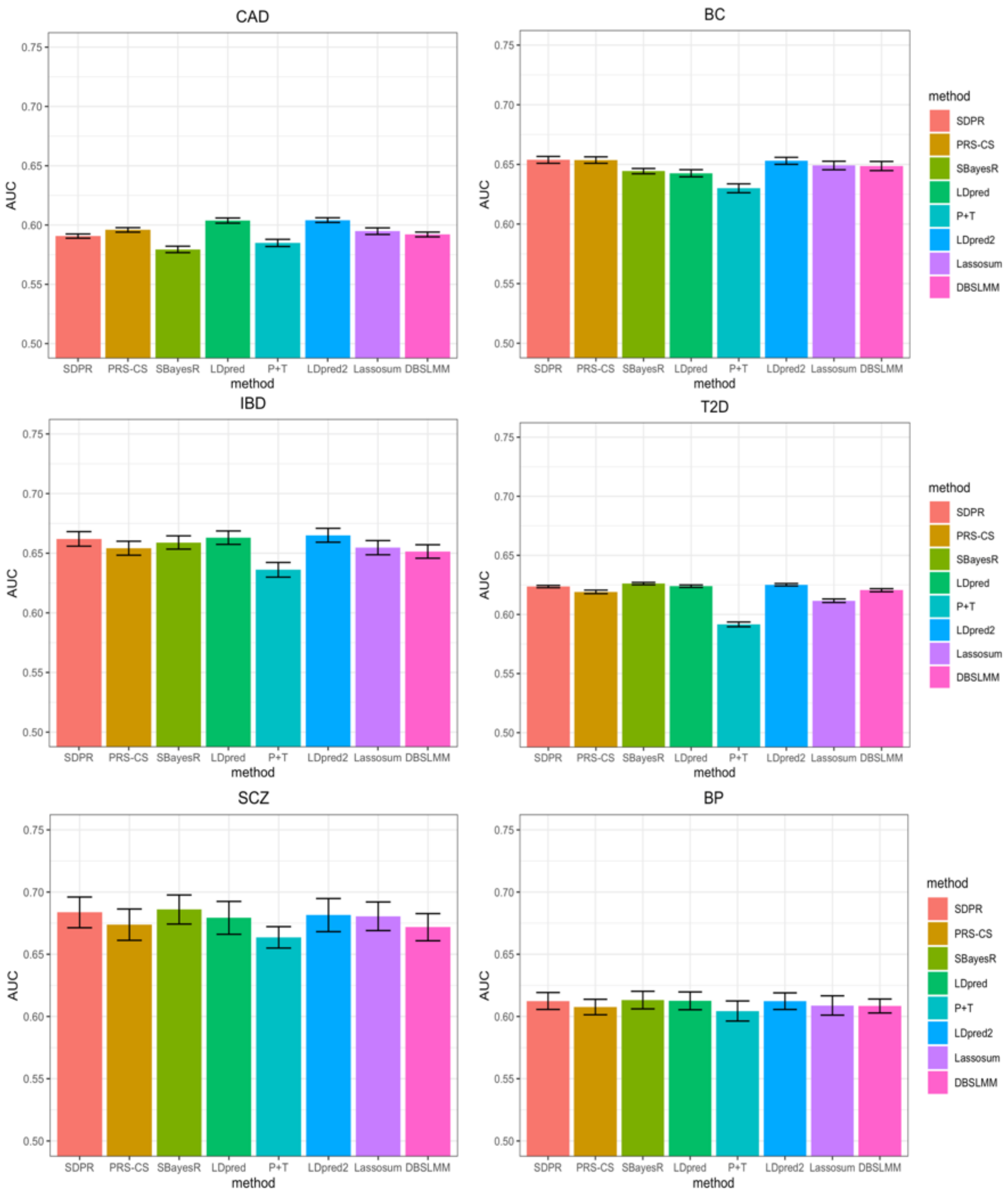


Figure 2.6. Prediction performance of different methods for 6 diseases in the UK biobank. Selected participants with corresponding diseases were randomly assigned to form validation and test dataset (Table 2.7). For PRS-CS, LDpred , P+T, LDpred2, lassosum and DBSLMM, parameters were tuned based on the performance on the

validation dataset. We repeated the split and tuning process for 10 times. The mean AUC across 10 random splits was reported in the Table 2.9.

Traits	SDPR	PRS-CS	SBayesR	LDpred	P+T	LDpred2	Lassosum	DBSLMM
CAD	0.591	0.596	0.579	0.604	0.584	0.604	0.594	0.592
BC	0.654	0.654	0.644	0.644	0.630	0.653	0.649	0.649
IBD	0.662	0.654	0.658	0.662	0.636	0.665	0.654	0.651
T2D	0.624	0.619	0.626	0.624	0.592	0.625	0.612	0.620
SCZ	0.684	0.673	0.686	0.679	0.664	0.681	0.680	0.672
BP	0.612	0.607	0.613	0.612	0.604	0.612	0.609	0.608

Table 2.9. The mean of AUC across 10 random splits for six diseases.

Consistent with simulations, SBayesR performed similarly to SDPR when there was no convergence issue (IBD, type 2 diabetes, schizophrenia, bipolar vs height, lipid traits). In general, PRS-CS performed better when the training sample size was large (height and breast cancer vs IBD and type 2 diabetes) and LDpred performed better when the training sample size was small (coronary artery disease, IBD vs height, breast cancer). LDpred2 performed significantly better than LDpred, achieving highly competitive performance. SDPR performed best among methods (PRS-CS auto, SBayesR, LDpred2 auto) without the need of parameter tuning (Table 2.10 and 2.11). Taken together, our design of the likelihood function and usage of Dirichlet process prior empowers SDPR with generally robust performance across different genetic architectures and training sample sizes.

Traits	SDPR	PRS-CS auto	SBayesR	LDpred2 auto
Height	0.271	0.271	0.226	0.224

BMI	0.093	0.089	0.082	0.070
HDL	0.116	0.114	0.033	0.063
LDL	0.147	0.141	0.026	0.113
Total cholesterol	0.146	0.144	0.032	0.074
Triglycerides	0.075	0.082	0.021	0.055

Table 2.10. The mean of variance of phenotypes explained by PRS across 10 random splits for six quantitative traits for methods without the need of parameter tuning.

Traits	SDPR	PRS-CS auto	SBayesR	LDpred2 auto
CAD	0.591	0.594	0.579	0.585
BC	0.654	0.652	0.644	0.617
IBD	0.662	0.652	0.658	0.616
T2D	0.624	0.619	0.626	0.602
SCZ	0.684	0.677	0.686	0.677
BP	0.612	0.610	0.613	0.610

Table 2.11. The mean of AUC across 10 random splits for six diseases for methods without the need of parameter tuning.

2.3.4 Computational time

SDPR is implemented in C++ to best utilize the resources of high-performance computing facilities. SDPR optimizes the speed of the computational bottleneck by using SIMD programming, parallelization over independent LD blocks, and high-performance linear algebra library. Besides, SDPR by default runs analysis on each chromosome in parallel because the genetic architecture may be different across chromosomes. We benchmarked the computational time and memory usage of each method on an Intel Xeon Gold 6240 processor (2.60 GHZ). For SDPR and PRS-CS, we paralleled computation over 22 chromosomes and used three threads per chromosome for the linear algebra library (22 ×3 = 66 threads in total). Time and memory usage were reported for the longest chromosome, which was the rate limiting step. For LDpred, SBayesR and P+T, no parallelization was used. LDpred2 was run in the genome-wide mode with 10 threads for parallel computation. DBSLMM and lassosum were run with 3 threads for parallel computation. The evaluation was based on a fixed number of MCMC iterations-1000 for SDPR and PRS-CS (default), 4000 for SBayesR (non-default but achieved generally good performance in simulations and real data application), 100 for LDpred (default), 1000 for LDpred2 (default). One should keep in mind that the number of MCMC iterations and threads for parallel computation affects the computation time significantly, though we did not explore it in this paper since each method also has different convergence and computational properties.

Table 2.12 shows that SDPR was able to finish the analysis in 15 minutes for most traits and required no more than 3 Gb of memory for each chromosome. SBayesR was also fast but the memory usage was significant for five diseases as no SNPs were removed to improve the convergence. The speed of PRS-CS, LDpred, P+T, LDpred2, lassosum, and DBSLMM was impeded by the need of iterating over tuning parameters. PRS-CS used less memory because the largest size of LD blocks output by ldetect was smaller compared with SDPR.

Trait	SDPR	PRS-CS	SBayesR	LDpred	P+T	LDpred2	Lassosum	DBSLMM
Height	0.20 (2.4)	2.5 (0.7)	0.92 (12.6)	5.0 (15.5)	0.6 (1.1)	5.5 (31.2)	0.50 (2.6)	1.7 (1.1)

BMI	0.18 (2.4)	2.8 (0.7)	0.50 (7.6)	4.9 (15.1)	0.5 (1.1)	5.4 (30.8)	0.45 (2.6)	0.60 (1.1)
HDL	0.20 (2.4)	1.6 (0.7)	0.68 (8.5)	5.1 (15.6)	0.5 (1.1)	3.9 (31.7)	0.41 (2.2)	0.44 (1.1)
LDL	0.22 (2.4)	2.2 (0.7)	0.67 (8.7)	5.1 (15.6)	0.6 (1.1)	5.5 (31.6)	0.42 (2.2)	0.61 (1.1)
Total cholesterol	0.25 (2.4)	2.2 (0.7)	0.48 (8.7)	5.1 (15.4)	0.5 (1.1)	4.1 (31.7)	0.40 (2.6)	0.60 (1.1)
Triglycerides	0.21 (2.4)	2.2 (0.7)	0.50 (8.3)	5.1 (15.5)	0.5 (1.1)	3.5 (31.6)	0.42 (2.6)	0.62 (1.1)
Coronary artery disease	0.23 (2.3)	1.9 (0.7)	0.39 (7.0)	4.7 (14.0)	0.3 (1.1)	3.5 (27.1)	0.33 (2.2)	0.77 (1.1)
Breast cancer	0.20 (2.9)	2.7 (0.7)	0.63 (42.3)	5.5 (16.4)	0.5 (1.1)	4.6 (37.1)	0.42 (2.7)	0.65 (1.1)
IBD	0.28 (2.8)	2.2 (0.7)	0.78 (39.5)	5.1 (16.0)	0.6 (1.1)	3.7 (33.4)	0.45 (2.7)	0.68 (1.1)
Type 2 diabetes	0.31 (2.9)	2.4 (0.7)	0.87 (47.4)	5.5 (17.4)	0.5 (1.1)	4.5 (37.2)	0.51 (2.8)	0.63 (1.2)
Schizophrenia	0.28 (2.7)	2.3 (0.7)	2.6 (42.1)	5.3 (16.4)	0.5 (1.1)	4.4 (36.8)	0.43 (2.3)	0.64 (1.1)
Bipolar	0.28 (2.8)	2.2 (0.7)	1.7 (43.8)	5.3 (16.3)	0.5 (1.1)	4.4 (36.8)	0.45 (2.6)	0.64 (1.1)

Table 2.12. Computational time and memory usage of different methods for 12 traits. The computational time is in hours. Memory usage of each method, as listed in the parenthesis, is measured in the unit of Gigabytes (Gb). We did not include the time of computing PRS in the validation and test datasets except for P+T, lassosum, LDpred2, and DBSLMM, because such computation was non-trivial for methods with a large grid of tuning parameters.

2.4 Discussion

Building on the success of genome wide association studies, polygenic prediction of complex traits has shown great promise with both public health and clinical relevance. Recently, there is growing interest in developing non-parametric or semi-parametric approaches that make minimal assumptions about the distribution of effect sizes to improve genetic risk prediction [37, 60, 61]. However, these methods either require access to individual-level data (DPR) [37], external training datasets (NPS) [60], or do no account for LD (So's method) [61]. Other widely used methods usually make specific parametric assumptions, and require external validation or pseudo-validation datasets to optimize the prediction performance [16, 21, 30]. To address the limitations of the existing methods, we have proposed a non-parametric method SDPR that is adaptive to different genetic architectures, statistically robust, and computationally efficient. Through simulations and real data applications, we have illustrated that SDPR is practically simple, fast yet effective to achieve competitive performance.

One of the biggest challenges of summary statistics-based method is how to deal with mismatch between summary statistics and reference panel. Based on our experience, misspecification of correlation of marginal effect sizes for SNPs in high LD can sometimes cause severe convergence issues of MCMC, especially for methods not relying on parameter tuning. Our investigation revealed that even when estimating LD from a perfectly matched reference panel, if SNPs were genotyped on different individuals, the correlation/covariance of marginal effect sizes in the summary statistics can be different from what is expected from the reference panel. We proposed a modified likelihood function to deal with this issue and observed improved convergence of MCMC. Our approach may be applied in a broader setting given that many summary statistics-based methods assume $\hat{\beta}\beta \sim N\left(R\beta, \frac{R}{N}\right)$ or $z|\beta \sim N(R\sqrt{N}\beta, R)$. When the sample size is small, the noise and heterogeneity of GWAS summary statistics poses more challenge for methods trying to learn every parameter from data (PRS-CS auto, LDpred2-auto, SBayesR, and SDPR). Under such circumstances, it is advantageous for methods like LDpred/LDpred2 to use an independent validation dataset to select the optimal parameters.

Although we have focused on the polygenic prediction of SDPR in this paper, it can provide estimation of heritability, genetic architecture, and posterior inclusion probability (PIP) for fine mapping. These issues will be fully explored in our future studies. SDPR can also be extended as a summary statistics-based tool to predict gene expression level for transcriptome wide association studies since a previous study has shown that individual level data based Dirichlet process model improves transcriptomic data imputation [62].

