# Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets

Sheng Yang[1,2,3] and Xiang Zhou[2,3,*]

Accurate construction of polygenic scores (PGS) can enable early diagnosis of diseases and facilitate the development of personalized medicine. Accurate PGS construction requires prediction models that are both adaptive to different genetic architectures and scalable to biobank scale datasets with millions of individuals and tens of millions of genetic variants. Here, we develop such a method called Deterministic Bayesian Sparse Linear Mixed Model (DBSLMM). DBSLMM relies on a flexible modeling assumption on the effect size distribution to achieve robust and accurate prediction performance across a range of genetic architectures. DBSLMM also relies on a simple deterministic search algorithm to yield an approximate analytic estimation solution using summary statistics only. The deterministic search algorithm, when paired with further algebraic innovations, results in substantial computational savings. With simulations, we show that DBSLMM achieves scalable and accurate prediction performance across a range of realistic genetic architectures. We then apply DBSLMM to analyze 25 traits in UK Biobank. For these traits, compared to existing approaches, DBSLMM achieves an average of 2.03%–101.09% accuracy gain in internal cross-validations. In external validations on two separate datasets, including one from BioBank Japan, DBSLMM achieves an average of 14.74%–522.74% accuracy gain. In these real data applications, DBSLMM is 1.03–28.11 times faster and uses only 7.4%–24.8% of physical memory as compared to other multiple regression-based PGS methods. Overall, DBSLMM represents an accurate and scalable method for constructing PGS in biobank scale datasets.

## Introduction

The polygenic score (PGS) for a phenotype, in its simplest form, is a weighted summation of the estimated genetic effect sizes across genome-wide single nucleotide polymorphisms (SNPs).[1–4] Through aggregating the contribution of many SNPs toward the phenotype of interest, PGS can be used to construct an individual's inherited component, which is his/her genetic predisposition, underlying the phenotype.[5–7] By estimating the genetic predisposition, PGS serves both as the earliest measurable predictor and as the most stable predictor for disease and disease-related complex traits.[8–10] PGS is commonly referred to as the polygenic risk score (PRS) when the phenotype of interest is a disease status.[2,11] PGS has a long-standing history both in animal breeding programs and in human genetics.[12] PGS has also been widely applied to a range of genetic applications that include disease risk prediction,[13–25] genetic prediction of complex traits,[14,15,17,19,26–30] prioritization of preventive interventions,[31–36] understanding missing heritability,[37–40] modeling polygenic adaptation,[41] genomic selection in animal breeding programs,[42,43] transcriptome-wide association studies (TWASs),[44–46] and, more recently, Mendelian randomization analysis.[47–49] Accurate construction of PGS can facilitate disease prevention and intervention at an early stage and can aid in the development of personalized medicine.

Various statistical methods have been developed for constructing PGS.[50] Different PGS methods often differ in their modeling assumptions on the SNP effect size distribution[15,28] (a detailed review is provided in the Supplemental Material and Methods). Previous studies have shown that a flexible effect size modeling assumption is key for constructing accurate PGS across a range of phenotypes with different genetic architectures.[15,28] However, achieving flexible effect size modeling is challenging computationally on large-scale genome-wide association studies (GWASs) that are being conducted today. Specifically, the sample size of GWASs has been steadily and rapidly increasing in the past years, with biobank scale datasets becoming increasingly common. These large biobank-scale datasets include UK Biobank (UKB),[51] BioBank Japan (BBJ),[52] China Kadoorie Biobank,[53] FINNGEN,[54] and All of Us,[55] each containing hundreds of thousands of individuals and tens of millions of genetic markers. Many existing PGS approaches, especially the ones with flexible modeling assumptions on the effect size distribution (e.g., Bayes alphabetic methods,[42] Bayesian sparse linear mixed model [BSLMM],[15] and Dirichlet process regression model [DPR][28]), often rely on computationally expensive algorithms such as Markov chain Monte Carlo (MCMC) and are thus unscalable to large-scale biobank datasets.[56] Indeed, only a limited number of PGS methods, often the ones with relatively simple effect size assumptions, are scalable to biobank datasets. Notable scalable PGS methods include SBLUP,[27] LDpred,[13] lassosum,[16] and C+T procedure.[18,57] These scalable PGS methods often use summary statistics in terms of marginal z-scores or marginal effect size estimates commonly available from biobank datasets. In addition, these methods also rely on a reference panel—obtained either directly from the data or from an external reference panel—to calculate the

[1]Department of Biostatistics, Nanjing Medical University, Nanjing, Jiangsu 211166, China; [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; [3]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA
*Correspondence: xzhousph@umich.edu

necessary linkage disequilibrium (LD) matrix for model fitting. However, even these scalable methods, with the only exception of the C+T procedure, are not easily applicable to biobank scale data. For example, LDpred and SBLUP, even with low computing cost settings, remain computationally costly and require a large amount of physical memory (> 50 GB) for constructing PGS. Such large physical memory requirements are not readily accessible even on large computing clusters. For example, the state-of-art University of Michigan Center for Statistical Center computing cluster, which has successfully hosted and is currently hosting many large-scale GWAS data and analyses (e.g., 1000 Genomes projects, TOPMed project, various T2D consortium projects, etc.), only has a limited number of nodes with large physical memory (Table S1). Consequently, there is a pressing need to develop computationally scalable PGS methods with both low memory cost and relatively flexible modeling assumptions to enable accurate PGS construction in large-scale biobank datasets.

Here, we develop a scalable computing method for accurate PGS construction in biobank scale datasets. Our method makes use of summary statistics and relies on a flexible modeling assumption on the effect size distribution similar to that used in BSLMM,[15] an accurate but unscalable prediction model commonly applied to GWASs and TWASs. The flexible effect size modeling assumption allows us to maintain robust and accurate prediction performance across a range of genetic architectures. Different from the MCMC approach used previously to fit BSLMM, however, our method relies on a simple deterministic search algorithm to yield an approximate analytic solution using summary statistics only, which, when paired with other algebraic innovations, results in orders of magnitude of computational speed improvement and physical memory savings as compared to BSLMM. We refer to our methods as the Deterministic BSLMM (DBSLMM). We examine the performance of DBSLMM and compare it with other scalable PGS methods in both simulations and applications to 25 traits in UKB.

## Material and Methods

### Overview of DBSLMM

DBSLMM is described in detail in the Supplemental Material and Methods. Briefly, DBSLMM follows closely the effect size assumption made in BSLMM, an accurate but unscalable method commonly used for PGS construction[15] and TWAS applications.[28,44,47] In particular, DBSLMM assumes that all SNPs have non-zero effects on the phenotype (i.e., in line with the polygenic/omnigenic assumption[58]), but that some SNPs have larger effect sizes than the others (i.e., in line with the core gene concept in the omnigenic model[58]). The DBSLMM modeling assumption on SNP effect size distribution represents a hybrid between the sparse modeling assumption (e.g., in Bayesian variable selection model[59]) and the polygenic/omnigenic modeling assumption (e.g., in linear mixed models[26]). By including both the sparse model and the polygenic model as special cases, the DBSLMM modeling assumption allows for adaptive PGS construction according to the underlying genetic architecture and thus can achieve accurate prediction performance across a range of phenotypes.[15] Different from BSLMM, however, DBSLMM does not rely on individual-level genotypes and phenotypes. Instead, DBSLMM requires only summary statistics in terms of marginal z-scores and a SNP correlation matrix that is constructed either directly from the data through a subsampling strategy or based on an external reference panel.[60] In addition, different from BSLMM, DBSLMM avoids the time-consuming MCMC algorithm for parameter estimation. Instead, DBSLMM first relies on a scalable deterministic searching algorithm to efficiently and effectively select a subset of SNPs with potentially large effects. Then, DBSLMM obtains the SNP effect size estimates for all genome-wide SNPs through an analytic solution. When further paired with a block-diagonal matrix approximation to the SNP correlation matrix as well as a fast preconditioned conjugate gradient algorithm for solving linear systems, the analytic solution allows us to construct PGS and performs genetic predictions of phenotypes in a computationally efficient fashion.[61,62] With these innovations, the computational complexity of DBSLMM becomes approximately linear respective to both the sample size and the SNP number, making DBSLMM scalable to biobank scale data with hundreds of thousands of individuals and tens of millions of SNPs. DBSLMM is freely available (see Web Resources).

### Compared Methods

We compare DBSLMM with four previously developed PGS methods that can be applied to analyze UKB scale data. All these four methods make use of summary statistics.

The first method is C+T, which uses informed clumping and p value thresholding. We use the PLINK software (v.1.90b6.9) to perform clumping and thresholding, where we set the region size to be 1 MB and the LD threshold to be $r^2 = 0.1$. In clumping, we explore ten different p value thresholds according to Lloyd-Jones et al.:[29] $5 \times 10^{-8}$, $1 \times 10^{-6}$, $1 \times 10^{-4}$, 0.001, 0.01, 0.05, 0.1, 0.2, 0.5, and 1.0. For example, when the p value threshold is 0.1, we use the PLINK *clump* command as *–clump-kb 1000 –clump-r2 0.1 –clump-p1 0.1*. We then use cross-validation to obtain the optimal p value threshold in both simulations and real data applications. With clumping, in the real data applications, the total number of SNPs analyzed ranges from 35 (for RA) to 157,269 (for CAD) in UKB.

The second method is LDpred. We use the LDpred python software (v.1.0.1) for fitting. LDpred contains two tuning parameters: the radius parameter and the non-zero effect proportion parameter. For the radius parameter, we set it to be the recommended value ($m$/3,000) in the simulations, with $m$ being the number of SNPs.[13] In the real data, due to memory and computational time constraints, we set the radius parameter to be 200, close to that used in Lloyd-Jones et al.[29] For the non-zero effect proportion parameter, we follow Vilhjálmsson et al.[13] and Lloyd-Jones et al.[29] and explore nine different choices for the non-zero effect proportion parameter: 1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 0.0003, and 0.0001. We used cross-validation to obtain the optimal proportion parameter among the nine choices in both simulations and real data applications. Besides the regular LDpred, we also examined a special case, LDpred-inf, which is

based on an infinitesimal model. LDpred-inf and LDpred by setting the proportion = 1 are the same model but have different fitting algorithms (analytic solution for LDpred-inf and MCMC for LDpred with proportion = 1). LDpred-inf does not require cross-validation.