Although our method has robust performance in comparison with other methods, we caution that currently for most traits the prediction accuracy is still limited for direct application in clinical settings. From our perspective, there are three factors that affect the prediction accuracy. First, how much heritability is explained by common SNPs for diseases and complex traits? Second, if diseases or complex traits have relatively moderate heritability, is the GWAS sample size large enough to allow accurate estimation of effect sizes? Third, if the above two conditions are met, is a method able to have good prediction performance? The first two questions have been discussed in the literatures [10, 35, 63]. As for method development, we have focused on addressing the third question in this paper, and think SDPR represents a solid step in polygenic risk prediction.

Finally, we provide two technical directions for further development of SDPR. First, SDPR may have better performance after incorporating functional annotation as methods utilizing functional annotation generally perform better [64]. Second, studies have shown that PRS developed using EUR GWAS summary statistics does not transfer well to other populations [65, 66]. We can further modify the likelihood function to account for different LD patterns across populations to improve the performance of trans-ethnic PRS.

Chapter 3

3.1 Introduction

Polygenic risk score (PRS) of a complex trait for a given individual is constructed by combining the estimated effect sizes of genetic markers across the genome for this individual. PRS has received great interest recently due to its ability to identify individuals with higher disease risk for more effective population screening, diagnosis, and monitoring [32]. However, PRSs for most diseases to date have been primarily developed for Europeans as most well-powered genome wide association studies (GWAS) have been performed in cohorts of European ancestry. There can be substantial reduction in prediction accuracy when the PRSs derived from European samples are directly applied to non-European populations, leading to possible health disparities [65, 66].

The limited generalizability of PRS across different populations may be attributed but not limited to a number of factors. First, there is a lack of well-powered GWAS for training PRS models in the non-European populations. Second, the pattern of linkage disequilibrium (LD) and the tagging of causal variants can be different across populations. Third, the allele frequencies of variants vary between populations and some variants can even be population specific. As a general rule, the effect sizes of rarer variants are harder to estimate and GWAS with larger sample size are required in order to provide accurate estimates. Fourth, the effect sizes of one variant can be null (i.e. no effect), population specific (non-zero in one population) or correlated in two populations [67, 68]. Therefore, the effect sizes estimated from European GWAS may or may not be directly transferable to other populations.

Great efforts have been made in recent years to improve the genetic diversity of GWAS [69, 70]. Increased availability of GWAS summary statistics and biobank data from non-European ancestries creates an opportunity for developing novel methods to improve the accuracy of PRS in different populations. One general approach is to first estimate effect sizes in each population separately, and then derive a linear combination of the estimated effect sizes from a validation dataset of the target population [71]. Other approaches include jointly modeling GWAS summary statistics from multiple populations under the assumption that the causal variants are largely shared across populations [72, 73].

Here we propose SDPRX, an extension of SDPR [74], that integrates GWAS summary statistics and LD matrices from two populations with effect sizes under a hierarchical Bayesian model. SDPRX characterizes the joint distribution of the effect sizes of a SNP (single nucleotide polymorphism) in two populations to be both null, population specific or shared with correlation. We compared the performance of SDPRX with existing methods through extensive simulations and applications to seven traits in the East Asian (EAS) and African (AFR) individuals from the UK Biobank (UKB) [41]. We show that SDPRX improves the prediction accuracy in non-European populations over the existing methods.

3.2 Methods

3.2.1 Model of SDPRX

The relationship between marginal effect sizes in the summary statistics and true effect sizes can be modeled as

$$\begin{aligned} \hat{\beta}_1 \mid \eta, \beta_1 &\sim N(R_1\, \eta\beta_1, R_1/N_1 + aI) \\ \hat{\beta}_2 \mid \eta, \beta_2 &\sim N(R_2\, \eta\beta_2, R_2/N_2 + aI) \qquad (3.1) \end{aligned}$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the marginal effect sizes, R_1 and R_2 are the LD matrices, and N_1 and N_2 are GWAS sample sizes for populations 1 and 2, respectively. Compared with the commonly used assumption $\hat{\beta}|\beta \sim N(R\beta, R/N)$, the function above has two variations [74]. First, it shrinks the off-diagonal covariance by a constant identity matrix aI to avoid the "blow up" of effect sizes of SNPs in high LD due to the mismatch between GWAS summary statistics and reference panel. Second, it introduces a redundant parameter η so that the choice of hyperparameters of the prior on the variance components does not constrain the posterior inference [46].

We specify the following joint distribution as the prior on the effect sizes (β_{j1}, β_{j2}) of each SNP *j* in the two populations:

$$\begin{aligned} \begin{pmatrix} \beta_{j1} \\ \beta_{j2} \end{pmatrix} \sim p_0 \begin{pmatrix} \delta_0 \\ \delta_0 \end{pmatrix} + p_1 \sum_{k=1}^{1000} \pi_{1k} \begin{pmatrix} N(0, \sigma_{1k}^2) \\ \delta_0 \end{pmatrix} + \\ p_2 \sum_{k=1}^{1000} \pi_{2k} \begin{pmatrix} \delta_0 \\ N(0, \sigma_{2k}^2) \end{pmatrix} + \\ p_3 \sum_{k=1}^{1000} \pi_{3k}\, N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_{3k}^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \qquad (3.2) \end{aligned}$$

This prior characterizes the genetic architecture of one trait in two populations by a mixture of four mutually exclusive components. The first term describes the effect sizes of one SNP as zero in both populations. The second, third and fourth terms represent the effect sizes of one SNP as non-zero in population 1 only, non-zero in population 2 only, or non-zero and correlated in both populations. We note that if a SNP is only present in one population, it will be assigned to the first term (null), or one of the second and third terms (population specific).

We further assigned a Dirichlet distribution prior on the probability of each SNP to be null (p_0), population 1 specific (p_1), population 2 specific (p_2) and shared with correlation (p_3).

$$(p_0, p_1, p_2, p_3) \sim Dir(1) \tag{3.3}$$

For the second (population 1 specific), third (population 2 specific) and fourth terms (shared with correlation), we used the truncated stick-breaking process to represent the variance components and probability of assignments [75]. For example, for the second term (population 1 specific) we had:

$$V_{1k} \sim \ Beta(1, \alpha_1), k = 1, \dots, 1000$$

$$\pi_{11} \ = \ V_{11}$$

$$\pi_{1k} = \prod_{m=1}^{k-1} (1 - V_{1m}) V_{1k}, k = 2, \dots 1000$$

$$\sigma_{1k}^2 \sim \ IG(.5, .5)$$

$$\alpha_1 \sim \ Gamma(0.1, 0.1). \tag{3.4}$$

Finally, we set $N_1 a = N_2 a = 1$ and let $\eta \sim N(0,10^6)$ [74]. The cross-population genetic correlation ρ can be obtained from software like Popcorn [76]. In simulations, we set ρ to be the true value. To reduce the computational burden, SDPRX partitioned the LD matrix (element-wise maximum of LD matrices from two populations) into approximately independent LD blocks [74]. A MCMC algorithm was designed to fit the model (Section 3.2.2). In practice, we used 1000 MCMC iterations and the first 200 iterations as the burn-in. The computational time for the longest chromosome was around 5 hours. The mean of posterior effect sizes $\eta\beta_1$ and $\eta\beta_2$ were outputted as the adjust weights for two populations. When an independent validation dataset is available, one can also perform a convex combination of the output weights (α increased from 0 to 1 by a step of 0.05) and select the best α to further optimize the performance.

$$\beta_{target} = \alpha\beta_1 + (1-\alpha)\beta_2 \tag{3.5}$$

3.2.2 MCMC algorithm

Here we describe our MCMC algorithm based on Gibbs sampling to obtain the posterior samples. For each SNP j, we introduce a vector $z_j = (m,k), m \in \{0,1,2,3\}, k \in \{1,2,\dots,1000\}$ indicating whether effect sizes are population specific and which variance component it is assigned to. For example, z_j equals (1,4) if the effect sizes of SNP j are population 1 specific and it is assigned to the fourth variance component.

Compute A_1, B_1, A_2, B_2: $A_1 = (R_1 + N_1 aI)^{-1} R_1$, $B_1 = R_1 A_1$. $A_2 = (R_2 + N_2 aI)^{-1} R_2$, $B_2 = R_2 A_2$.

Sampling z_j: For each LD block, we first integrate out β_1 and β_2 to derive the conditional probability of SNP j whose effect sizes are correlated in two populations and assigned to the kth variance component:

$$P\big(z_j = (3,k)\big|.\big)$$

$$\propto \int\int p\big(\hat{\beta}_1\big|\beta_{1j},\eta\big)\, p\big(\hat{\beta}_2\big|\beta_{2j},\eta\big)\, p\big(\beta_{1j},\beta_{2j}\big|z_j = (3,k),\sigma_{3k}^2\big)\, d\beta_{1j}d\beta_{2j}$$

$$\times\ P\big(z_j = (3,k)\big)$$

$$\propto \int\int \exp\left\{-\frac{1}{2}\big(\hat{\beta}_1 - \eta R_1\beta_1\big)^T (R_1/N_1 + aI)^{-1}\big(\hat{\beta}_1 - \eta R_1\beta_1\big)\right\} \exp\left\{-\frac{1}{2}\big(\hat{\beta}_2\right.$$

$$- \eta R_2\beta_2\big)^T (R_2/N_2 + aI)^{-1}\big(\hat{\beta}_2$$

$$\left.- \eta R_2\beta_2\big)\right\} \frac{1}{2\pi\sigma_{3k}^2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{\beta_{1j}^2 + \beta_{2j}^2 - 2\rho\beta_{1j}\beta_{2j}}{\sigma_{3k}^2}\right]\right\} d\beta_{1j}d\beta_{2j}$$

$$\times \pi_{3k} p_3$$

$$\propto \int\int \exp\left\{-\frac{N_1}{2}\eta^2\beta_1^T B_1\beta_1 + N_1\eta{\hat{\beta}_1}^T A_1\beta_1\right\} \exp\left\{-\frac{N_2}{2}\eta^2\beta_2^T B_2\beta_2\right.$$

$$\left.+ N_2\eta{\hat{\beta}_2}^T A_2\beta_{12}\right\} \frac{1}{2\pi\sigma_{3k}^2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{\beta_{1j}^2 + \beta_{2j}^2 - 2\rho\beta_{1j}\beta_{2j}}{\sigma_{3k}^2}\right]\right\} d\beta_{1j}d\beta_{2j}$$

$$\times \pi_{3k} p_3$$

$$\propto \int\int \exp\left\{-\frac{N_1}{2}\eta^2 B_{1,jj}\beta_{1j}^2 - N_1\eta^2 \sum_{i\neq j} B_{1,ij}\beta_{1i}\beta_{1j}\right.$$

$$\left. + N_1\eta \sum_{i} A_{1,ij}\hat{\beta}_{1i}\beta_{1j} \right\} \exp\left\{-\frac{N_2}{2}\eta^2 B_{2,jj}\beta_{2j}^2 - N_2\eta^2 \sum_{i\neq j} B_{2,ij}\beta_{2i}\beta_{2j}\right.$$

$$\left. + N_2\eta \sum_{i} A_{2,ij}\hat{\beta}_{2i}\beta_{2j} \right\} \frac{1}{2\pi\sigma_{3k}^2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{\beta_{1j}^2+\beta_{2j}^2-2\rho\beta_{1j}\beta_{2j}}{\sigma_{3k}^2}\right]\right\} d\beta_{1j}d\beta_{2j}$$

$$\times \pi_{3k}p_3$$

$$\propto \int\int \exp\{-a_{jk1}\beta_{1j}^2 - a_{jk2}\beta_{2j}^2 + N_1 b_{1j} + N_2 b_{2j} + c_k\beta_{1j}\beta_{2j}\} \frac{1}{2\pi\sigma_{3k}^2\sqrt{1-\rho^2}} d\beta_{1j}d\beta_{2j}$$

$$\times \pi_{3k}p_3$$

$$\propto \frac{1}{\left(4a_{jk1}a_{jk2}-c_k^2\right)^{\frac{1}{2}}\sigma_{3k}^2} \exp\{a_{jk1}\mu_{jk1}^2 + a_{jk2}\mu_{jk2}^2 - c_k\mu_{jk1}\mu_{jk2}\} \times \frac{\pi_{3k}p_3}{\sqrt{1-\rho^2}} \tag{3.6}$$

where

$$b_{1j} = \eta\sum_{i} A_{1,ij}\hat{\beta}_{1i} - \eta^2\sum_{i\neq j} B_{1,ij}\beta_{1i}$$

$$b_{2j} = \eta\sum_{i} A_{2,ij}\hat{\beta}_{2i} - \eta^2\sum_{i\neq j} B_{2,ij}\beta_{2i}$$

$$a_{jk1} = \frac{N_1}{2}\eta^2 B_{1,jj} + \frac{1}{2\sigma_{3k}^2(1-\rho^2)}$$

$$a_{jk2} = \frac{N_2}{2}\eta^2 B_{2,jj} + \frac{1}{2\sigma_{3k}^2(1-\rho^2)}$$

$$\mu_{jk1} = \frac{2a_{jk2}N_1 b_{1j} + c_k N_2 b_{2j}}{4a_{jk1}a_{jk2} - c_k^2}$$

$$\mu_{jk2} = \frac{2a_{jk1}N_2b_{2j} + c_kN_1b_{1j}}{4a_{jk1}a_{jk2} - c_k^2}$$

$$c_k = \frac{\rho}{(1-\rho^2)\sigma_{3k}^2}$$

We next derive the conditional probability of SNP j whose effect sizes are population specific or null. It can be viewed as the special case to evaluate the last integrand by setting $\rho = 0, \beta_{2j} = 0$ (population 1 specific), $\rho = 0, \beta_{1j} = 0$ (population 2 specific), and $\beta_{1j} = \beta_{2j} = 0$ (both null).

$$P\big(z_j = (1,k)\big|.\big) \propto \frac{1}{\sqrt{N_1\eta^2B_{1,jj}\sigma_{1k}^2 + 1}} \exp\left\{\frac{N_1^2b_{1j}^2}{N_1\eta^2B_{1,jj} + \sigma_{1k}^{-2}}\right\} \times \pi_{1k}p_1$$

$$P\big(z_j = (2,k)\big|.\big) \propto \frac{1}{\sqrt{N_2\eta^2B_{2,jj}\sigma_{2k}^2 + 1}} \exp\left\{\frac{N_2^2b_{2j}^2}{N_2\eta^2B_{2,jj} + \sigma_{2k}^{-2}}\right\} \times \pi_{2k}p_2$$

$$P\big(z_j = (0{,}0)\big|.\big) \propto p_0$$

We use log-exp-sum trick to avoid numerical overflow. Note that because SNPs in different LD blocks are approximately independent, we can sample their assignments in parallel. For population 1 specific SNPs, we only need to evaluate $P\big(z_j = (1,k)\big|.\big)$ and $P\big(z_j = (0{,}0)\big|.\big)$.