The third method is SBLUP. We use the GCTA software (v.1.92.2) to fit SBLUP.[27] SBLUP requires the user to specify the LD window size and input a heritability estimate. We set LD window size to be 200, the maximum number we can run given the CSG cluster memory and computation time constraints. For the heritability parameter, we follow Lloyd-Jones et al.[29] and obtain the heritability estimate using LD score regression (LDSC).[63]

The fourth method is lassosum. We use the lassosum R package (v.0.4.4) for fitting. We use the same block-wise LD matrix computed for DBSLMM to serve as the reference LD matrix for lassosum.[62] We use cross-validation to obtain the optimal penalty parameter and weight of the LD reference panel for lassosum.

Besides the above four PGS methods, in a small-scale simulation, we also compare our method with the standard BSLMM. We use the GEMMA software (v.0.98) to fit the standard BSLMM, which uses MCMC for parameter inference.[15] To allow for efficient computation, we set the number of burn-in steps in MCMC as 6,000 and the number of MCMC steps as 2,000. Note that the standard BSLMM requires individual-level genotype and phenotype data.

Finally, after obtaining the effect size estimates in the training data, we calculate PGS in the test set using the *score* function in PLINK.[57]

## Simulations

We performed simulations to examine the performance of DBSLMM and compared it with existing approaches. To do so, we randomly obtained 12,000 individuals with European ancestry from the UKB data (data description in the next section). We selected the first 100,000 SNPs on chromosome 1 for these individuals. Among these SNPs, we randomly selected causal ones, simulated their effect sizes, and generated phenotypes, all using the phenotype simulation tool in GCTA.[64] To cover a range of possible genetic architectures, we considered four different simulation scenarios: scenario I (polygenic), scenario II (sparse), scenario III (hybrid I), and scenario IV (hybrid II).

Scenario I is a polygenic scenario, where all SNPs are assumed to be causal. Scenario II is a sparse scenario, where we randomly selected 0.1% SNPs to be causal. In both scenarios I and II, we simulated the causal SNP effects from the same distribution. We considered three different effect size distributions: a normal distribution $N(0,\ h^2/m)$, a scaled *t*-distribution with four degrees of freedom $t(df = 4, h^2/m)$ (in line with df = 4 in Bayesian alphabetic models), and a Laplace distribution $L\left(0, \sqrt{\frac{h^2}{2m}}\right)$; here $m$ is again the number of SNPs and $(h^2/m)$ is the per-SNP variance for all three distributions. The parameters in the three distributions were set so that the causal SNPs in total explain a fixed proportion of phenotypic variance, $h^2$ (i.e., SNP heritability). We set $h^2$ to be either 0.1, 0.2, or 0.5, representing low, moderate, and high SNP heritability, respectively. In total, we explored nine different simulation settings for each of scenarios I and II (3 SNP heritability settings × 3 distributions).

Scenarios III and IV are two hybrid scenarios of the previous two scenarios. In scenarios III and IV, all SNPs are causal, but their effect sizes come from a mixture of two different distributions. In particular, we randomly selected 0.1% of SNPs to have either moderate (scenario III) or large (scenario IV) effects while assigning the remaining SNPs to have small effects. We set the proportion of genetic variance explained by the large-effects term (PGE) to be 0.2 in scenario III and 0.5 in scenario IV. We first simulated the large effects from one of the three different distributions (normal, *t*, or Laplace) with variance $PGEh^2/m_l$, where $m_l$ is the number of large effect SNPs. We then simulated the additional effects for all SNPs from the distribution similar to large effect SNPs with variance $(1 - PGE)h^2/m$. We again set $h^2$ to be either 0.1, 0.2, or 0.5. Therefore, we explored a total of nine simulation settings for each of scenario III and IV (3 SNP heritability settings × 3 distributions).

In each simulation setting, we performed ten simulation replicates. The small number of simulation replicates is due to computational reasons and follows exactly that of Mak et al.,[16] Privé et al.,[25] and Lloyd-Jones et al.[29] In each replicate, we divided samples into three different datasets: a training set with 10,000 individuals, a validation set with 1,000 individuals, and a test set with the remaining 1,000 individuals. The proportion of samples in the three sets follows exactly that of Lloyd-Jones et al.[29] Besides these three datasets, we also randomly selected 500 individuals from the validation data and treated them as a reference panel for computing the LD matrix. With the simulated data, we first applied the linear regression model implemented in the GEMMA software to the training data to obtain marginal z-scores.[65] We then paired the marginal z-scores from the training set with the LD correlation matrix computed from the reference panel and fitted different PGS methods. During the fitting process, whenever necessary (i.e., for LDpred, lassosum, and C+T), we used the validation set for parameter tuning. Due to the small number of SNPs in simulations, we found it challenging to obtain a reasonably accurate SNP heritability estimate from standard SNP heritability estimation tools such as LDSC. Therefore, we followed Lloyd-Jones et al.[29] and supplied the true SNP heritability for all compared methods.

Besides these main simulations, we also performed small-scale simulations to directly compare our method with BSLMM. The main small-scale simulation consisted of 2,200 individuals, with 2,000 in the training set and 200 in the test set. The detailed simulation and PGS construction steps followed what has been described above. In particular, we used normal distribution to simulate causal SNP effect sizes and explored the three SNP heritability settings. Besides the main small-scale simulations, we also explored six other small-scale simulations with different sample sizes in the training data: 200, 500, 1,000, 2,000, 5,000, and 10,000. In these side small-scale simulations, we explored only scenario I and measured computational time and memory usage for BSLMM and DBSLMM.

## UKB Data

We obtained genotype and phenotype data from the UKB. The UKB data contain genotype information for 502,618 individuals. We followed the same quality control (QC) procedures described in Lloyd-Jones et al.,[29] UK Biobank – Neale lab, and UK Biobank (see Web Resources) for sample and SNP QC. An overview of the QC process is displayed in Figure S1. Specifically, for sample QC, we retained individuals (1) who have genotypes successfully measured, (2) who are included in the genotype principal component (PC) computation, and (3) who have a white British ancestry,

as is evident either through self-reporting "White British" or being within 20 standard deviations away from the European cluster center based on two leading PCs. In addition, we excluded individuals (1) who have more than ten putative third-degree relatives based on the kinship table, (2) who have sex chromosome aneuploidy, as is evident by the inconsistency between the reported sex and the sex inferred through genotypes, and (3) who are redacted and thus do not have a corresponding ID in the phenotype data. For SNP QC, we focused our analysis on autosome SNPs following Lloyd et al.[29] and the UK Biobank 2015 release (see Web Resources). We retained SNPs with a high genotype calling confidence, as is evident by the maximum probability across the three genotypes being larger than 0.9. We filtered out SNPs (1) with a minor allele frequency (MAF) $< 0.01$, (2) with a Hardy-Weinberg equilibrium (HWE) test p value $< 10^{-7}$, (3) with an imputation information score $< 0.8$, (4) with a proportion of missingness ($P_m$) $> 0.05$, or (5) that are a duplicated SNP. After these QC steps, we retained a total of 337,198 individuals and 9,428,411 SNPs for analysis.

Besides genotype information, we also obtained traits. Specifically, following Márquez-Luna et al.,[14] Privé et al.,[25] Lloyd-Jones et al.,[29] Kichaev et al.,[66] and the Neale lab (see Web Resources), we obtained 16 continuous traits that have an observed SNP heritability estimated to be above 0.1 and 9 binary traits that have a prevalence between 0.01 and 0.3. The 16 continuous traits include standing height (SH, n = 335,473), platelet count (PLT, n = 326,219), bone mineral density (BMD, n = 193,397), basal metabolic rate (BMR, n = 330,306), body mass index (BMI, n = 335,106), red blood cell count (RBC, n = 326,220), age at menarche (AM, n = 180,061), RBC distribution width (RDW, n = 326,218), eosinophils count (EOS, n = 325,653), white blood cell count (WBC, n = 326,216), forced vital capacity (FVC, n = 306,637), forced expiratory volume (FEV1) versus FVC ratio (FFR, n = 306,637), waist-hip ratio (WHR, n = 335,568), neuroticism score (NS, n = 273,107), systolic blood pressure (SBP, n = 313,972), and years of education (YE, n = 225,898) (see Neale lab in Web Resources).[14,29,66] The nine binary traits include prostate cancer (PRCA, n = 147,408, prevalence = 0.05), tanning ability (TA, n = 329,458, prevalence = 0.20), type II diabetes (T2D, n = 329,355, prevalence = 0.04), coronary artery disease (CAD, n = 238,284, prevalence = 0.05), rheumatoid arthritis (RA, n = 232,309, prevalence = 0.02), breast cancer (BRCA, n = 170,148, prevalence = 0.07), asthma (AS, n = 306,381, prevalence = 0.14), morning person (MP, n = 300,143, prevalence = 0.27), and depression (MDD, n = 284,252, prevalence = 0.08). Among the nine binary traits, for the seven diseases, following Privé et al.,[25] we treated either self-reported or ICD10 cases as 1 and others as 0. For TA, we treated "get very tanned" as 1 and others as 0. For MP, we treated "definitely a morning person" as 1 and others as 0. An overview of the phenotypes analyzed in the paper is shown in Table S2 for continuous traits and Table S3 for binary traits.

## Cross Validation in UKB

For one phenotype at a time, we performed 5-fold cross-validation in UKB to evaluate the performance of different PGS methods. First, we randomly selected 1,000 individuals (500 males and 500 females) to serve as a validation dataset.[13] We randomly selected 500 individuals from the validation data to serve as the reference panel in which we computed the SNP correlation matrix. Besides the validation data, we partitioned the remaining individuals randomly into five equal-sized disjoint subsets, each containing 35,975–67,154 individuals for the 16 continuous traits and 29,526–65,994 individuals for the 9 binary traits (number of individuals varies across the phenotypes). We performed cross-validation by treating four of these subsets as the training data and the remaining subset as the test data. We repeated the cross-validation process five times, with each of the five subsets used exactly once as the test data. In each cross-validation, we retained SNPs with an MAF $> 0.01$ in all these datasets, resulting in an approximately 9 million SNPs for each analysis.