Sampling β_1, β_2: For SNPs that are non-causal in any populations, we simply set the corresponding entries of β_1 and β_2 as zero. We then jointly sample the effect sizes of causal SNPs in one independent LD block. We introduce two indexes γ_1 and γ_2 such that β_{1,γ_1} and β_{2,γ_2} are non-zero. We combine β_{1,γ_1} and β_{2,γ_2} into one vector β_γ, which follows a bivariate normal distribution with mean 0 and variance-covariance matrix Σ_0.

The jth diagonal entry of Σ_0 is $\sigma_{z_j}^2$. If effect sizes of one SNP are non-zero with correlation in two populations, then $\Sigma_{0,ij} = \Sigma_{0,\mathrm{ji}} = \rho\sigma_{z_j}^2$. Other entries of Σ_0 are zero. Note that the special structure of Σ_0 allows an analytical solution of Σ_0^{-1}. We next derive the conditional likelihood as:

$$
\begin{aligned}
& p\left(\beta_{1,\gamma_1}, \beta_{2,\gamma_2} \middle| \,.\right) \\
& \propto \exp\left\{-\frac{N_1}{2}\eta^2\beta_1^T B_1\beta_1\right. \\
& \left. + \eta\hat{\beta}_1{}^T A_1\beta_1\right\} \exp\left\{-\frac{N_2}{2}\eta^2\beta_2^T B_2\beta_2 + \eta\hat{\beta}_2{}^T A_2\beta_2\right\} \exp\left\{-\frac{1}{2}\left(\beta_{1,\gamma_1}\ \beta_{2,\gamma_2}\right)^T \Sigma_0^{-1}\left(\beta_{1,\gamma_1}\ \beta_{2,\gamma_2}\right)\right\} \\
& \quad \propto \exp\left\{-\frac{1}{2}\eta^2\left(\beta_{1,\gamma_1}\ \beta_{2,\gamma_2}\right)^T \begin{pmatrix} N_1 B_{1,\gamma_1} & 0 \\ 0 & N_2 B_{2,\gamma_2} \end{pmatrix} \left(\beta_{1,\gamma_1}\ \beta_{2,\gamma_2}\right)\right. \\
& \qquad\qquad \left. + \eta\left(N_1\hat{\beta}_1^T A_{1,\gamma_1} \quad N_2\hat{\beta}_2^T A_{2,\gamma_2}\right)\right\} \exp\left\{-\frac{1}{2}\left(\beta_{1,\gamma_1}\ \beta_{2,\gamma_2}\right)^T \Sigma_0^{-1}\left(\beta_{1,\gamma_1}\ \beta_{2,\gamma_2}\right)\right\} \\
& \quad \propto \exp\left\{-\frac{1}{2}\eta^2\beta_\gamma^T B_\gamma\beta_\gamma + \eta\hat{\beta}^T A_\gamma\beta_\gamma\right\} \exp\left\{-\frac{1}{2}\beta_\gamma^T\Sigma_0^{-1}\beta_\gamma\right\} \\
& \quad = MVN(\eta\Sigma A_\gamma^T\hat{\beta}_\gamma,\ \Sigma)
\end{aligned}
\tag{3.7}
$$

where $\Sigma = \left(\eta^2 B_\gamma + \Sigma_0^{-1}\right)^{-1}$, $A_\gamma = \left(N_1\hat{\beta}_1^T A_{1,\gamma_1} \quad N_2\hat{\beta}_2^T A_{2,\gamma_2}\right)$, $B_\gamma = \begin{pmatrix} N_1 B_{1,\gamma_1} & 0 \\ 0 & N_2 B_{2,\gamma_2} \end{pmatrix}$.

A_{1,γ_1} is the submatrix by selecting columns from matrices A_1 based on the index γ_1. B_{1,γ_1} is the submatrix by selecting rows and columns from matrices B_1 based on the index γ_1.

Sampling η: The full conditional likelihood is

$p(\eta|.)$

$$\propto \exp\left\{-\frac{1}{2}N_1\eta^2\sum\beta_1^T B_1\beta_1\right.$$

$$\left.+N_1\eta\sum\hat{\beta}_1^T A_1\beta_1\ \right\}\exp\left\{-\frac{1}{2}N_2\eta^2\sum\beta_2^T B_2\beta_2+N_2\eta\sum\hat{\beta}_2^T A_2\beta_2\ \right\}\exp\left\{-\frac{\eta^2}{2\times 10^{-6}}\right\}$$

$$=N\left(\frac{N_1(\sum\hat{\beta}_1^T A_1\beta_1)+N_2(\sum\hat{\beta}_2^T A_2\beta_2)}{N_1(\sum\beta_1^T B_1\beta_1)+N_2(\sum\beta_2^T B_2\beta_2)+10^{-6}},\ \frac{1}{N_1(\sum\beta_1^T B_1\beta_1)+N_2(\sum\beta_2^T B_2\beta_2)+10^{-6}}\right)$$

$$(3.8)$$

Sampling $\sigma_{1k}^2, \sigma_{2k}^2, \sigma_{3k}^2$: The full conditional likelihood is

$$p(\sigma_{1k}^2|.)\propto\prod_{j:z_j=(1,k)}\frac{1}{\sigma_{1k}}\exp\left\{-\frac{\beta_{1j}^2}{2\sigma_{1k}^2}\right\}\sigma_{1k}^{-2(.5-1)}\exp\left\{-\frac{.5}{\sigma_{1k}^2}\right\}$$

$$=IG(\frac{M_{1k}}{2}+.5,\frac{\sum_{j:z_j=(1,k)}\beta_{1j}^2}{2}+.5)$$

$$p(\sigma_{2k}^2|.)\propto\prod_{j:z_j=(2,k)}\frac{1}{\sigma_{2k}}\exp\left\{-\frac{\beta_{2j}^2}{2\sigma_{2k}^2}\right\}\sigma_{2k}^{-2(.5-1)}\exp\left\{-\frac{.5}{\sigma_{2k}^2}\right\}$$

$$=IG(\frac{M_{2k}}{2}+.5,\frac{\sum_{j:z_j=(2,k)}\beta_{2j}^2}{2}+.5)$$

$$p(\sigma_{3k}^2|.)\propto\prod_{j:z_j=(3,k)}\frac{1}{\sigma_{3k}}\exp\left\{-\frac{\beta_{1j}^2+\beta_{2j}^2-2\rho\beta_{1j}\beta_{2j}}{2(1-\rho^2)\sigma_{3k}^2}\right\}\sigma_{3k}^{-2(.5-1)}\exp\left\{-\frac{.5}{\sigma_{3k}^2}\right\}$$

$$=IG(\frac{M_{3k}}{2}+.5,\frac{\sum_{j:z_j=(3,k)}\beta_{1j}^2+\beta_{2j}^2-2\rho\beta_{1j}\beta_{2j}}{2(1-\rho^2)}+.5)$$

$$(3.9)$$

where $M_{1k} = \sum_j I\left(z_j = (1,k)\right), M_{2k} = \sum_j I\left(z_j = (2,k)\right), M_{3k} = \sum_j I\left(z_j = (3,k)\right)$ and I is the indicator function.

Sampling $V_{mk}, m \in \{1,2,3\}, k \in \{1,2,\dots,1000\}$: The full conditional likelihood is

$$\begin{aligned} p(V_{mk}|.) &\propto V_k^{M_{mk}}(1-V_k)^{M_{m(k+1)}+\cdots+M_{1000}+\alpha_M-1} \\ &= Beta(1+M_{mk}, \alpha + \sum_{l=k+1}^{1000} M_{ml}\ ) \end{aligned} \tag{3.10}$$

for j=1,2,3 and k = 1, …, 999. V_{m1000} equals 1 according to the definition of the truncated stick-breaking process.

Computing $\pi_{mk}, m \in \{1,2,3\}$: The prior probability can be computed as

$$\begin{aligned} \pi_{m1} &= V_{m1} \\ \pi_{mk} &= \prod_{l=1}^{k-1}(1-V_{ml})\, V_{mk} \ \ (k \geq 2) \end{aligned}$$

Sampling p_0, p_1, p_2, p_3: The conditional distribution is:

$$p_0, p_1, p_2, p_3|. \sim Dir(M_0+1, M_1+1, M_2+1, M_3+1) \tag{3.11}$$

where $M_0 = \sum_j I\left(z_j = (0,0)\right), M_1 = \sum_j I\left(z_j = (1,.)\right), M_2 = \sum_j I\left(z_j = (2,.)\right), M_{3k} = \sum_j I\left(z_j = (3,.)\right)$. Note that we exclude population specific variants when computing M_0, M_1, M_2, M_3.

Sampling $\alpha_m, m \in \{1,2,3\}$: The full conditional likelihood is

$$p(\alpha_m|.) \propto \prod_{l=1}^{1000-1} \alpha_m (1 - V_{ml})^{\alpha_m - 1} \alpha_m^{.1-1} \exp\{-.1 \times \alpha_m\}$$

$$= Gamma(0.1 + 1000 - 1, 0.1 - \sum_{k=1}^{1000-1} \log{(1 - V_{mk})}) \tag{3.12}$$

We record the effect sizes $\eta\beta_1$ and $\eta\beta_2$ together with the heritability $h_1^2 = \beta_1^T R_1 \beta_1$ and $h_2^2 = \beta_2^T R_2 \beta_2$ for each iteration and compute the average of all posterior samples as the final estimator. We note that h_1^2 and h_2^2 together with the maximum of effect sizes can be used to assess whether the algorithm converges.

3.2.3 Existing methods

We compared the performance of SDPRX with three other methods: (1) PRS-CSx as implemented in the PRS-CSx software; (2) LDpred2 as implemented in the bigsnpr package; (3) XPASS as implemented in the XPASS package. For PRS-CSx, the global shrinkage parameter was specified as {1e-6, 1e-4, 1e-2, 1, auto}. For LDpred2, we ran LDpred2-inf, LDpred2-auto and LDpred2-grid, and reported the best performance of three options. The grid of hyperparameters was set as non-sparse, p in a sequence of 21 values from 10^{-5} to 1 on a log-scale, and h^2 within {0.7, 1, 1.4} of h^2_{LDSC}. For XPASS, population specific effects were included in both populations ($p < 10^{-10}$, clump_r2 = 0.1, clump_kb = 1000). In real data analysis, we also performed a linear regression on the validation dataset to learn the weights for combination of effect sizes.

3.2.4 Simulations

We first evaluated the prediction performance of each method via simulations across different genetic architectures and training sample sizes. We focused on four methods—SDPRX, PRS-CSx [73], LDpred2 [18] and XPASS [72]. To simulate individual-level genotypes from the 1000 Genomes Phase 3 haplotype, we first randomly selected 3,000 SNPs from the first 30,000 common SNPs (MAF > 0.05 in EAS, EUR and AFR) on chromosomes 1 to 10. The curated haplotypes reduced the computational burden of Hapgen2, while still provided a good representation of the real population structure. We then used Hapgen2 to simulate individual-level genotypes from the curated haplotypes. The simulated genotypes all passed the quality control (MAF > 0.05, genotype missing rate < 0.1, pHWE > 10^{-6}). The training cohort consisted of 40K EUR individuals and varying sample sizes (10K, 20K, 40K) of EAS and AFR individuals. The reduced sample size of non-EUR populations aligns with the fact that the sample size of most non-EUR GWAS is smaller than EUR GWAS. The validation and test datasets consisted of 5K individuals of each population.

The genetic architecture was simulated for two populations (EUR + EAS or EUR + AFR) as follows.

$$\begin{pmatrix}\beta_{j1}\\ \beta_{j2}\end{pmatrix} \sim (1-p_1-p_2-p_3)\begin{pmatrix}\delta_0\\ \delta_0\end{pmatrix} + p_1\begin{pmatrix}N\left(0,\frac{0.2h^2}{Mp_1}\right)\\ \delta_0\end{pmatrix} +$$
$$p_2\begin{pmatrix}\delta_0\\ N\left(0,\frac{0.2h^2}{Mp_2}\right)\end{pmatrix} +$$
$$p_3\, N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \frac{0.8h^2}{Mp_3}\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right),$$

where $h^2 = 0.3$, $M = 30{,}000$. Effect sizes of one SNP in two populations can be both zero, population specific (non-zero in population 1 or population 2) or correlated with the cross-population genetic correlation. We fixed the total heritability to be 0.3 and assumed that 80% of the total heritability was explained by SNPs with correlated effect sizes between the two populations. We considered three scenarios by increasing the proportions of population-specific and shared causal variates: (1) $p_1 = p_2 = p_3 = 0.0005$, (2) $p_1 = p_2 = p_3 = 0.005$, (3) $p_1 = p_2 = p_3 = 0.05$. We also varied the cross-population genetic correlation ρ in {0.4, 0.6, 0.8}. Phenotypes were then generated from simulated effect sizes using GCTA-sim, and marginal linear regression analysis was performed on the training data to obtain summary statistics using PLINK2 [14, 50]. Each simulation setting was repeated 10 times. The validation dataset was used to estimate LD matrix for each method and tune parameters for LDpred2 and PRS-CSx. The prediction performance was assessed by the square of Pearson correlation of PRS and simulated phenotype in the independent test dataset.