In the analysis, we first obtained marginal z-scores in the training data by fitting a standard linear regression using the GEMMA software.[65] For each continuous trait, following the Neale lab (see Web Resources), we first fitted linear regression models to remove the effects of the top ten genotype PCs and sex and obtained phenotype residuals. We then transformed phenotype residuals to a standard normal distribution through quantile-quantile normalization. For each binary trait, we directly fitted linear regression models for one SNP at a time by treating the top ten genotype PCs and sex as covariates to obtain the marginal z-scores. With the marginal z-scores from the training data and the SNP correlation matrix from the reference panel, we then fitted different prediction methods to obtain SNP effect size estimates. When necessary (i.e., for LDpred, lassosum and C+T), we used the validation data to select the optimal tuning parameters. Afterward, we supplied the estimated SNP effects from different methods to the test data to construct PGS. For continuous traits, we evaluated the performance of different PGS methods in the test data using Pearson correlation ($R^2$) and mean square error (MSE; calculated by *Metrics* R package v.0.1.4). For binary traits, we evaluated the performance of different PGS methods in the test data using area under curve (AUC; calculated by *pROC* R package v.1.15.3) and Brier score (calculated by *scoring* R package v.0.6).

Besides the above analyses, we also computed a theoretical $R^2$ under the infinitesimal model in the following form:[67]

$$E(R^2) = \frac{h^2}{1 + m_i/(nh^2)} \qquad \text{(Equation1)}$$

where $n$ is sample size, $m_i$ is the number of independent SNPs included in the model, and $h^2$ is the SNP heritability. Following Yang et al.,[68] we estimated $m_i$ as the total number of SNPs divided by the mean LD score of these SNPs. We treated the expected $R^2$ in Equation 1 as a baseline prediction performance obtained using an infinitesimal model under ideal situations.

Finally, we evaluated the performance of different methods on predicting extreme phenotypes. To do so, for each continuous trait in turn, we ordered individuals by their trait values and divided them into ten equal-sized groups. We combined the first group (i.e., individuals with lowest trait values) and the tenth group (i.e., individuals with highest trait values) and performed 5-fold cross validation on them to examine the performance of different PGS methods. As a comparison, we also randomly selected an equal number of individuals (i.e., 20%) and performed 5-fold cross validation there.

## External Validation outside UKB

Besides evaluating PGS methods through cross-validations within UKB, we also evaluated the performance of different PGS methods by external validation. In particular, we trained different PGS methods in each of the five training sets in UKB as described above

and validated their performance in two external datasets with summary statistics. In addition, besides training PGS methods in each of the five training sets, we also performed side analysis where we trained PGS methods using the entire UKB data consisting of all five folds to validate their performance in external data.

The first external data consists of GWAS summary statistics for individuals with European ancestry. The data were obtained from GWAS-ALTAS.[69] We focused on phenotypes that are analyzed in the present study, that contain summary statistics with allele information, and that are measured on non-UKB samples. With the three criteria, we obtained six traits that include SH (n = 253,288),[70] PLT (n = 4,250),[71] BMI (n = 339,224),[72] RBC (n = 4,250),[71] EOS (n = 4,250),[71] and WBC (n = 4,250).[71] Here, SH was adjusted for the first 20 PCs. BMI was adjusted for age, age$^2$, and any necessary specific covariates (i.e., genotype-derived PC). Four blood measurements were adjusted for age, sex, and time of blood collection (including a square component). The details of the GWAS summary statistics for the six traits are provided in Table S4. We intersected SNPs for the six traits, resulting in an overlap set of 1,765,807 SNPs for analysis.

The second external data consists of GWAS summary statistics from the BBJ on individuals with East Asian ancestry.[73–75] We obtained summary statistics for the same six traits listed above: SH (n = 159,095),[74] PLT (n = 108,208),[73] BMI (n = 158,284),[75] RBC (n = 108,794),[73] EOS (n = 62,076),[73] and WBC (n = 107,964).[73] Here, SH in BBJ was adjusted for age, age$^2$, sex, and top ten genotype PCs. BMI and four blood measurements were adjusted for age, sex, top ten genotype PCs, and disease status (affected versus non-affected) for the 47 target diseases in BBJ, as well as any necessary trait-specific covariates. The details of GWAS summary statistics for the six traits are provided in Table S5. As above, we intersected SNPs for the six traits, resulting in an overlap set of 5,654,625 SNPs for analysis.

We examined each trait in each external dataset one at a time. Because of the different SNP number of the UKB data and the two external validation datasets, for each trait in the external data, we used the common SNPs that appeared in both the training data of UKB and the external test data for model fitting. We aligned SNP alleles to be consistent between the training data in UKB and the external test data. We then applied the fitted methods to the external test data and evaluated their prediction performance. Because the two external data consist only of summary statistics, we relied on the following strategy to evaluate prediction performance. Specifically, we denoted the unobserved individual-level phenotype vector in the external data as $\tilde{\boldsymbol{y}}$, the unobserved individual-level genotype matrix as $\tilde{\boldsymbol{X}}$, the observed summary statistics in terms of z-scores as $\tilde{\boldsymbol{z}}$, and the sample size as $\tilde{n}$. We first applied the PGS method to the UKB training/validation data and obtained $\widehat{\beta}_j$ as the estimated effect size of $j^{\text{th}}$ SNP. If we had observed the individual-level genotype matrix in the external data, we would have constructed PGS directly as $\widehat{\tilde{\boldsymbol{y}}} = \tilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}$. Subsequently, we would have been able to evaluate the prediction performance by computing $R^2$, with $R$ being

$$R = cor\left(\tilde{\boldsymbol{y}}, \widehat{\tilde{\boldsymbol{y}}}\right) = \frac{cov\left(\tilde{\boldsymbol{y}}, \widehat{\tilde{\boldsymbol{y}}}\right)}{\sqrt{var\left(\tilde{\boldsymbol{y}}\right) var\left(\widehat{\tilde{\boldsymbol{y}}}\right)}} = \frac{\frac{1}{\tilde{n}}\tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}}{\sqrt{\frac{1}{\tilde{n}}\widehat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}}} = \frac{\tilde{\boldsymbol{z}}^T\widehat{\boldsymbol{\beta}}}{\sqrt{\widehat{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}}} \quad \text{(Equation 2)}$$

where $\boldsymbol{\Sigma} = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}/\tilde{n}$ is the SNP correlation matrix in the test data. In the above equation, we assumed that the phenotype vector and each column of the genotype matrix in the test data were centered and standardized to have a mean of zero and a standard deviation of one. The above equation allowed us to compute $R^2$ in the test data using summary statistics in the form of $\tilde{\boldsymbol{z}}$ and a SNP correlation matrix $\boldsymbol{\Sigma}$. Note that a similar version of the above equation is also provided in a recent preprint.[19] In order to compute test $R^2$ in Equation 2 in a computationally efficient fashion, we also approximated $\boldsymbol{\Sigma}$ with a block diagonal matrix as described in the details of DBSLMM.[62] In particular, for the first external data, because these summary statistics are from individuals of European ancestry, we used the same 500 individuals in the UKB validation data to construct the SNP correlation matrix. We also used the same block information of EUR (1,703 blocks for the whole genome)[62] to compute the block diagonal SNP correlation matrix as used in DBSLMM. For the second external data, because these summary statistics are from individuals of East Asian ancestry, we used 504 individuals with East Asian ancestry (EAS) from the 1000 Genomes Project to construct SNP correlation matrix.[76] We retained SNPs with an MAF > 0.01, an HWE test p value $< 10^{-3}$, and non-missing genotypes. We also constructed a block diagonal SNP correlation based on EAS block information (1,445 blocks for the whole genome).[62]

## Results

### Simulations

We performed simulations to evaluate the performance of DBSLMM and compare it with existing PGS approaches.

We first considered a set of small-scale simulations with 100,000 SNPs and 2,200 individuals to directly compare the prediction performance and computational requirements of DBSLMM with BSLMM. The results showed that DBSLMM was only slightly inferior to BSLMM in terms of prediction accuracy: across all four simulation scenarios, DBSLMM incurred only an average of 1.55% accuracy loss in terms of prediction $R^2$, with prediction $R^2$ values highly correlated with that from BSLMM (Pearson correlation coefficient = 0.969; Figures 1A and 1B). However, DBSLMM was 3.25 times faster than BSLMM when n = 2,000. The computational gain brought by DBSLMM over BSLMM became much more appreciable with increasing sample sizes (Figure 1C). For example, on a moderate-sized sample (n = 10,000), DBSLMM was 64.72 times faster than BSLMM, while using only a fraction of physical memory (8.0%) as BSLMM (Figure 1D).