3.2.5 UK Biobank analysis

We downloaded GWAS summary statistics from GIANT, DIAGRAM, GLGC, BBJ, and PAGE consortia [42, 54, 58, 69, 70, 77-79]. We followed the guideline of LDHub to perform quality control on the GWAS summary statistics for each population [52]. We removed strand ambiguous (A/T and G/C) SNPs, insertions and deletions (INDELs), and SNPs with an effective sample size less than 0.67 times the 90th percentile of sample size. We did not restrict to SNPs present in two GWAS summary statistics so that population specific SNPs would be retained. Table 3.1 shows the number of SNPs present in the summary statistics for each trait after intersecting with reference panel and test dataset. The number of SNPs may not be optimal to achieve the best performance for each trait, but it did allow a fair comparison of different methods. We used the 1000 Genomes EUR, EAS and AFR samples as the LD reference panel for EUR, EAS and AFR (admixed populations for PAGE study) summary statistics respectively. For UK Biobank, we first performed principal component analysis (PCA) together with 1000 Genomes samples. We then trained a random forest classifier to assign UK Biobank samples to one of five super populations (EUR, EAS, AFR, SAS, AMR) based on top 10 PCs (Figure 3.1). We retained 2091 unrelated EAS and 6829 unrelated AFR samples with a predicted probability greater than 0.9 to form the validation and test datasets. We finally performed quality control (MAF > 0.01, genotype missing rate < 0.05, INFO > 0.8, pHWE > 1e-10) to obtain a total of 802,212 Hapmap3 (HM3) SNPs for EAS and 753,052 HM3 SNPs for AFR.

	GWAS sample size (EUR/EAS/AFR)	1KG HM3 &GWAS & UKB SNPs (EAS)	1KG HM3 & GWAS & UKB SNPs (AFR)	UKB EAS sample size (EAS)	UKB AFR sample size (AFR)
Height	252,230/159,095/49,781	523,930	433,973	2,081	6,727
BMI	233,766/158,284/49,335	536,830	433,973	2,078	6,715
HDL	885,540/116,404/90,804	515,898	433,846	398	1,610
LDL	840,006/79,693/87,559	545,790	433,967	440	1,710
TC	929,732/144,579/92,554	335,414	433,844	440	1,714
TG	860,547/81,071/89,467	545,812	433,858	440	1,714
T2D	156,109/191,764/14,480	538,716	433,973	1,263	4,809

Table 3.1. Summary of sample size and SNPs in GWAS summary statistics and UK Biobank datasets. The union of SNPs in GWAS summary statistics of two populations passing the quality control were intersected with the 1000 Genomes Hapmap3 reference panel and UK Biobank to form the final SNP list.

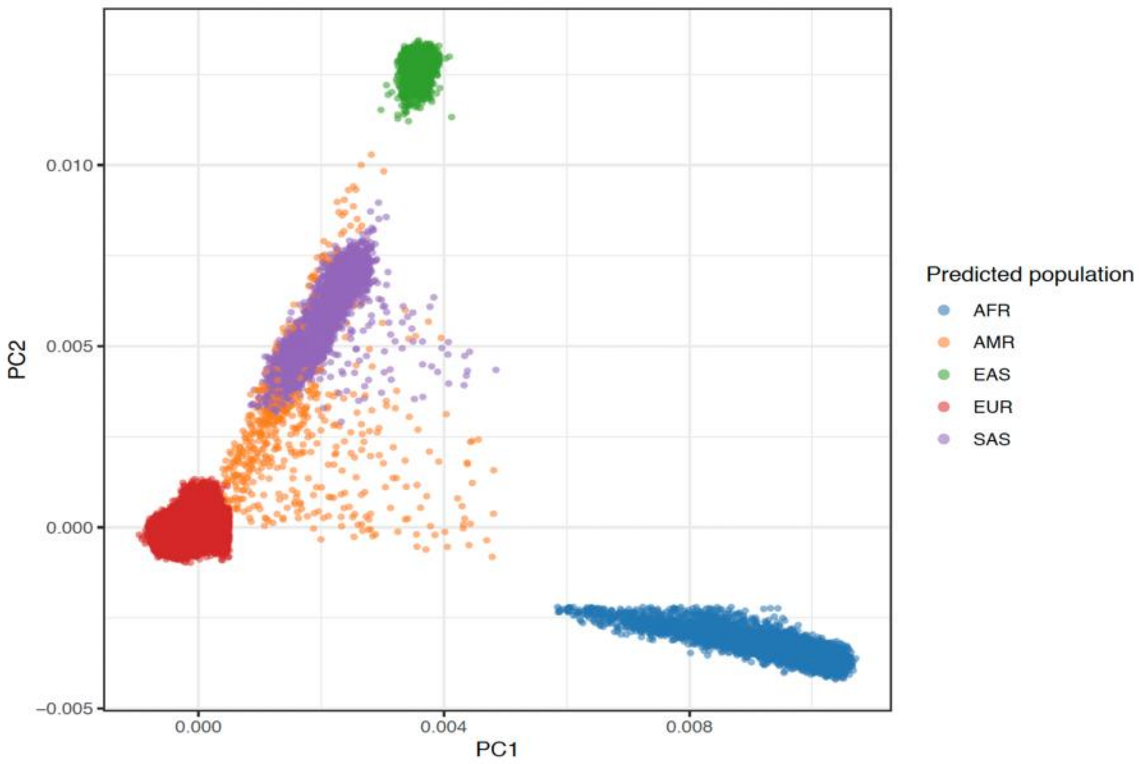


Figure 3.1. Principal component analysis of UK Biobank individuals. A random forest classifier was trained to assign each individual to one of five super populations (European, East Asian, African, Admixed American, South Asian) by a random forest classifier. We retained 2091 unrelated EAS and 6829 unrelated AFR samples with a predicted probability greater than 0.9 to form the validation and test dataset.

Phenotypes were selected based on the relevant data fields (50 for height, 21001 for BMI, 30780 for LDL, 20760 for HDL, 20690 for TC, 30870 for TG, and ICD-10 codes of E11.X, or ICD-9 codes of K51.X, or self-reported history of type 2 diabetes). For six quantitative traits, we reported the prediction R2 of PRS (variance explained by PRS) defined as $R^2 = 1 - \frac{SS1}{SS0}$, where $SS0$ is the sum of squares of the residuals of the restricted linear regression model with covariates (an intercept, age, sex, top 10 PCs of the genotype data), and SS1 is the sum of squares of the residuals of the full linear regression model (covariates above and PRS). For one binary trait, we reported the AUC of PRS only for better comparison of different methods.

3.2.6 Code availability

SDPRX is available at https://github.com/eldronzhou/SDPRX. The code used in this paper is available at https://github.com/eldronzhou/SDPRX_paper.

3.3 Results

3.3.1 Simulations

We focused on the results in EAS and AFR since our main purpose is to jointly utilize EUR GWAS data to improve the performance of PRS in non-EUR populations. Overall, all methods performed better as the proportion of causal SNPs decreased (Figure 3.2 and Table 3.2). Under a highly sparse genetic architecture (Scenario 1), the increase of sample size provided minimal benefits since the effect size per causal SNP was large enough for accurate estimation. In contrast, the improvement with an increasing sample size became apparent when the genetic architecture was polygenic (Scenario 3). Among all methods, XPASS did not perform well as the simulated data violated its assumption that all SNPs are causal. LDpred2 had descent accuracy when the genetic architecture was sparse or the sample size was large. However, there was clear advantage of cross population methods (SDPRX and PRS-CSx) over LDpred2 when the genetic architecture was polygenic (Scenario 3) and the sample size was small (10K and 20K). Results were similar for lower genetic correlation (Figure 3.3 and 3.4; Table 3.3 and 3.4). These results suggest that jointly modeling EUR and non-EUR GWAS can improve the prediction accuracy in non-EUR populations if non-EUR GWAS alone was not well powered. We can see that SDPRX outperformed the other methods in most cases.

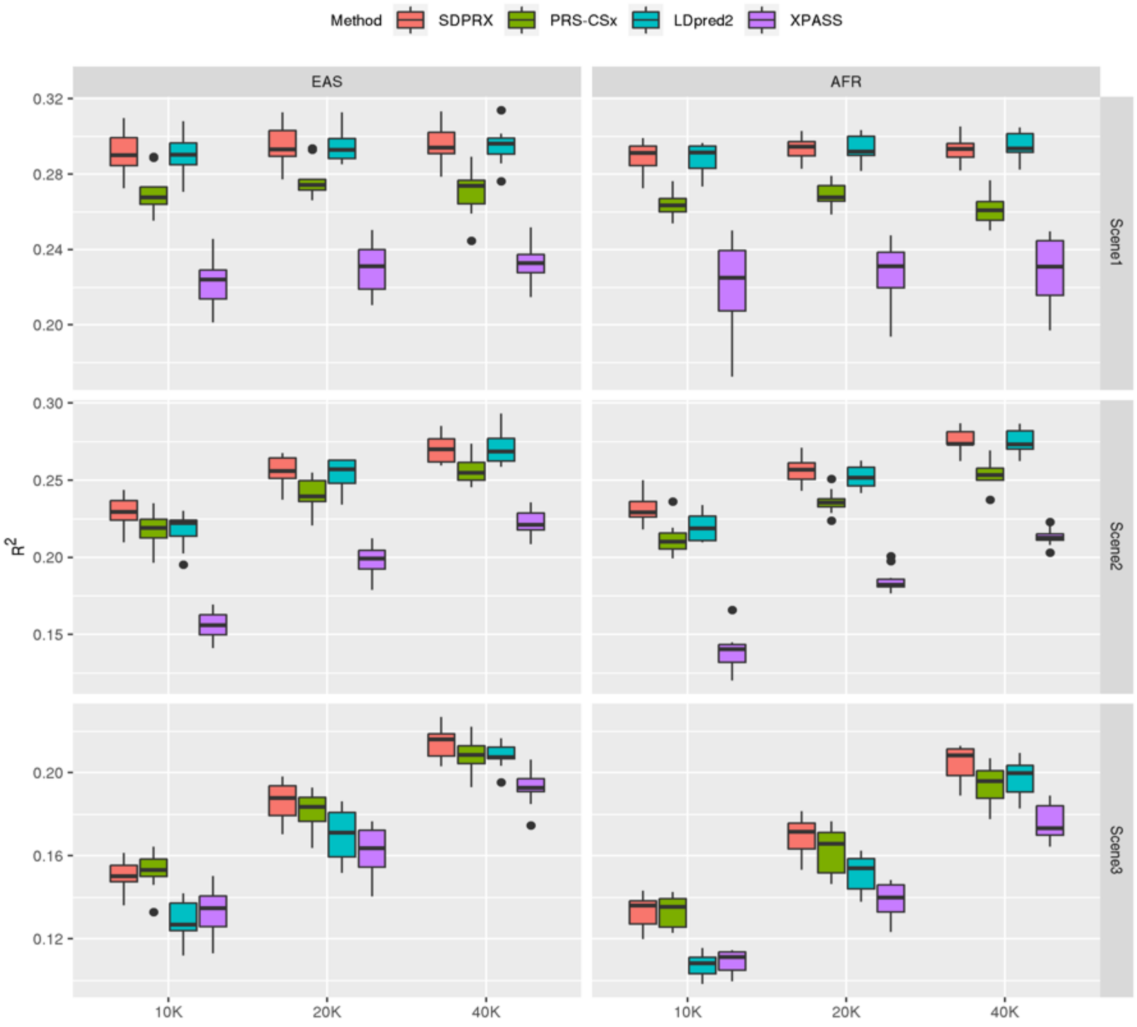


Figure 3.2. Prediction performance of different methods on simulated data with high cross-population genetic correlation. The proportion of SNPs with population 1 specific,

population 2 specific and correlated effect sizes was equally set to be 0.05% (Scenario 1), 0.5% (Scenario 2) and 5% (Scenario 3). The cross-population genetic correlation was set to be 0.8 and the heritability was 0.3. Simulation in each scenario was repeated for 10 times. For each boxplot, the central mark is the median and the lower and upper edges represent the 25th and 75th percentiles.

		EAS			AFR		
		10K	20K	40K	10K	20K	40K
Scene 1	SDPRX	0.290	0.293	0.294	0.291	0.295	0.293
	PRS-CSx	0.268	0.274	0.274	0.263	0.268	0.261
	LDpred2	0.290	0.293	0.296	0.292	0.292	0.294
	XPASS	0.224	0.231	0.233	0.225	0.231	0.231
Scene 2	SDPRX	0.230	0.256	0.270	0.229	0.257	0.274
	PRS-CSx	0.219	0.240	0.255	0.210	0.236	0.253
	LDpred2	0.222	0.257	0.269	0.219	0.252	0.273
	XPASS	0.156	0.199	0.221	0.140	0.182	0.213
Scene 3	SDPRX	0.150	0.188	0.216	0.136	0.172	0.208
	PRS-CSx	0.153	0.183	0.209	0.136	0.166	0.196
	LDpred2	0.127	0.171	0.207	0.108	0.154	0.200
	XPASS	0.135	0.164	0.193	0.111	0.140	0.173

Table 3.2. The median of square of Pearson correlation across 10 replications when the cross-population genetic correlation was 0.8.