Next, we considered a set of large-scale simulations with 12,000 individuals to compare DBSLMM with five PGS methods that are also scalable to biobank scale datasets. These other compared PGS approaches include C+T, LDpred-inf, LDpred, SBLUP, and lassosum. Here, in each simulation setting, following Lloyd-Jones et al.,[29] we divided individuals into three non-overlap sets: a training set (n = 10,000), a validation set (n = 1,000), and a test set (n = 1,000). We fitted all methods on the training set based on summary statistics, and when necessary (e.g., for LDpred, lassosum, and C+T), cross-validated the hyper-parameters in the validation set. We evaluated the prediction performance of different methods in the test set by computing $R^2$ or MSE, across ten replicates in each

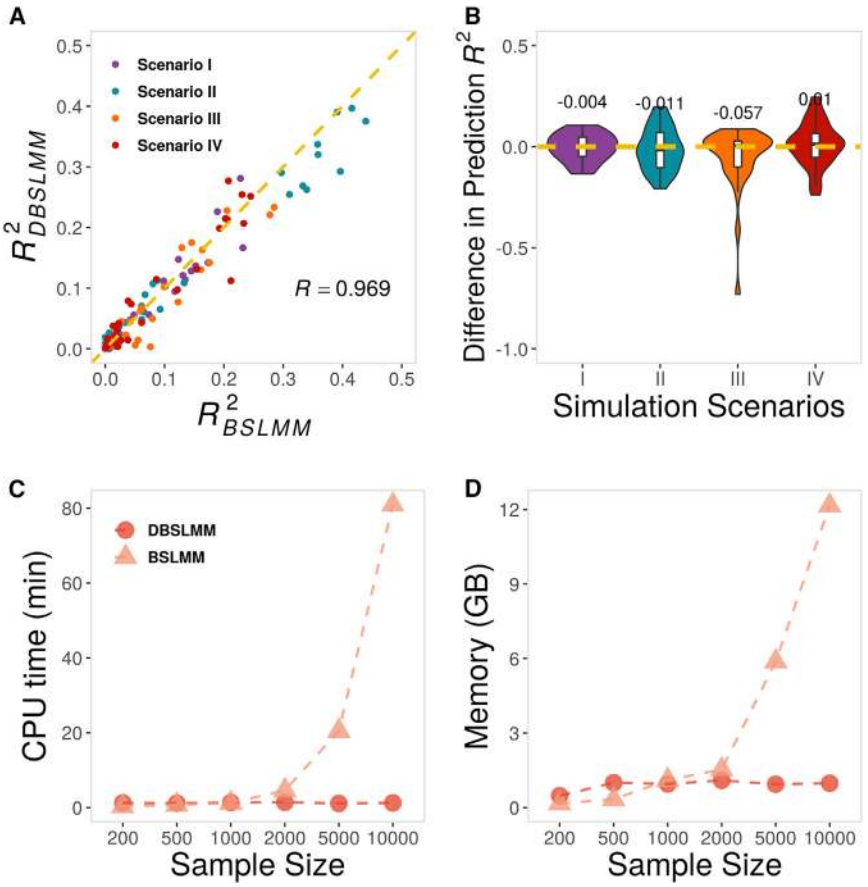


**Figure 1. Comparison between DBSLMM and BSLMM in Small-Scale Simulations**

(A) Prediction $R^2$ in the test data for DBSLMM (y axis) is similar to that for BSLMM (x axis) across four simulation scenarios (four colors). Each scenario consists of nine simulation settings, each with ten simulation replicates. Correlation between $R^2_{DBSLMM}$ and $R^2_{BSLMM}$ across all replicates and all settings is shown in the panel.

(B) Violin plot shows the scaled difference in terms of prediction $R^2$ between DBSLMM and BSLMM (y axis) for the four simulation scenarios (x axis; four colors). The scaled prediction difference is computed in each simulation replicate in each simulation setting. For the scaled prediction difference, we computed $R^2_{DBSLMM} - R^2_{BSLMM}$ and divided it with the true heritability.

(C) CPU time (y axis) for BSLMM (triangle) and DBSLMM (dot) are shown for increasing sample sizes.

(D) Memory usage (y axis) for BSLMM (triangle) and DBSLMM (dot) are shown for increasing sample sizes.

Computation in (C) and (D) is based on 100,000 SNPs and is performed with one thread of an Intel Xeon CPU E5-2683 v3.

simulation setting (Figures 2 and S2–S9 for $R^2$; Figures S10–S18 for $R^2$ difference; and Figures S19–S21 for MSE).

Overall, DBSLMM was either the most accurate (in 20 simulation settings) or the second most accurate (14) method in 34 out of 36 simulation settings in terms of the average prediction $R^2$ across simulation replicates. The performance of DBSLMM was followed by SBLUP, which was ranked as the best (11) or second best (6) method in 17 out of 36 simulation settings. lassosum (second best in 13) and C+T (best in 5 and second best in 2) also performed reasonably well. LDpred-inf was the second best method only in one simulation setting while LDpred was neither the best nor the second best in any of these simulation settings, which may partially reflect the unstable MCMC fitting algorithm underlie LDpred as demonstrated in a previous study.[25] In the 20 settings where DBSLMM was the best, on average, DBSLMM was 7.06% more accurate (median = 6.28%, range = 0.26%–18.40%) than the second best method. In the 16 settings where DBSLMM was not the best, on average, DBSLMM was 3.72% less accurate (median = 3.37%, range = 0.03%–8.12%) than the best method. Compared to each individual PGS method, across all simulation settings, DBSLMM on average improved prediction accuracy upon SBLUP, LDpred-inf, LDpred, lassosum, and C+T by 155.24%, 35.69%, 27.87%, 10.59%, and 35.84%, respectively.

Importantly, the performance of DBSLMM was relatively stable across different genetic architectures. Specifically, in the polygenic and the two-hybrid scenarios (scenarios I, III, and IV), the performance of DBSLMM was similar to or better than the polygenic models LDpred-inf and SBLUP, whereas the sparse models LDpred, C+T, and lassosum did not fare well (Figures 2A, 2C, and 2D). In the sparse scenario (scenario II), the performance of DBSLMM was also similar to or better than LDpred, C+T, and lassosum, whereas the polygenic models LDpred-inf and SBLUP did not fare well (Figure 2B). The overall robust performance of DBSLMM across genetic architectures is likely due to its flexible modeling assumption on the SNP effect sizes, which allows DBSLMM to be adaptive to the underlying genetic architecture and achieve robust performance across different settings.

A careful examination of the simulation results provides further insights. First, the results based on MSE were generally consistent with that based on $R^2$ for most PGS methods, with the only exception of lassosum. For lassosum, its MSE in the test data was extremely large in non-sparse scenarios (I, III, and IV), often orders of magnitudes larger than that of the other methods. For example, in the baseline setting (i.e., $h^2 = 0.1$ and normal effect size distribution), the average MSE for lassosum across simulation replicates were 189.767, 1.585, 98.732, and 56.224, for scenarios I–IV, respectively (Figure S19). As a comparison, the average MSE for DBSLMM was only 0.998, 1.028, 0.986, and 0.991. The poor performance of lassosum in terms of

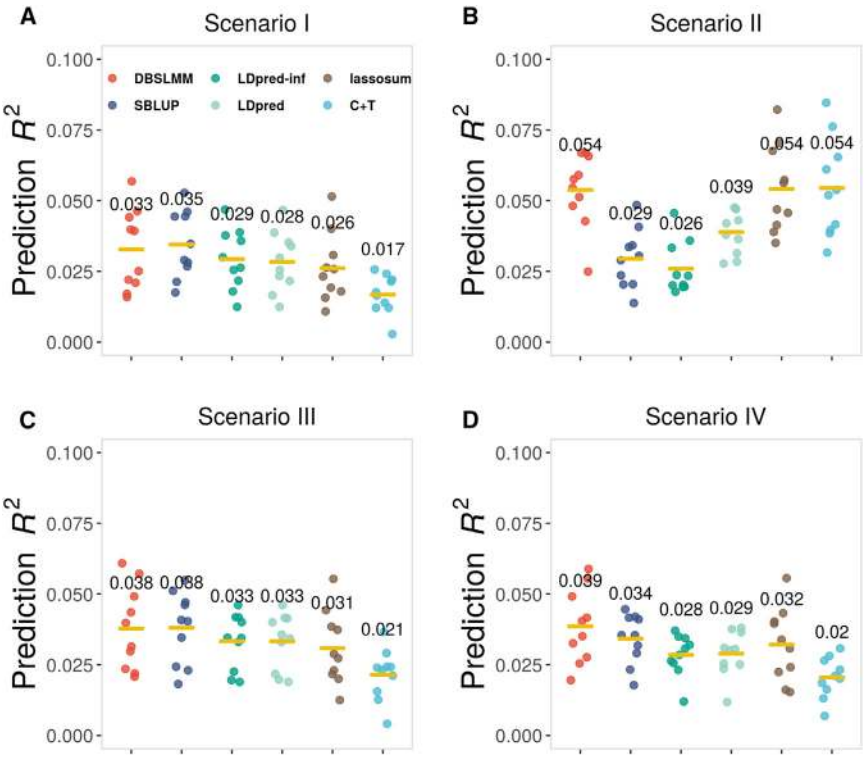


**Figure 2. Comparison of Six PGS Methods in Their Prediction Performance in Large-Scale Simulations**

Jitter plots show the prediction $R^2$ across ten replicates for different methods in each the four simulation scenarios. Compared methods include DBSLMM (orange red), SBLUP (steel blue), LDpred-inf (dark cyan), LDpred (light green), lassosum (sienna), C+T (medium turquoise). Results are shown for the baseline simulation setting with a normal effect size distribution and with heritability = 0.1. Solid lines represent the mean of prediction $R^2$ across ten replicates, with the numerical number also displayed above each method.

MSE under non-sparse settings likely reflects a mismatch between the sparse modeling assumption made in lassosum and the actual polygenic genetic architecture. Indeed, while the SNP effect size estimates from lassosum are reasonably accurate under sparse scenario (II), they become rather inaccurate under non-sparse scenarios (I, III, IV; Figure S22). The estimation inaccuracy of lassosum under non-sparse scenarios also appears to be dependent on both LD score (Figure S23) and MAF (Figure S24).

Second, among the three parameters examined in the simulations (i.e., the SNP heritability $h^2$, effect size distribution, and PGE in scenarios III and IV), two of them ($h^2$ and PGE) influenced the performance of different PGS methods. For $h^2$, the performance of different PGS methods, with one exception (SBLUP), generally improved with increasingly high $h^2$. For SBLUP, its performance increased when $h^2$ increased from 0.1 to 0.2 but became stable or slightly reduced when $h^2$ increased further to 0.5. For example, in the baseline setting in scenario I, the prediction $R^2$ for SBLUP was 0.035, 0.088, and 0.070, for $h^2 = 0.1$, 0.2, 0.5, respectively (Figures 1, S2, and S3). For $\rho$, the performance of different PGS methods, with one exception (SBLUP), also generally improved with increasingly large PGE. For SBLUP, its prediction performance dependency on PGE changed with respect to the effect size distribution: its performance decreased with respect to $\rho$ under a normal or a Laplace distribution, but increased under a *t*-distribution (Figures 1, S4, and S7).