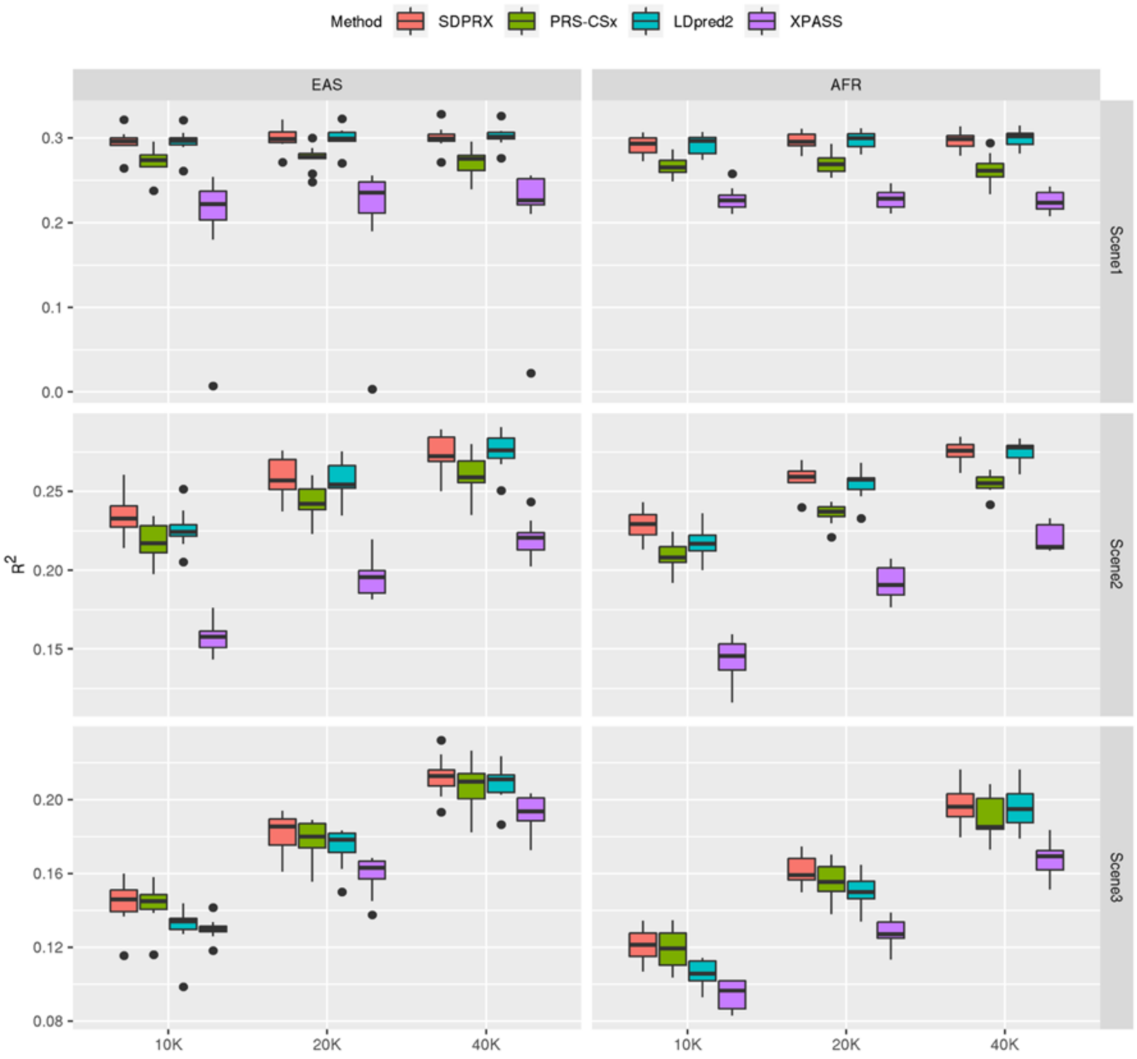


Figure 3.3. Prediction performance of different methods on simulated data with moderate cross-population genetic correlation. The proportion of SNPs with population 1 specific, population 2 specific and correlated effect sizes was equally set to be 0.05% (Scenario 1), 0.5% (Scenario 2) and 5% (Scenario 3). The cross-population genetic correlation was set to be 0.6 and the heritability was 0.3. Simulation in each scenario was repeated for 10 times. For each boxplot, the central mark is the median and the lower and upper edges represents the 25th and 75th percentiles.

		EAS			AFR		
		10K	20K	40K	10K	20K	40K
Scene 1	SDPRX	0.296	0.299	0.300	0.293	0.296	0.299
	PRS-CSx	0.274	0.278	0.276	0.265	0.269	0.262
	LDpred2	0.297	0.299	0.302	0.297	0.300	0.302
	XPASS	0.222	0.236	0.226	0.226	0.228	0.223
Scene 2	SDPRX	0.233	0.257	0.272	0.229	0.259	0.276
	PRS-CSx	0.217	0.242	0.259	0.208	0.237	0.255
	LDpred2	0.225	0.254	0.276	0.217	0.258	0.278
	XPASS	0.158	0.196	0.221	0.146	0.191	0.215
Scene 3	SDPRX	0.146	0.186	0.213	0.121	0.159	0.196
	PRS-CSx	0.145	0.180	0.210	0.119	0.156	0.185
	LDpred2	0.134	0.178	0.211	0.106	0.150	0.195
	XPASS	0.130	0.163	0.194	0.096	0.127	0.169

Table 3.3. The median of square of Pearson correlation across 10 replications when the cross-population genetic correlation was 0.6.

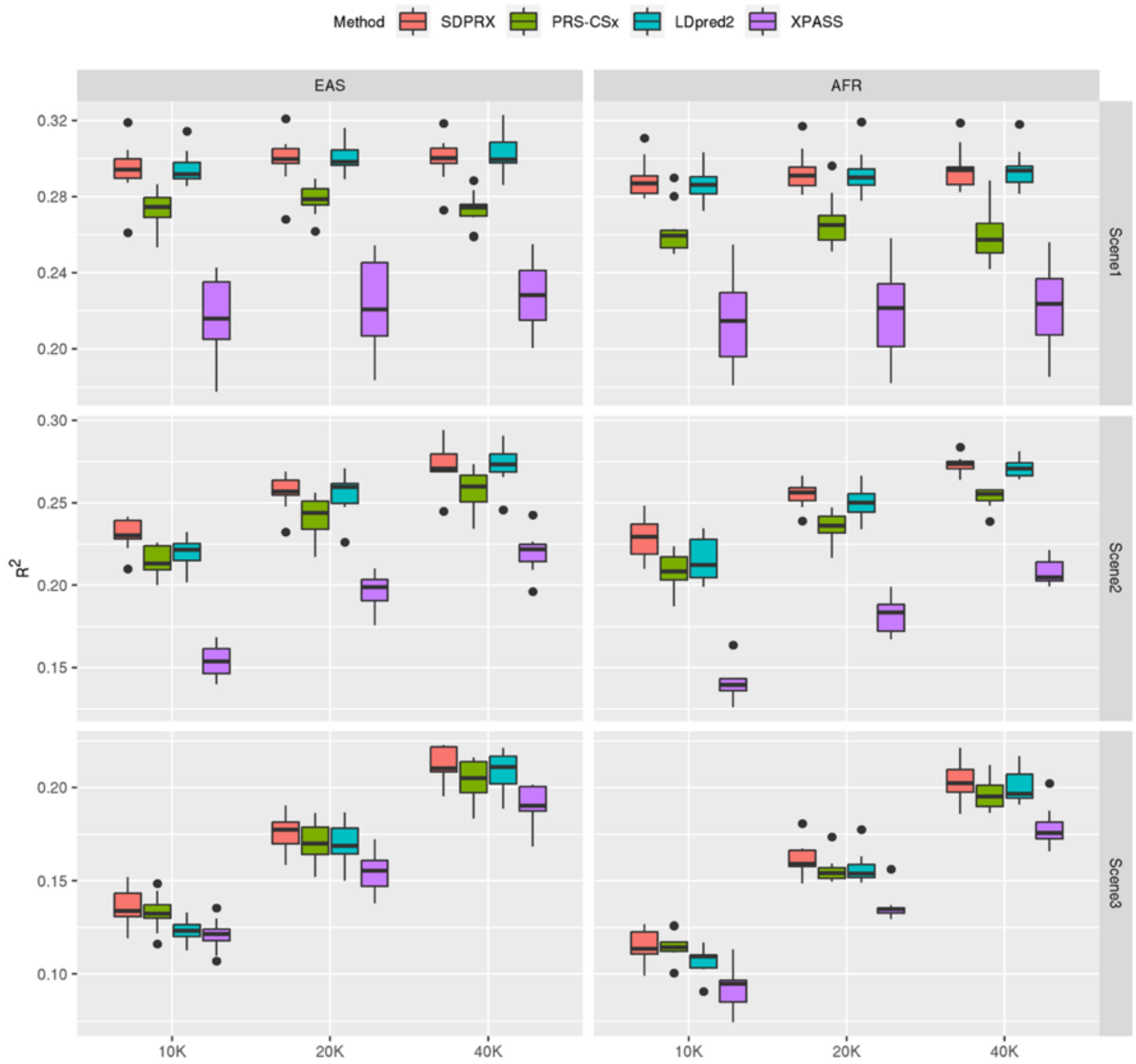


Figure 3.4. Prediction performance of different methods on simulated data with low cross-population genetic correlation. The proportion of SNPs with population 1 specific, population 2 specific and correlated effect sizes was equally set to be 0.05% (Scenario 1), 0.5% (Scenario 2) and 5% (Scenario 3). The cross-population genetic correlation was set to be 0.4 and the heritability was 0.3. Simulation in each scenario was repeated for 10 times. For each boxplot, the central mark is the median and the lower and upper edges represents the 25th and 75th percentiles.

		EAS			AFR		
		10K	20K	40K	10K	20K	40K
Scene 1	SDPRX	0.294	0.300	0.300	0.287	0.291	0.294
	PRS-CSx	0.275	0.279	0.274	0.260	0.265	0.257
	LDpred2	0.292	0.298	0.300	0.286	0.290	0.294
	XPASS	0.216	0.221	0.228	0.215	0.222	0.224
Scene 2	SDPRX	0.230	0.257	0.271	0.229	0.256	0.274
	PRS-CSx	0.213	0.244	0.260	0.208	0.236	0.256
	LDpred2	0.222	0.260	0.273	0.212	0.250	0.271
	XPASS	0.154	0.199	0.222	0.140	0.184	0.205
Scene 3	SDPRX	0.134	0.177	0.210	0.114	0.159	0.202
	PRS-CSx	0.132	0.170	0.205	0.115	0.154	0.195
	LDpred2	0.123	0.169	0.211	0.109	0.154	0.197
	XPASS	0.122	0.155	0.190	0.095	0.135	0.176

Table 3.4. The median of square of Pearson correlation across 10 replications when the cross-population genetic correlation was 0.4.

3.3.2 Prediction performance for UK Biobank traits

We next compared the performance of SDPRX with other methods in predicting six quantitative traits (height, body mass index, high-density lipoproteins, low-density lipoproteins, total cholesterol, and triglycerides) and one binary trait (type 2 diabetes) for EAS and AFR individuals in UK Biobank. We first investigated the prediction accuracy of each method in EAS (Figure 3.5 and Table 3.5) without learning the linear combination of effect sizes. Consistent with simulations, SDPRX achieved the highest prediction accuracy in all but one trait and an average of 20% increase in R^2 compared with the second-best method. The average improvement of SDPRX over LDpred2 was 22%, suggesting that jointly modeling EUR and EAS GWAS summary statistics indeed provided benefits compared with using EAS GWAS summary statistics alone. We then linearly combined EUR and EAS effect sizes for each method by weights learned on the validation dataset. SDPRX remained the best method except for LDL with an average of 12% improvement over the second-best method (Figure 3.6 and Table 3.6).

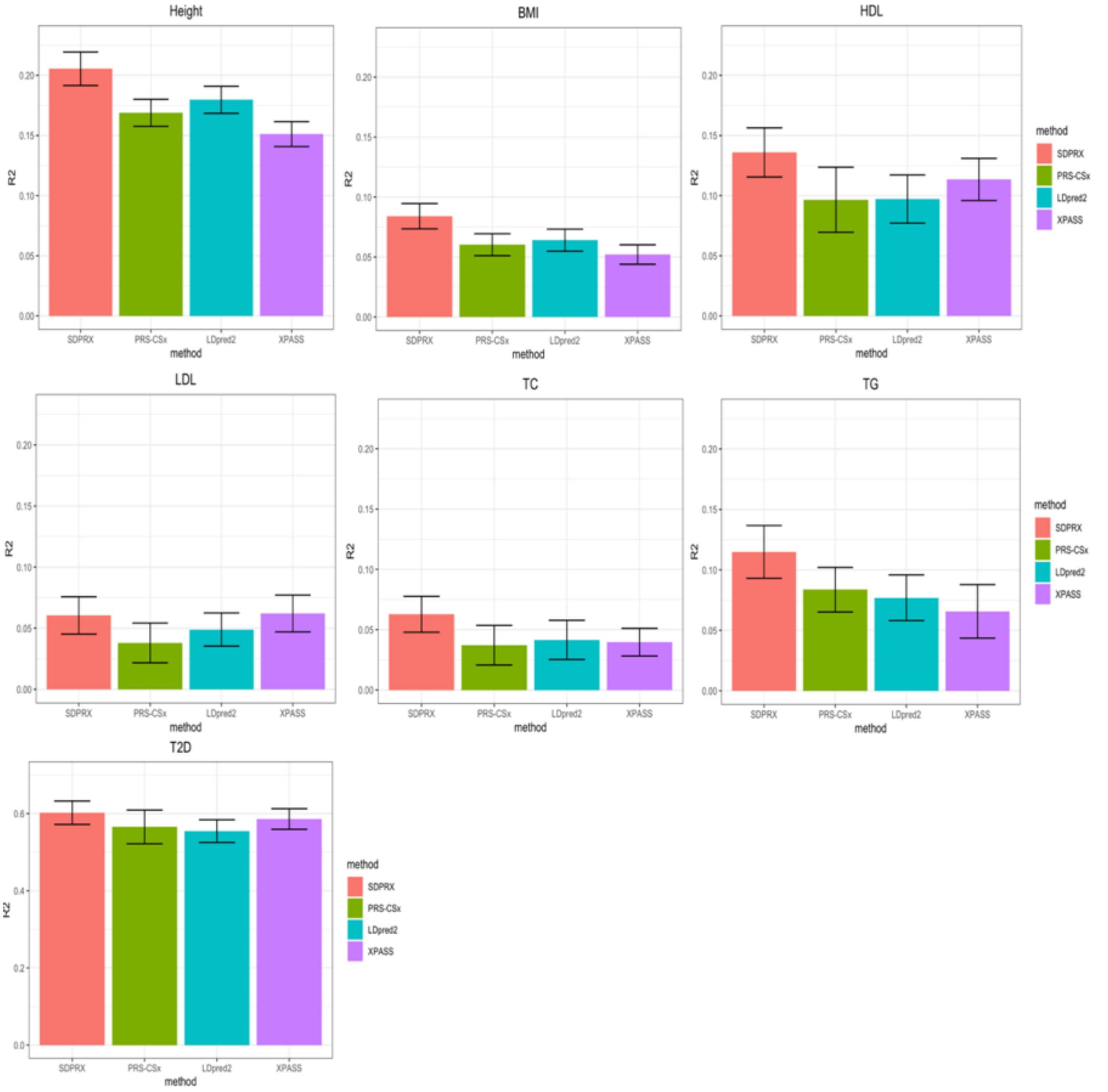


Figure 3.5. Prediction performance of different methods for six quantitative traits and one binary trait in EAS samples from UK Biobank without the linear combination of effect sizes. Selected participants with corresponding phenotypes were randomly split

to form the validation (1/3) and test datasets (2/3). The mean and standard deviation of R^2 (quantitative trait) and AUC (binary trait) across 20 random splits are showed on the bar plot.