Finally, we examined the computing time and memory requirement of different methods in the simulated data with SNP number ranging from 10,000 to ~720,000 (all SNPs on chromosome 1; Figure 3). Among the compared methods, DBSLMM was faster than SBLUP, LDpred, and lassosum, with computational gain increases with increasing number of SNPs. For example, it took SBLUP, LDpred, and lassosum 104.48, 154.12, and 18.51 min, respectively, to analyze ~720,000 SNPs on chromosome 1. In contrast, it took DBSLMM only 7.03 min to analyze the data there, representing 14.87-, 21.94-, and 2.63-fold speed gain over these three methods. In addition, DBSLMM required only a small amount of physical memory. For example, SBLUP, LDpred, and lassosum required 57.10, 7.62, and 3.37 GB memory for analyzing ~720,000 SNPs, respectively. In contrast, DBSLMM used only 2.48 GB, which is 4.3%, 32.5%, and 73.5% of that required by the three methods. Certainly, the simplest approach, C+T, used the least amount of memory and computing time, though its performance did suffer in many polygenic settings as shown above.

### Applications to UKB: Internal Cross Validation

We applied DBSLMM and other methods to analyze UKB data. The detailed description of the 16 continuous traits is shown in Table S2, and the detailed description of the 9 binary traits is shown in Table S3. For each trait in turn, we performed 5-fold cross validation. For continuous traits, we evaluated the methods' performance by computing $R^2$ (Figure 4) and MSE (Figure S25). For binary traits, we evaluated the methods' performance by computing AUC (Figure S26) and Brier score (Figure S27).

Overall, consistent with the simulations, we found that DBSLMM achieved the best performance among all PGS methods. Specifically, it outperformed all other PGS methods in 11 of the 16 continuous traits in terms of prediction $R^2$ and in 6 of the 9 binary traits in terms of AUC. It ranked as the second-best PGS method for 3 other continuous traits and for 3 remaining binary traits. The overall performance of DBSLMM was followed by

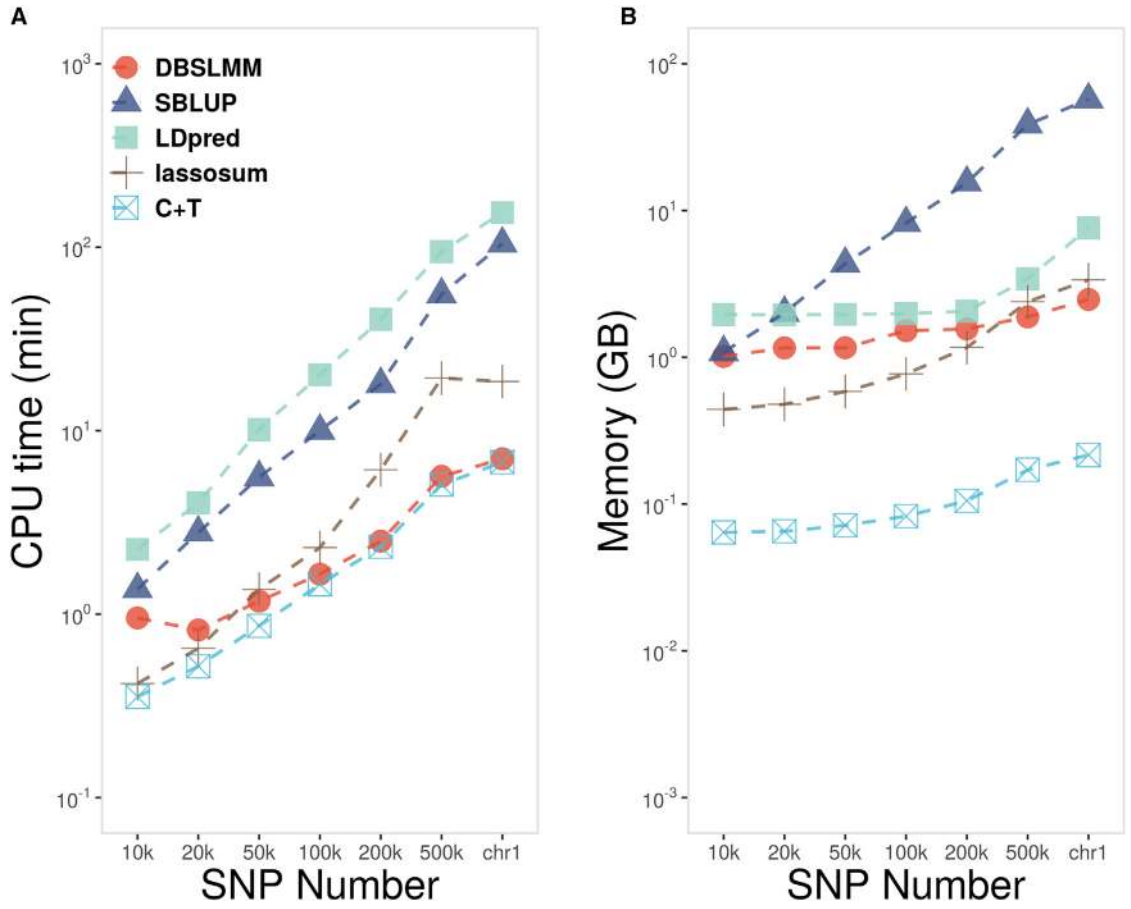


**Figure 3. Comparison of Computing Cost for Different PGS Methods**
CPU time (A) and memory cost (B) for different methods are shown with respect to increasing SNP numbers (x axis). chr1 on x axis represents ~720,000 SNPs. Compared methods include DBSLMM (orange-red), SBLUP (steel blue), LDpred (dark cyan), lassosum (sienna), C+T (medium turquoise). Computation is based on the baseline simulation setting of scenario IV and is performed with one thread of an Intel Xeon CPU E5-2683 v3. Note that the y axis is on log scale for both panels.

lassosum (best for 5 continuous traits and 1 binary trait; second-best for 8 continuous traits and 3 binary traits), SBLUP (best for 2 binary traits; second-best for 2 continuous traits and 1 binary trait), and C+T (second-best for 3 continuous traits and 1 binary trait). LDpred and LDpred-inf did not fare well: LDpred was the worst method for 9 continuous traits and 2 binary traits while LDpred-inf was the worst method for 1 continuous trait and 1 binary trait. In the 17 traits where DBSLMM was the best, on average, DBSLMM was 9.17% more accurate (median = 4.57%, range = 0.31%–27.79%) than the second-best PGS method. In the 8 traits where DBSLMM was not the best, on average, DBSLMM was 6.55% less accurate (median = 6.44%, range = 0.18% to 16.19%) than the best method. On average, for continuous traits, DBSLMM improved prediction upon SBLUP, LDpred-inf, LDpred, lassosum, and C+T by 44.35%, 93.47%, 101.09%, 8.19%, and 38.50%, respectively. For binary traits, DBSLMM improved prediction upon these methods by an average of 2.03%, 5.15%, 7.48%, 4.71%, and 5.67%, respectively. In addition, presumably due to its flexible effect size assumption, DBSLMM often outperformed the theoretical $R^2$ under an infinitesimal model with the corresponding sample size for almost all traits (Figure 4). In contrast, as expected, the two infinitesimal models (SBLUP and LDpred-inf) yielded similar prediction $R^2$ as the theoretical $R^2$.

As expected, the performances of all PGS methods in continuous traits were positively correlated with their SNP heritability (Pearson correlation in the range of 0.930–0.969 for continuous traits; Figure S28). In addition, the performance of all PGS methods improved on predicting individuals with extreme phenotypes, although their ranking remained similar (Figure S29). Moreover, consistent with the simulations, for continuous traits, the performance of lassosum measured by MSE could be much worse than that measured by $R^2$. In particular, lassosum became the worst method for 14 out of 16 continuous traits when measured by MSE. For these 14 traits, the MSE from lassosum was on average 26 times larger than the second worst method. As a specific example, lassosum performed reasonably well for SH in terms of prediction $R^2$ (0.255 for lassosum; 0.302 for DBSLMM). However, lassosum performed poorly for SH in terms of MSE (256.44 for lassosum; 1.24 for DBSLMM). Similarly, for