Traits	SDPRX	PRS-CSx	LDpred2	XPASS
Height	0.213	0.203	0.202	0.175
BMI	0.084	0.059	0.081	0.052
HDL	0.136	0.104	0.108	0.127
LDL	0.059	0.036	0.028	0.061
Total cholesterol	0.063	0.039	0.032	0.033
Log triglycerides	0.116	0.102	0.114	0.071
Type 2 diabetes	0.591	0.571	0.548	0.586

Table 3.5. The mean of variance of phenotypes explained by PRS for six quantitative trait and AUC for one binary trait in EAS across 20 random splits with the linear combination of effect sizes.

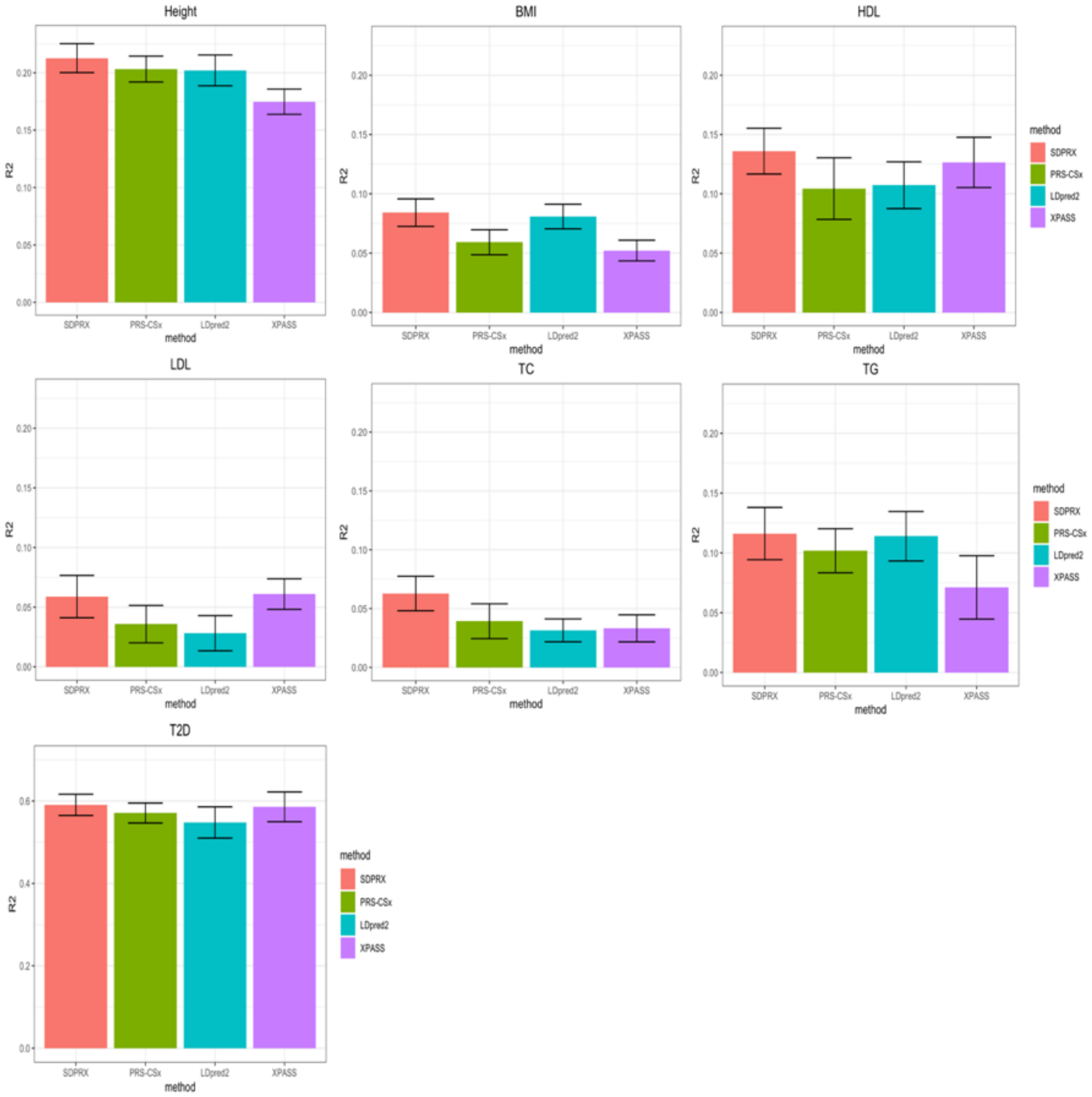


Figure 3.6. Prediction performance of different methods for six quantitative traits and one binary trait in EAS samples from UK Biobank with the linear combination of effect sizes. Selected participants with corresponding phenotypes were randomly split to form the validation (1/3) and test datasets (2/3). The mean and standard deviation of R^2 (quantitative trait) and AUC (binary trait) across 20 random splits are showed on the bar plot.

Traits	SDPRX	PRS-CSx	LDpred2	XPASS
Height	0.067	0.059	0.055	0.049
BMI	0.028	0.021	0.020	0.019
HDL	0.097	0.083	0.085	0.077
LDL	0.139	0.131	0.129	0.085
Total cholesterol	0.125	0.102	0.123	0.074
Log triglycerides	0.041	0.036	0.038	0.032
Type 2 diabetes	0.560	0.550	0.545	0.541

Table 3.6. The mean of variance of phenotypes explained by PRS for six quantitative trait and AUC for one binary trait in AFR across 20 random splits without the linear combination of effect sizes.

Results for AFR were similar to results for EAS (Figure 3.7 and Table 3.7). SDPRX performed the best in all traits regardless of learning the linear combination of effect sizes. The average improvement of SDPRX over the second-best method was 11% before the linear combination, and 14% after the linear combination (Figure 3.8 and Table 3.8).

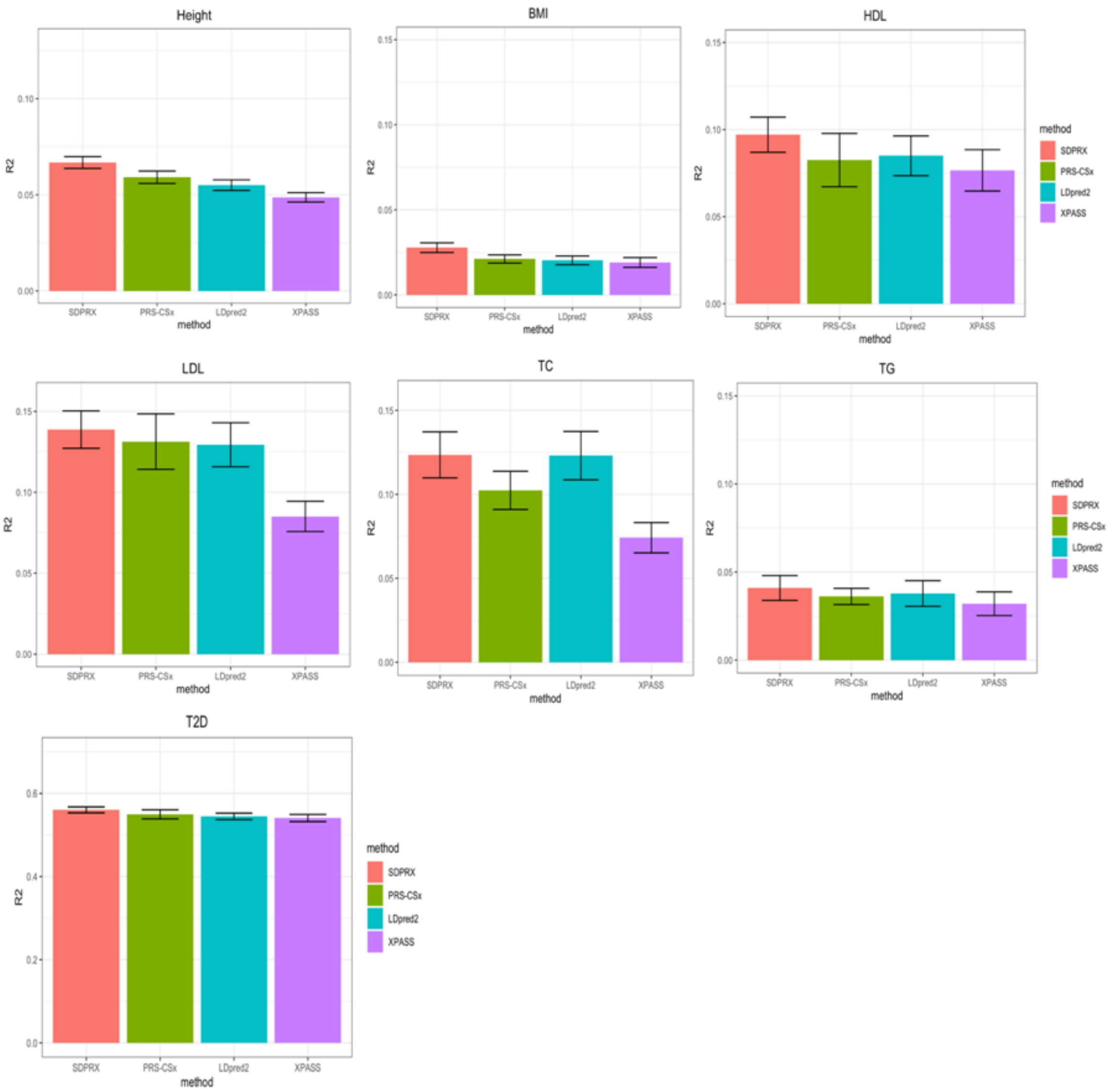


Figure 3.7. Prediction performance of different methods for six quantitative traits and one binary trait in AFR samples from UK Biobank without the linear combination of effect sizes. Selected participants with corresponding phenotypes were randomly split to form the validation (1/3) and test datasets (2/3). The mean and standard deviation of R^2 (quantitative trait) and AUC (binary trait) across 20 random splits are showed on the bar plot.

Traits	SDPRX	PRS-CSx	LDpred2	XPASS
Height	0.067	0.059	0.055	0.049
BMI	0.028	0.021	0.020	0.019
HDL	0.097	0.083	0.085	0.077
LDL	0.139	0.131	0.129	0.085
Total cholesterol	0.125	0.102	0.123	0.074
Log triglycerides	0.041	0.036	0.038	0.032
Type 2 diabetes	0.560	0.550	0.545	0.541

Table 3.7. The mean of variance of phenotypes explained by PRS for six quantitative trait and AUC for one binary trait in AFR across 20 random splits without the linear combination of effect sizes.

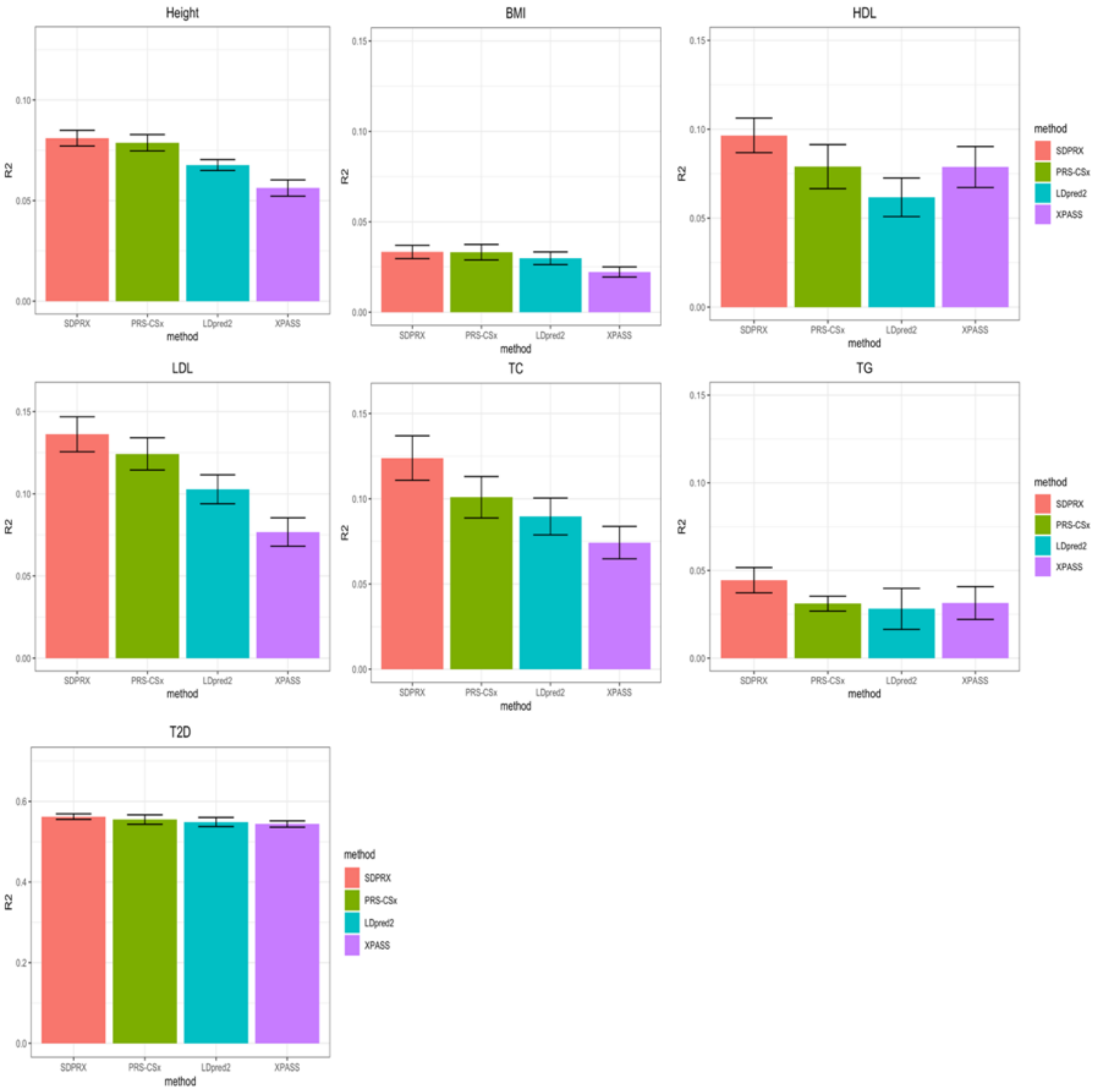


Figure 3.8. Prediction performance of different methods for six quantitative traits and one binary trait in EAS samples from UK Biobank with the linear combination of effect sizes. Selected participants with corresponding phenotypes were randomly split to form the validation (1/3) and test dataset (2/3). The mean and standard deviation of R^2 (quantitative trait) and AUC (binary trait) across 20 random splits are showed on the bar plot.

Traits	SDPRX	PRS-CSx	LDpred2	XPASS
Height	0.081	0.079	0.068	0.056
BMI	0.033	0.033	0.030	0.022
HDL	0.097	0.079	0.062	0.079
LDL	0.136	0.125	0.103	0.076
Total cholesterol	0.124	0.101	0.090	0.074
Log triglycerides	0.044	0.031	0.028	0.031
Type 2 diabetes	0.562	0.555	0.549	0.544

Table 3.8. The mean of variance of phenotypes explained by PRS for six quantitative trait and AUC for one binary trait in AFR across 20 random splits with the linear combination of effect sizes.

We next compared the prediction accuracy of the same trait between EAS and AFR. For height, BMI and T2D, the prediction was better in EAS than AFR because AFR GWAS summary statistics were largely derived from an admixed population with a lower sample size. Among four lipid traits, the performance of LDL and TC was lower in EAS while the performance of HDL and TG was lower in AFR. A glimpse into the number of SNPs with relatively large posterior effect sizes in EAS and AFR (absolute value greater than 0.01) may explain the discrepancy above (Figure 3.10). There were more SNPs with AFR specific effect sizes for LDL and TC, and more SNPs had EAS specific effect sizes for HDL and TG. Hence, both sample size and genetic architecture of two populations affect the transferability of PRS.

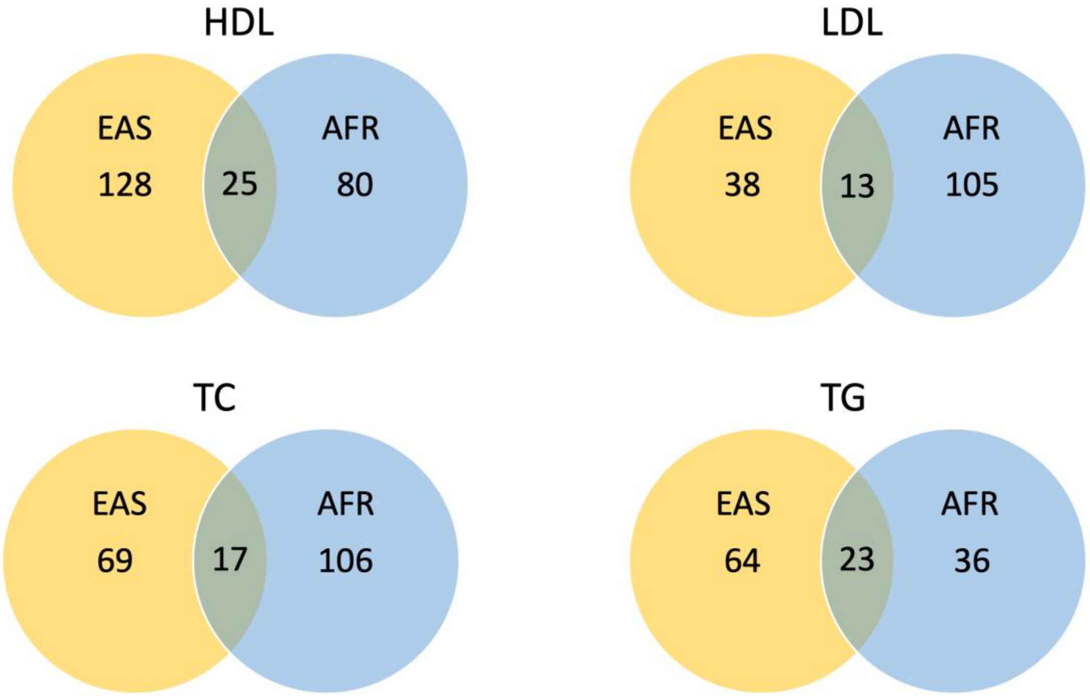


Figure 3.9. Venn diagram showing the number of SNPs with relatively large posterior effect sizes (absolute value greater than 0.01) output by SDPRX in EAS and AFR.

3.4 Discussion

SDPRX takes GWAS summary statistics from two populations as input, and thus is able to leverage shared information from two populations to better estimate the effect sizes of SNPs compared with single population method like LDpred2. The prior assumption made by SDPRX is more general than XPASS and PRS-CSx. Unlike SDPRX, XPASS assumes that the genetic architecture is polygenic and all SNPs have non-zero effect sizes, while empirically methods assuming only part of SNPs having non-zero effect sizes often have better performance [18]. SDPRX differs with PRS-CSx in two aspects. First, it explicitly allows SNPs to have both population specific and shared effect sizes whereas PRS-CSx assumes all SNPs are shared. Second, SDPRX directly incorporates the cross-population genetic correlation into the model for better estimation of shared effect sizes. These points, taken together, may explain why SDPRX outperformed the other methods in both simulation and real data analyses.

Although SDPRX improves the prediction accuracy in non-EUR populations, it is far from overcoming the gap between performance of PRS in EUR and non-EUR populations. We think developing computational methods alone will not be able to solve this issue, and there are two points that may explain the gap based on the results presented in this paper. First, the sample sizes of non-EUR GWAS are limited. Results in EAS were overall better than results in AFR due to the larger sample size of EAS GWAS. Second, other factors like genetic architecture may be different for some traits in two populations. For example, the performance of HDL, LDL, TC and TG was different in EAS and AFR in spite of similar GWAS sample sizes. We also note that social, environmental and familial factors were not considered in this study since we primarily focused on comparison of methods, though they may play an important role in the transferability of PRS [66].

Lastly, we note three limitations of our current work that we will address in the future. First, we restricted to HM3 SNPs for an easy comparison of different methods, which is not optimal as it might not include some informative SNPs. Second, SDPRX is currently not designed for admixed populations, which is challenging as the LD pattern would be heterogenous and difficult to capture using a single LD matrix. To our knowledge, how to connect the marginal effect sizes in the GWAS summary statistics derived from admixed populations with true effect sizes is also less clear, which may deal with the adjustment of local ancestry and covariates [80, 81]. Third, methods utilizing functional annotation have shown to improve the performance in both single and cross population settings [64, 82]. Incorporating functional annotation may further improve the performance of SDPRX.

Appendix A

In this appendix, we derive the likelihood function when SNPs are typed on different individuals, motivated by the observation in the Table 2.1.

Claim: If each SNP j is genotyped on N_j individuals, define the matrix H whose elements are $H_{ii} = \frac{1}{N_i}$ and $H_{ij} = \frac{N_{s,ij}}{N_i N_j}$ $(i \neq j)$, where $N_{s,ij}$ is the number of shared individuals genotyped for SNPs i and j. Then the likelihood function can be evaluated as

$$\hat{\beta}|\beta \sim N(R\beta, R \circ H)$$

where ° is the Hadamard product.

Proof: Let $S_j$be the set of individuals on which SNP j is genotyped ($N_j = |S_j|$), then

$$E[\widehat{\beta_j}|\beta] = E\left[\frac{X_j^T y}{Nj}\middle|\beta\right]$$

$$= \sum_{i \in S_j} \frac{X_{ji}}{N_j}\left(\sum_k X_{ik}\beta_k\right)$$

$$= \sum_k \left(\sum_{i \in S_j} \frac{X_{ji}X_{ik}}{N_j}\right)\beta_k$$

$$= \sum_k R_{jk}\beta_k$$

$$(A.1)$$

For $i \neq j$, we have

$$cov[\hat{\beta}_i, \hat{\beta}_j\ |\beta] = E\left[\frac{X_i^T \epsilon_i X_j^T \epsilon_j}{N_i N_j}\middle|\beta\right]$$

$$= \frac{1}{N_i N_j} E\left[X_i^T \epsilon_i \epsilon_j^T X_j\right]$$

$$= \frac{1}{N_i N_j} X_i^T E\left[\epsilon_i \epsilon_j^T\right] X_j$$

$$= \frac{1}{N_i N_j} \sum_{k \in S_i \cap S_j} X_{ik} X_{jk}$$

$$= \frac{R_{ij} N_{s,ij}}{N_i N_j}$$

$$(A.2)$$

It is trivial to check that when the sample size of all SNPs is same, the derived likelihood function is the same as equation (2.2). Furthermore, the correlation of marginal effect sizes in the GWAS summary statistics will be less than the correlation in the reference panel, depending on how many individuals are overlapped ($N_{s,ij}$) for two SNPs ($cor = \frac{R_{ij} N_{s,ij}}{\sqrt{N_i N_j}} \leq \frac{R_{ij} \min(N_i, N_j)}{\sqrt{N_i N_j}} \leq R_{ij}$). As an extreme example, for two SNPs that are in perfect LD ($R_{ij} = 1$), if they are genotyped on completely nonoverlapped individuals ($N_s = 0$), the correlation of their effect sizes in the summary statistics would be zero.

Appendix B

This appendix provides overview of key properties of Dirichlet process, which is useful to understand the development of the model in chapter 2 and 3. The writing style is casual and we refer to Kevin Murphy's textbook for a more rigorous treatment [83].

B.1 Abstract definition of Dirichlet process

Dirichlet process (DP) is parameterized by a concentration parameter $\alpha > 0$ and a base distribution H over a space Θ [84]. DP can be viewed as a distribution over distribution. When a distribution G is drawn from a DP, we denote $G \sim DP(H, \alpha)$. For any finite partition $(T_1, \dots, T_k)$ of Θ, it satisfies the requirement that

$$\big(G(T_1), \dots, G(T_k)\big) \sim Dir\big(\alpha H(T_1), \dots, \alpha H(T_k)\big) \qquad (B.1)$$

The definition is abstract and requires some explanation. For example, suppose base distribution H is $N(0,1)$, we first draw countably infinite θ_k $(k = 1, \dots, \infty)$ from $N(0,1)$. The drawn distribution G from $DP(H, \alpha)$ is discrete such that $P(G(\theta) = \theta_k) = \pi_k$ and $\sum_{k=1}^{\infty} \pi_k = 1$. If we divide the real line ($\Theta = \mathbb{R}$) into k disjoint partitions $\{T_1 = (-\infty, a_1], T_2 = (a_2, a_3], \dots, T_k = (a_k, \infty)\}$, then $(\sum_{\theta_k \in T_1} \pi_k, \dots, \sum_{\theta_k \in T_k} \pi_k)$ has a joint Dirichlet distribution $Dir(\alpha\Phi(a_1), \dots, \alpha(1 - \Phi(\mathrm{a_k}))$ where Φ is the cumulative distribution function of standard normal distribution. In practice, most application of Dirichlet process does not directly deal with this abstract definition (B.1). Instead, it deals with two equivalent constructive representation of DP: Chinese restaurant process (CRP) and stick-breaking process.

B.2 Chinese restaurant process

There are many ways to derive the CRP. One intuitive way is to start with the finite Gaussian mixture model and let the number of clusters approaches to infinity. Such infinite Gaussian mixture model converges to DP and has CRP property in the cluster assignment [85].