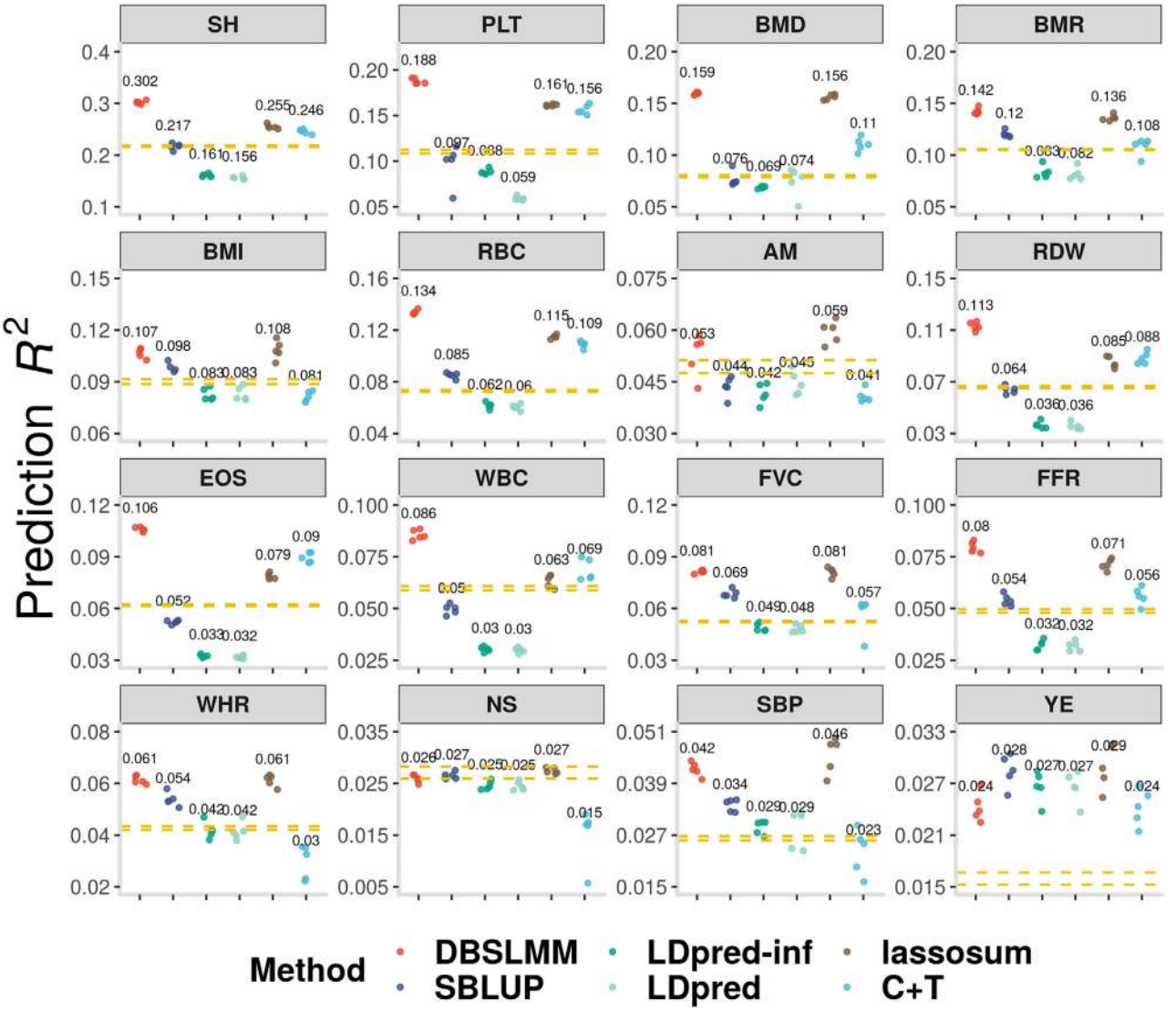


**Figure 4. Prediction Performance of Six PGS Methods for 16 Continuous Traits in UKB Cross-validation**
Methods include DBSLMM (orange-red), SBLUP (steel blue), LDpred-inf (dark cyan), LDpred (light green), lassosum (sienna), and C+T (medium turquoise). Title in each panel shows the abbreviation of 16 continuous traits: standing height (SH), platelet count (PLT), bone mineral density (BMD), basal metabolic rate (BMR), body mass index (BMI), red blood cell count (RBC), age at menarche (AM), RBC distribution width (RDW), eosinophils count (EOS), white blood cell count (WBC), forced vital capacity (FVC), forced expiratory volume (FEV1) versus FVC ratio (FFR), waist hip ratio (WHR), neuroticism score (NS), systolic blood pressure (SBP), and years of education (YE). The jitter plot in each panel displays the prediction $R^2$ for each method in the test set across five folds. The mean prediction $R^2$ across the five folds is displayed above the jitter plot. Dashed lines in each panel represent the maximum and minimum theoretical expected prediction $R^2$ under an infinitesimal model.

binary traits, the performance of lassosum was ranked as the worst method for 3 out of 9 binary traits measured by Brier score.

Finally, we examined the computing time and memory cost of different methods in the real data applications for SH in 5-fold cross validation (Table 1). With one CPU thread, DBSLMM was 19.43, 22.97, and 1.07 times faster and used 7.4%, 8.9%, and 23.4% of physical memory as compared to the three multivariable regression-based PGS methods, SBLUP, LDpred, and lassosum, respectively. DBSLMM was also implemented with parallel computing capability. With five CPU threads, DBSLMM was 28.11 and 1.03 times faster and used 18.8% and 24.8% of physical memory as compared to SBLUP and lassosum, respectively. In addition, the DBSLMM algorithm is reasonably robust with respect to the choice of the hyper-parameters: while the main analyses of DBSLMM were carried out by fixing two hyper-parameters to pre-defined values (the p value and LD threshold in the selection algorithm was set to be 1e−6 and 0.1, respectively), tuning these two hyper-parameters in the validation data yielded similar results (Figures S30 and S31).

### Applications to UKB: External Validation

The above results are based on cross-validation within UKB. To examine whether the PGS constructed by the methods

**Table 1. Computational Cost of Different PGS Methods in UKB**

| | One Thread | | Five Threads | |
|---|---|---|---|---|
| Methods | Memory Usage (GB) | CPU Time (min) | Memory Usage (GB) | CPU Time (min) |
| DBSLMM | 4.68 | 62.45 | 11.80 | 35.88 |
| SBLUP | 62.91 | 1,213.42 | 62.91 | 1,008.71 |
| LDpred | 52.70 | 1,434.38 | – | – |
| lassosum | 20.03 | 66.89 | 47.60 | 37.12 |
| C+T | 0.23 | 47.90 | – | – |

Table lists the method name (1st column), memory usage (2nd or 4th column), and computing time (3rd or 5th column) for analyzing one trait in UKB (with ~9 million SNPs). Memory usage and CPU time are recorded based on either one thread (2nd and 3rd columns) or five threads (4th and 5th columns) of an Intel Xeon CPU E5-2683 v3. LDpred and C+T do not have parallel computing capacity, so the computing time and memory usage on five threads are not recorded for these two methods.

above can be extrapolated into other datasets or other ethnic groups, we performed two external validation analyses. Briefly, we fitted different PGS models in each of the five training/validation datasets in UKB and then examined the performance of the PGS methods in each of the two external datasets: one on individuals with European ancestry (Table S4) and another on individuals with East Asian ancestry from BBJ (Table S5). Therefore, we obtained five prediction values in each external data, one from each of the five training/validation data in UKB. We focused our validation analysis on six traits, including SH, PLT, BMI, RBC, EOS, and WBC, which are available in both external datasets. Importantly, both external datasets consist only of summary statistics. Therefore, we extended the formula for computing $R^2$ in the test set to use summary statistics (Equation 2). We summarize the prediction $R^2$ results across the five different UKB training data for the two external data separately: Figure 5 for the first external data and Figure S32 for the second external data.

Overall, consistent with the simulations and UKB cross-validation results, DBSLMM was the best PGS method for five out of six traits in the first external dataset and for four out of six traits in the second external dataset, and was ranked as the second-best PGS method for one trait in the second set. The overall performance of DBSLMM was followed by SBLUP (best for two traits in the second set; second-best for one trait in the first set and two traits in the second set), C+T (second-best for three traits in the first set and one trait in the second set), LDpred-inf (second-best for two traits in the first set and one trait in the second set) and LDpred (second-best for one trait in the second set). lassosum did not fare well: it was the worst methods for three traits in the first data and one trait in the second data. For the five traits in the first data where DBSLMM performed the best, DBSLMM was on average 12.22% more accurate (median = 9.39%; range = 7.02%–21.01%) than the next best method (Figure 5). For the one trait (BMI) in first data where DBSLMM was not the best method, its performance was 18.68% less accurate than the best method. For the four traits in the second data where DBSLMM performed the best, DBSLMM was on average 12.09% more accurate (median = 11.20%; range = 0.99%–24.97%) than the next best method. For the two traits in the second data where DBSLMM was not the best method (BMI and PLT), DBSLMM was 4.76% and 18.64% less accurate than the best method, respectively. On average, in the first external data, DBSLMM improved prediction accuracy upon SBLUP, LDpred-inf, LDpred, lassosum, and C+T by 95.72%, 19.61%, 28.41%, 59.42%, and 23.65%, respectively. In the second external dataset, DBSLMM improved prediction accuracy upon these methods by 522.74%, 14.74%, 25.37%, 33.12%, and 43.79%, respectively. Finally, using the entire UKB data consisting of all five folds as the training data (instead of using only four out of the five folds) yielded consistent and slightly more accurate prediction results (Table S6).

Comparing the method performance in the external data versus that in the UKB cross-validation provides us with further insights. First, as expected, the performance of all PGS methods in the external data was worse than that in UKB cross-validation, and more so in the second external data than in the first external data (Figures 4, 5, and S32). For example, for SH, the average prediction $R^2$ for six methods (i.e., DBSLMM, SBLUP, LDpred-inf, LDpred, lassosum, and C+T) were 0.302, 0.217, 0.161, 0.156, 0.255, and 0.246 in UKB cross-validation, respectively. These prediction $R^2$ values reduced to 0.256, 0.048, 0.235, 0.176, 0.211, and 0.179 in the first external data, and further reduced to 0.118, 0.004, 0.117, 0.112, 0.078, and 0.082 in the second external data, respectively. Second, the relative performance of LDpred-inf and LDpred in the external data were slightly better than that in the UKB cross-validation. Third, while lassosum ranked the second in UKB cross-validation, it had a relatively low performance in the external validations and was ranked neither as the best nor the second-best method across all six traits in UKB data. Finally, comparing between the two external datasets, the rank of the six PGS methods were largely similar for four traits (SH, BMI, EOS, and WBC) but not the other two (PLT and RBC). For example, SBLUP was close to the worst method for PLT in the first external data but became the best in the second external data. Similarly, LDpred-inf was close to the worst method for RBC in the first set but became the third in the second set.