Given $G \sim DP(H, \alpha)$, $x_1, \dots, x_{N-1}$ are $N - 1$ iid observations drawn from the distribution G which take K distinct values $\theta_1, \dots, \theta_K$. The Chinese restaurant process (CRP) asserts that the predictive distribution of next observation x_N given $x_1, \dots, x_{N-1}$ is:

$$p(x_N = \theta | x_1, \dots, x_{N-1}, \alpha) = \frac{1}{\alpha + N - 1}\left(\alpha H(\theta) + \sum_{k=1}^{K} N_k \delta_{\theta_k}(\theta)\right) \qquad (B.2)$$

where N_k is the number of previous observations equal to θ_k.

This is called Polya urn or Blackwell-MacQueen sampling scheme. Equivalently, this asserts that $x_1, \dots, x_{N-1}$ has the clustering property:

$$P(z_N = k | z_1, \dots, z_{N-1}, \alpha) = \begin{cases} \dfrac{N_k}{N - 1 + \alpha} & \text{if k is an old cluster} \\ \dfrac{\alpha}{N - 1 + \alpha} & \text{if k is a new cluster} \end{cases}$$

where z_i is the cluster assignment of x_i, N_k is the number of previous observations assigned to cluster k. The analogy of Chinese restaurant process follows that when a person enters the restaurant, he may choose to join an existing table with probability proportional to the number of people already sitting at this table (N_k); otherwise, with a probability that diminishes as more people enter the room (due to the $1/(N - 1 + \alpha)$ term), he may choose to sit at a new table. If he joins an existing table, he takes the parameter of that table; if he sits at a new table, he draws a new parameter from the base distribution. Although deriving the Gibbs sample to fit the model is relatively straightforward, the dependency structure of the CRP makes it time-consuming when applied to large-scale data.

B.3 Stick-breaking process

Let $\{\pi_k\}_{k=1}^{\infty}$ be an infinite sequence of mixture weights which is derived from process below:

$$V_k \sim \mathrm{Beta}(1, \alpha)$$

$$\pi_k = V_k \prod_{l=1}^{k-1} V_l = V_k \left(1 - \sum_{l=1}^{k-1} \pi_l\right) \qquad (B.3)$$

Now define $G(\theta) = \sum_{k=1}^{\infty} \delta_{\theta_k}(\theta)$ where $\delta(.)$ is the Kronecker delta function and$\theta_k \sim H$. It can be shown that this construction of G satisfies equation (B.1) [86]. In practice, it is impossible to draw infinite many θ_k and one has to specify an upper bound K on the number of θ_k drawn from the base distribution, such that $G(\theta) = \sum_{k=1}^{K} \delta_{\theta_k}(\theta)$. This is a reasonable approximation because as k increases, π_k converges to 0 and only first few π_k would actually matter. We chose $K = 1000$ for SDPR and SDPRX and found the approximation worked well. Compared with CRP, stick-breaking process has the advantage that θ_k is drawn independently of $\theta_1, \ldots, \theta_{k-1}$, which allows the design of a parallel algorithm for efficient sampling when applied to large-scale data. For this reason, we used stick-breaking process rather than CRP when designing SDPR and SDPRX.

Bibliography:

1. Rommens, J.M., et al., *Identification of the cystic fibrosis gene: chromosome walking and jumping.* Science, 1989. **245**(4922): p. 1059-65.
2. *A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group.* Cell, 1993. **72**(6): p. 971-83.
3. Galton, F., *Regression towards mediocrity in hereditary stature.* The Journal of the Anthropological Institute of Great Britain and Ireland, 1886. **15**: p. 246-263.
4. Fisher, R.A., *XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.* Transactions of the Royal Society of Edinburgh, 1919. **52**(2): p. 399-433.
5. Visscher, P.M. and M.E. Goddard, *From R.A. Fisher's 1918 Paper to GWAS a Century Later.* Genetics, 2019. **211**(4): p. 1125-1130.
6. Boyle, E.A., Y.I. Li, and J.K. Pritchard, *An Expanded View of Complex Traits: From Polygenic to Omnigenic.* Cell, 2017. **169**(7): p. 1177-1186.
7. Loh, P.R., et al., *Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis.* Nat Genet, 2015. **47**(12): p. 1385-92.
8. Purcell, S.M., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-52.
9. Wray, N.R., et al., *Pitfalls of predicting complex traits from SNPs.* Nat Rev Genet, 2013. **14**(7): p. 507-15.
10. Lee, S.H., et al., *Estimating missing heritability for disease from genome-wide association studies.* Am J Hum Genet, 2011. **88**(3): p. 294-305.
11. Choi, S.W. and P.F. O'Reilly, *PRSice-2: Polygenic Risk Score software for biobank-scale data.* Gigascience, 2019. **8**(7).
12. Ma, Y. and X. Zhou, *Genetic prediction of complex traits with polygenic scores: a statistical review.* Trends Genet, 2021. **37**(11): p. 995-1011.
13. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nat Genet, 2010. **42**(7): p. 565-9.
14. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.
15. Robinson, M.R., et al., *Genetic evidence of assortative mating in humans.* Nature Human Behaviour, 2017. **1**: p. 0016.
16. Vilhjalmsson, B.J., et al., *Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores.* Am J Hum Genet, 2015. **97**(4): p. 576-92.
17. Jiang, J., et al., *On high-dimensional misspecified mixed model analysis in genome-wide association study.* The Annals of Statistics, 2016. **44**(5): p. 2127-2160.
18. Privé, F., J. Arbel, and B.J. Vilhjálmsson, *LDpred2: better, faster, stronger.* Bioinformatics, 2020.
19. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-288.

20. Qian, J., et al., *A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank.* PLoS Genet, 2020. **16**(10): p. e1009141.
21. Mak, T.S.H., et al., *Polygenic scores via penalized regression on summary statistics.* Genet Epidemiol, 2017. **41**(6): p. 469-480.
22. George, E.I. and R.E. McCulloch, *Variable selection via Gibbs sampling.* Journal of the American Statistical Association, 1993. **88**(423): p. 881-889.
23. Zhou, X. and M. Stephens, *Genome-wide efficient mixed-model analysis for association studies.* Nat Genet, 2012. **44**(7): p. 821-4.
24. Zhu, X. and M. Stephens, *BAYESIAN LARGE-SCALE MULTIPLE REGRESSION WITH SUMMARY STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES.* Ann Appl Stat, 2017. **11**(3): p. 1561-1592.
25. Zhou, X., P. Carbonetto, and M. Stephens, *Polygenic modeling with Bayesian sparse linear mixed models.* PLoS genetics, 2013. **9**(2): p. e1003264.
26. Moser, G., et al., *Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model.* PLoS genetics, 2015. **11**(4): p. e1004969.
27. Lloyd-Jones, L.R., et al., *Improved polygenic prediction by Bayesian multiple regression on summary statistics.* Nature Communications, 2019. **10**(1): p. 5086.
28. Habier, D., et al., *Extension of the bayesian alphabet for genomic selection.* BMC Bioinformatics, 2011. **12**: p. 186.
29. Carvalho, C.M., N.G. Polson, and J.G. Scott, *The horseshoe estimator for sparse signals.* Biometrika, 2010. **97**(2): p. 465-480.
30. Ge, T., et al., *Polygenic prediction via Bayesian regression and continuous shrinkage priors.* Nat Commun, 2019. **10**(1): p. 1776.
31. Lewis, C.M. and E. Vassos, *Polygenic risk scores: from research tools to clinical instruments.* Genome Med, 2020. **12**(1): p. 44.
32. Khera, A.V., et al., *Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations.* Nat Genet, 2018. **50**(9): p. 1219-1224.
33. Mavaddat, N., et al., *Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes.* Am J Hum Genet, 2019. **104**(1): p. 21-34.
34. Richardson, T.G., et al., *An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome.* Elife, 2019. **8**.
35. Zhang, Y., et al., *Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits.* Nature Genetics, 2018.
36. Ferguson, T.S., *A Bayesian Analysis of Some Nonparametric Problems.* Ann. Statist., 1973. **1**(2): p. 209-230.
37. Zeng, P. and X. Zhou, *Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models.* Nat Commun, 2017. **8**(1): p. 456.
38. Berisa, T. and J.K. Pickrell, *Approximately independent linkage disequilibrium blocks in human populations.* Bioinformatics, 2016. **32**(2): p. 283-5.

39. Gelman, A., et al., *Using Redundant Parameterizations to Fit Hierarchical Models.* Journal of Computational and Graphical Statistics, 2008. **17**(1): p. 95-122.
40. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
41. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.
42. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nat Genet, 2014. **46**(11): p. 1173-86.
43. Willer, C.J., et al., *Discovery and refinement of loci associated with lipid levels.* Nat Genet, 2013. **45**(11): p. 1274-1283.
44. Speed, D. and D.J. Balding, *SumHer better estimates the SNP heritability of complex traits from summary statistics.* Nat Genet, 2019. **51**(2): p. 277-284.
45. Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics, 1999. **55**(4): p. 997-1004.
46. Gelman, A., *Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper).* Bayesian Anal., 2006. **1**(3): p. 515-534.
47. Schafer, J. and K. Strimmer, *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article32.
48. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
49. Yang, S. and X. Zhou, *Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets.* Am J Hum Genet, 2020. **106**(5): p. 679-693.
50. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.
51. Yengo, L., et al., *Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry.* Hum Mol Genet, 2018. **27**(20): p. 3641-3649.
52. Zheng, J., et al., *LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis.* Bioinformatics, 2017. **33**(2): p. 272-279.
53. Lijoi, A., I. Prünster, and S.G. Walker, *On Consistency of Nonparametric Normal Mixtures for Bayesian Density Estimation.* Journal of the American Statistical Association, 2005. **100**(472): p. 1292-1296.
54. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology.* Nature, 2015. **518**(7538): p. 197-206.
55. Mehta, N.N., *Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease.* Circ Cardiovasc Genet, 2011. **4**(3): p. 327-9.
56. Michailidou, K., et al., *Association analysis identifies 65 new breast cancer risk loci.* Nature, 2017. **551**(7678): p. 92-94.

57. Liu, J.Z., et al., *Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.* Nat Genet, 2015. **47**(9): p. 979-986.

58. Scott, R.A., et al., *An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans.* Diabetes, 2017. **66**(11): p. 2888-2902.

59. *Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes.* Cell, 2018. **173**(7): p. 1705-1715.e16.

60. Chun, S., et al., *Non-parametric Polygenic Risk Prediction via Partitioned GWAS Summary Statistics.* The American Journal of Human Genetics, 2020. **107**(1): p. 46-59.

61. So, H.C. and P.C. Sham, *Improving polygenic risk prediction from summary statistics by an empirical Bayes approach.* Sci Rep, 2017. **7**: p. 41262.

62. Nagpal, S., et al., *TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits.* The American Journal of Human Genetics, 2019. **105**(2): p. 258-266.

63. Chatterjee, N., et al., *Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies.* Nat Genet, 2013. **45**(4): p. 400-5, 405e1-3.

64. Hu, Y., et al., *Leveraging functional annotations in genetic risk prediction for human complex diseases.* PLOS Computational Biology, 2017. **13**(6): p. e1005589.

65. Duncan, L., et al., *Analysis of polygenic risk score usage and performance in diverse human populations.* Nature Communications, 2019. **10**(1): p. 3328.

66. Martin, A.R., et al., *Clinical use of current polygenic risk scores may exacerbate health disparities.* Nat Genet, 2019. **51**(4): p. 584-591.

67. Shi, H., et al., *Population-specific causal disease effect sizes in functionally important regions impacted by selection.* Nat Commun, 2021. **12**(1): p. 1098.

68. Shi, H., et al., *Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data.* Am J Hum Genet, 2020. **106**(6): p. 805-817.

69. Wojcik, G.L., et al., *Genetic analyses of diverse populations improves discovery for complex traits.* Nature, 2019. **570**(7762): p. 514-518.

70. Graham, S.E., et al., *The power of genetic diversity in genome-wide association studies of lipids.* Nature, 2021. **600**(7890): p. 675-679.

71. Weissbrod, O., et al., *Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores.* medRxiv, 2021: p. 2021.01.19.21249483.

72. Cai, M., et al., *A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits.* Am J Hum Genet, 2021. **108**(4): p. 632-655.

73. Ruan, Y., et al., *Improving Polygenic Prediction in Ancestrally Diverse Populations.* medRxiv, 2021: p. 2020.12.27.20248738.

74. Zhou, G. and H. Zhao, *A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics.* PLoS Genet, 2021. **17**(7): p. e1009697.

75. Ishwaran, H. and L.F. James, *Gibbs Sampling Methods for Stick-Breaking Priors.* Journal of the American Statistical Association, 2001. **96**(453): p. 161-173.

76. Brown, B.C., et al., *Transethnic Genetic-Correlation Estimates from Summary Statistics.* Am J Hum Genet, 2016. **99**(1): p. 76-88.

77. Suzuki, K., et al., *Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population.* Nat Genet, 2019. **51**(3): p. 379-386.

78. Akiyama, M., et al., *Genome-wide association study identifies 112 new loci for body mass index in the Japanese population.* Nat Genet, 2017. **49**(10): p. 1458-1467.

79. Akiyama, M., et al., *Characterizing rare and low-frequency height-associated variants in the Japanese population.* Nat Commun, 2019. **10**(1): p. 4393.

80. Atkinson, E.G., et al., *Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power.* Nat Genet, 2021. **53**(2): p. 195-204.

81. Luo, Y., et al., *Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations.* Hum Mol Genet, 2021. **30**(16): p. 1521-1534.

82. Amariuta, T., et al., *Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements.* Nat Genet, 2020. **52**(12): p. 1346-1354.

83. Murphy, K.P., *Machine Learning: A Probabilistic Perspective*. 2012: The MIT Press.

84. Ferguson, T.S., *A Bayesian analysis of some nonparametric problems.* The annals of statistics, 1973: p. 209-230.

85. Rasmussen, C., *The infinite Gaussian mixture model.* Advances in neural information processing systems, 1999. **12**.

86. Sethuraman, J., *A constructive definition of Dirichlet priors.* Statistica sinica, 1994: p. 639-650.