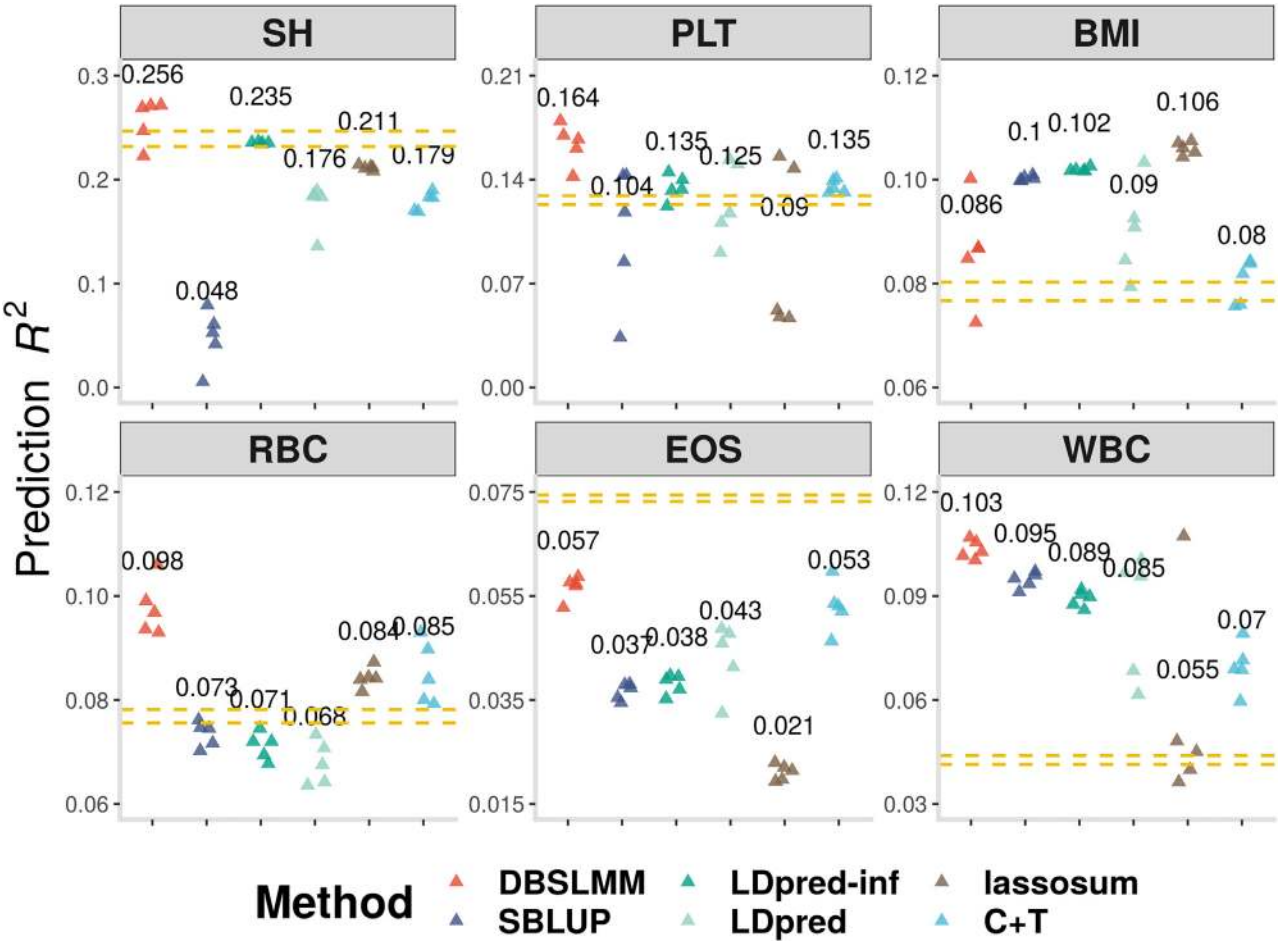


**Figure 5. Prediction Performance of Six PGS Methods for Six Continuous Traits in the External Validation Data with European Ancestry**

Compared methods include DBSLMM (orange-red), SBLUP (steel blue), LDpred-inf (dark cyan), LDpred (light green), lassosum (sienna), and C+T (medium turquoise). Title in each panel shows the abbreviation of six continuous traits: standing height (SH), platelet count (PLT), body mass index (BMI), red blood cell count (RBC), eosinophils count (EOS), and white blood cell count (WBC). The jitter plot in each panel displays the prediction $R^2$ for each method in the test set across five folds. The mean prediction $R^2$ across the five folds is displayed above the jitter plot. Dashed lines in each panel represent the maximum and minimum theoretical expected prediction $R^2$ under an infinitesimal model.

## Discussion

We have presented a new method, DBSLMM, for accurate and scalable construction of PGS in large-scale biobank datasets. The inferred SNP weights from DBSLMM not only predict well in European populations but also adapts reasonably well to East Asians. Indeed, we show that European weights can be used to predict BBJ traits, though with an approximate 3.3-fold decrease in accuracy as compared to predicting the same traits in UKB. The accuracy reduction observed in the present study is largely consistent with previous studies where the European weights incurred an approximately 2.0-fold accuracy reduction for predicting East Asians as compared to predicting Europeans for anthropometric and blood-panel traits[77] as well as for BMI and height.[78] Overall, we believe DBSLMM strikes an appealing balance between computational tractability and prediction accuracy for PGS applications.

We have compared DBSLMM with several other PGS methods, including SBLUP, LDpred, lassosum, and C+T, in both simulations and applications to UKB. In the comparison, due to computing cluster and resource limitation, we have restricted our analysis for all methods on using a maximum of 64 GB memory. Subsequently, we have to restrict the radius parameter in LDpred to be 200, the window size parameter in SBLUP to be 200, and the number of independent LD blocks in DBSLMM and lassosum to be within 2,000. These parameters effectively determine the LD distance considered in the model. With these parameter settings, we found that the prediction $R^2$ for the two polygenic models (LDpred-inf and SBLUP) remains largely consistent with the theoretical $R^2$ expected from an infinitesimal model, suggesting that these parameter choices were sufficient for LDpred-inf and SBLUP to achieve the optimal prediction performance. In addition, our results for the existing PGS methods were also largely consistent

with previous applications of the same method on the same trait.[29] However, we acknowledge that higher LD distance and larger LD block size will likely improve the prediction accuracy for at least some of the PGS methods such as DBSLMM and lassosum.

We have examined a relatively simple searching strategy to identify SNPs with potentially large effect sizes in DBSLMM. Our searching strategy is based on a clumping algorithm and is computationally efficient. However, other more sophisticated variable search and screening algorithms may lead to more accurate selection of large-effect SNPs, potentially improving PGS accuracy further. For example, lasso,[79] smoothly clipped absolute deviation (SCAD),[80] elastic net,[81] and sure independence screening (SIS)[82] can all be used to select SNPs with potentially large effects. Our analytic solution can be applied to SNPs with large effects obtained from any searching strategy. Therefore, pairing our method with other searching strategies in the future can have added benefits.

Our modeling assumption represents a direct attempt for modeling the omnigenic hypothesis that was proposed recently.[58] Specifically, our model categorizes SNPs into two groups: a small group of SNPs with large effect sizes and a large group of SNPs with small effect sizes. Such SNP categorization is equivalent to assuming that all SNPs have non-zero effects, while a small proportion of them have additional effects. The assumption that all SNPs have non-zero effects attempts to model the omnigenic hypothesis that all genes/SNPs have non-zero effects. The assumption that a small subset of SNPs have additional effects also attempts to model the omnigenic hypothesis that a small subset of genes, termed core genes, have additional effects. The set of core genes was hypothesized in the omnigenic model to directly underlie disease etiology and contribute disproportionally to disease and disease-related complex traits. In DBSLMM, we can also compute a statistic $\rho$ to quantify the proportion of genetic variance explained by large effect SNPs in the trait (details in Supplemental Material and Methods). $\rho$ is a value between 0 and 1 and effectively measures how "polygenic/omnigenic" the given trait is. Specifically, if $\rho$ is small and close to be zero, then the genetic variance of the trait is largely explained by a large number of small-effect SNPs. Consequently, a polygenic PGS model may work preferentially well for the trait. In contrast, if $\rho$ is large and close to be one, then the genetic variance of the trait is largely explained by large-effect SNPs. Consequently, a sparse PGS model may work preferentially well. Because DBSLMM can take advantage of both large- and small-effect SNPs in a data adaptive fashion, DBSLMM can work reasonably well across a range of $\rho$ values. In addition, DBSLMM becomes a sparse model as $\rho$ approaches to 1 and becomes a polygenic model as $\rho$ approaches to 0.[15] In real data applications, the performance gain brought by DBSLMM over polygenic PGS methods is highly positively correlated with the statistics $\rho$, while the performance gain brought by DBSLMM over sparse PGS methods is negatively correlated with $\rho$ (Figure S33).

Certainly, while the effect size distribution assumption in DBSLMM is relatively flexible, more flexible modeling assumptions exist. For example, DPR uses a non-parametric effect size distribution assumption that effectively categorizes SNPs into infinitely many groups *a priori*.[28] LDpred-funct and AnnoPred attempts to incorporate SNP functional annotations into modeling effect size distribution.[14,20] These different approaches can all improve prediction accuracy across a range of genetic architectures. Therefore, it would be ideal to extend the current deterministic searching strategy and analytic solution for our model to other, more flexible, PGS models. For example, we could potentially extend the searching strategy to select different groups of SNPs, with each group of SNPs having different magnitude of effect sizes, as in DPR. We could also impose different level of penalty on the effect sizes of SNPs from different groups. An analytic solution may be derived from some of these models but may be non-trivial for many others. Nevertheless, exploring the use of the deterministic searching strategy and analytic forms of solutions to other more flexible PGS models, either in line with the above methods or in other ways, will likely yield fruitful results in the future.

## Acknowledgments

This study was supported by the National Institutes of Health (NIH) grant R01HG009124 and National Science Foundation (NSF) grant DMS1712933. This study has been conducted using UK Biobank resource under Application Number 30686. UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government, and the Northwest Regional Development Agency. It has also had funding from the Welsh Assembly Government, British Heart Foundation, and Diabetes UK. The authors also thank Dr. Kirsten Herold at the UM-SPH writing lab for her helpful editorial suggestions.

## Declaration of Interests

The authors declare no competing interests.

## Web Resources

DBSLMM and GEMMA, http://www.xzlab.org/software.html
GCTA, https://gcta.freeforums.net
GWAS-ATLAS, https://atlas.ctglab.nl/
lassosum, https://github.com/tshmak/lassosum#lassosum-
LDpred, https://github.com/bvilhjal/ldpred
LDSC, https://github.com/bulik/ldsc
PLINK, https://www.cog-genomics.org/plink/
UKB, https://www.ukbiobank.ac.uk

UK Biobank – Neale lab, http://www.nealelab.is/uk-biobank

UK BioBank, http://www.ukbiobank.ac.uk/wp-content/uploads/2017/07/ukb_genetic_file_description.txt

UK BioBank 2015 Release, http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf

## References

1. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.
2. Owens, D.K., Davidson, K.W., Krist, A.H., Barry, M.J., Cabana, M., Caughey, A.B., Doubeni, C.A., Epling, J.W., Jr., Kubik, M., Landefeld, C.S., et al.; US Preventive Services Task Force (2019). Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer: US Preventive Services Task Force Recommendation Statement. JAMA *322*, 652–665.
3. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.
4. So, H.-C., Kwan, J.S., Cherny, S.S., and Sham, P.C. (2011). Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. Am. J. Hum. Genet. *88*, 548–565.
5. Toulopoulou, T., Zhang, X., Cherny, S., Dickinson, D., Berman, K.F., Straub, R.E., Sham, P., and Weinberger, D.R. (2019). Polygenic risk score increases schizophrenia liability through cognition-relevant pathways. Brain *142*, 471–485.
6. de Los Campos, G., Vazquez, A.I., Hsu, S., and Lello, L. (2018). Complex-trait prediction in the era of big data. Trends Genet. *34*, 746–754.
7. Khera, A.V., Chaffin, M., Wade, K.H., Zahid, S., Brancale, J., Xia, R., Distefano, M., Senol-Cosar, O., Haas, M.E., Bick, A., et al. (2019). Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. Cell *177*, 587–596.e9.
8. de los Campos, G., Gianola, D., and Allison, D.B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. *11*, 880–886.
9. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348.
10. Selzam, S., Ritchie, S.J., Pingault, J.-B., Reynolds, C.A., O'Reilly, P.F., and Plomin, R. (2019). Comparing Within- and Between-Family Polygenic Score Prediction. Am. J. Hum. Genet. *105*, 351–363.
11. Fritsche, L.G., Beesley, L.J., VandeHaar, P., Peng, R.B., Salvatore, M., Zawistowski, M., Gagliano Taliun, S.A., Das, S., LeFaive, J., Kaleba, E.O., et al. (2019). Exploring various polygenic risk scores for skin cancer in the phenomes of the Michigan genomics initiative and the UK Biobank with a visual catalog: PRSWeb. PLoS Genet. *15*, e1008202.
12. Wray, N.R., Kemper, K.E., Hayes, B.J., Goddard, M.E., and Visscher, P.M. (2019). Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. Genetics *211*, 1131–1141.
13. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592.
14. Márquez-Luna, C., Gazal, S., Loh, P.-R., Furlotte, N., Auton, A., and Price, A.L. (2018). Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. bioRxiv. https://doi.org/10.1101/375337.
15. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.
16. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., and Sham, P.C. (2017). Polygenic scores via penalized regression on summary statistics. Genet. Epidemiol. *41*, 469–480.
17. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776.
18. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P.; and International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
19. Zhao, Z., Yi, Y., Wu, Y., Zhong, X., Lin, Y., Hohman, T.J., Fletcher, J., and Lu, Q. (2019). Fine-tuning Polygenic Risk Scores with GWAS Summary Statistics. bioRxiv. https://doi.org/10.1101/810713.
20. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput. Biol. *13*, e1005589.
21. Hu, Y., Lu, Q., Liu, W., Zhang, Y., Li, M., and Zhao, H. (2017). Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS Genet. *13*, e1006836.
22. Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: Polygenic Risk Score software. Bioinformatics *31*, 1466–1468.
23. Choi, S.W., and O'Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience *8*, 8.
24. Speed, D., and Balding, D.J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. *24*, 1550–1557.
25. Privé, F., Vilhjálmsson, B.J., Aschard, H., and Blum, M.G.B. (2019). Making the Most of Clumping and Thresholding for Polygenic Scores. Am. J. Hum. Genet. *105*, 1213–1221.
26. VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. *91*, 4414–4423.
27. Robinson, M.R., Kleinman, A., Graff, M., Vinkhuyzen, A.A., Couper, D., Miller, M.B., Peyrot, W.J., Abdellaoui, A., Zietsch, B.P., and Nolte, I.M. (2017). Genetic evidence of assortative mating in humans. Nat. Hum. Behav., 1.
28. Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat. Commun. *8*, 456.
29. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. *10*, 5086.

30. So, H.-C., and Sham, P.C. (2017). Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Sci. Rep. *7*, 41262.
31. Gibson, G. (2019). On the utilization of polygenic risk scores for therapeutic targeting. PLoS Genet. *15*, e1008060.
32. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. *19*, 581–590.
33. Torkamani, A., and Topol, E. (2019). Polygenic Risk Scores Expand to Obesity. Cell *177*, 518–520.
34. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.
35. Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J.P., Chen, T.-H., Wang, Q., Bolla, M.K., et al.; ABCTB Investigators; kConFab/AOCS Investigators; and NBCS Collaborators (2019). Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am. J. Hum. Genet. *104*, 21–34.
36. Fritsche, L.G., Gruber, S.B., Wu, Z., Schmidt, E.M., Zawistowski, M., Moser, S.E., Blanc, V.M., Brummett, C.M., Kheterpal, S., Abecasis, G.R., and Mukherjee, B. (2018). Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. Am. J. Hum. Genet. *102*, 1048–1061.
37. Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B., and de los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. PLoS Genet. *7*, e1002051.
38. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al.; LifeLines Cohort Study (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.
39. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
40. Young, A.I. (2019). Solving the missing heritability problem. PLoS Genet. *15*, e1008222.
41. Rosenberg, N.A., Edge, M.D., Pritchard, J.K., and Feldman, M.W. (2018). Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. Evol. Med. Public Health *2019*, 26–34.
42. Habier, D., Fernando, R.L., Kizilkaya, K., and Garrick, D.J. (2011). Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics *12*, 186.
43. Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics *157*, 1819–1829.
44. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
45. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., Im, H.K.; and GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
46. Nagpal, S., Meng, X., Epstein, M.P., Tsoi, L.C., Patrick, M., Gibson, G., De Jager, P.L., Bennett, D.A., Wingo, A.P., Wingo, T.S., and Yang, J. (2019). TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. Am. J. Hum. Genet. *105*, 258–266.
47. Yuan, Z., Zhu, H., Zeng, P., Yang, S., Sun, S., Yang, C., Liu, J., and Zhou, X. (2019). Testing and controlling for horizontal pleiotropy with the probabilistic Mendelian randomization in transcriptome-wide association studies. bioRxiv. https://doi.org/10.1101/691014.
48. Cheng, Q., Yang, Y., Shi, X., Yang, C., Peng, H., and Liu, J. (2019). MR-LDP: a two-sample Mendelian randomization for GWAS summary statistics accounting linkage disequilibrium and horizontal pleiotropy. bioRxiv. https://doi.org/10.1101/684746.
49. Richardson, T.G., Harrison, S., Hemani, G., and Davey Smith, G. (2019). An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. eLife *8*, e43657.
50. Choi, S.W., Heng Mak, T.S., and O'Reilly, P.F. (2018). A guide to performing Polygenic Risk Score analyses. bioRxiv. https://doi.org/10.1101/416545.
51. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.
52. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al.; BioBank Japan Cooperative Hospital Group (2017). Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. *27* (3S), S2–S8.
53. Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., Li, L.; and China Kadoorie Biobank (CKB) collaborative group (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int. J. Epidemiol. *40*, 1652–1666.
54. Locke, A.E., Steinberg, K.M., Chiang, C.W.K., Service, S.K., Havulinna, A.S., Stell, L., Pirinen, M., Abel, H.J., Chiang, C.C., Fulton, R.S., et al.; FinnGen Project (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. Nature *572*, 323–328.
55. Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., Dishman, E.; and All of Us Research Program Investigators (2019). The "All of Us" Research Program. N. Engl. J. Med. *381*, 668–676.
56. Kim, H., Grueneberg, A., Vazquez, A.I., Hsu, S., and de Los Campos, G. (2017). Will Big Data Close the Missing Heritability Gap? Genetics *207*, 1135–1145.
57. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
58. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An expanded view of complex traits: from polygenic to omnigenic. Cell *169*, 1177–1186.
59. Guan, Y., and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann. Appl. Stat. *5*, 1780–1815.

60. Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann. Appl. Stat. *11*, 2027–2051.
61. Kaasschieter, E.F. (1988). Preconditioned conjugate gradients for solving singular systems. J. Comput. Appl. Math. *24*, 265–275.
62. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285.
63. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
64. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
65. Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods *11*, 407–409.
66. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging polygenic functional enrichment to improve GWAS power. Am. J. Hum. Genet. *104*, 65–75.
67. Daetwyler, H.D., Villanueva, B., and Woolliams, J.A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE *3*, e3395.
68. Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A.V., Ingelsson, E., O'Connell, J.R., Mangino, M., et al.; GIANT Consortium (2011). Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. *19*, 807–812.
69. Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. *51*, 1339–1348.
70. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; and LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
71. Ferreira, M.A., Hottenga, J.-J., Warrington, N.M., Medland, S.E., Willemsen, G., Lawrence, R.W., Gordon, S., de Geus, E.J., Henders, A.K., Smit, J.H., et al. (2009). Sequence variants in three loci influence monocyte counts and erythrocyte volume. Am. J. Hum. Genet. *85*, 745–749.
72. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al.; LifeLines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; and International Endogene Consortium (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206.
73. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.
74. Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., Matsuda, K., Ikegawa, S., Takahashi, A., Kanai, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. Nat. Commun. *10*, 4393.
75. Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat. Genet. *49*, 1458–1467.
76. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
77. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.
78. Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P.M., and Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. bioRxiv. https://doi.org/10.1101/2020.01.14.905927.
79. Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. B *58*, 267–288.
80. Fan, J., and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. J. Am. Stat. Assoc. *96*, 1348–1360.
81. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. *67*, 301–320.
82. Fan, J., and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. Ann. Stat. *38*, 3567–3604.