Polygenic Prediction via Bayesian Regression and Continuous Shrinkage Priors

Tian Ge[1,2,3], Chia-Yen Chen[1,2,3,4], Yang Ni[5], Yen-Chen Anne Feng[1,2,3,4], Jordan W. Smoller[1,2,3]

[1]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; [2]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA; [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02138, USA; [4]Analytic and Translational Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; [5]Department of Statistics, Texas A&M University, College Station, TX 77843, USA

Address correspondence to:

Tian Ge

Psychiatric and Neurodevelopmental Genetics Unit

Center for Genomic Medicine

Massachusetts General Hospital

Email: tge1@mgh.harvard.edu

Abstract

Polygenic prediction has shown promise in identifying individuals at high risk for complex diseases, and may become clinically useful as the predictive performance of polygenic risk scores (PRS) improves. Here, we present PRS-CS, a novel polygenic prediction method that infers posterior SNP effect sizes using GWAS summary statistics and an external linkage disequilibrium (LD) reference panel. PRS-CS utilizes a high-dimensional Bayesian regression framework, and is distinct from previous work by placing a continuous shrinkage (CS) prior on SNP effect sizes, which is robust to varying genetic architectures, provides substantial computational advantages, and enables multivariate modeling of local LD patterns. Simulation studies using data from the UK Biobank show that PRS-CS outperforms existing methods across a wide range of effect size distributions, especially when the training sample size is large. We apply PRS-CS to predict six complex diseases and six quantitative traits in the Partners HealthCare Biobank, and further demonstrate the improvement of PRS-CS in prediction accuracy over alternative methods.

Introduction

Polygenic risk scores (PRS), which summarize the effects of genome-wide genetic markers to measure the genetic liability to a trait or a disorder, have shown promise in predicting human complex traits and diseases, and may facilitate early detection, risk stratification, and prevention of common complex diseases in healthcare settings[1,2].

To maximize the translational potential of PRS, statistical and computational methods are needed that can (1) jointly model genetic markers across the genome to make full use of the available information while accounting for local linkage disequilibrium (LD) structures; (2) accommodate varying effect size distributions across complex traits and diseases, from highly polygenic genetic architectures (e.g., height and schizophrenia), to a mixture of small effect sizes and clusters of genetic loci that have moderate to larger magnitudes of effects (e.g., autoimmune diseases and Alzheimer's disease); (3) produce prediction from GWAS summary statistics without access to individual-level data; and (4) retain computational scalability.

To date, most applications calculate PRS from a subset of the genetic markers after pruning out SNPs in LD and applying a P-value threshold to GWAS summary statistics[3]. Although this approach has advantages in terms of computational and conceptual simplicity, and has been used to predict genetic liability across a broad phenotypic spectrum, recent studies have shown that this conventional method for PRS construction discards information and limits prediction accuracy[4]. More sophisticated Bayesian polygenic prediction methods that rely on GWAS summary statistics, including LDpred[4] and the normal-mixture model recently developed[5,6], can incorporate genome-wide markers and accommodate varying genetic architectures, and thus have enhanced performance and flexibility. However, the type of prior on SNP effect sizes used in these methods, known as discrete mixture priors, imposes daunting computational challenges and may result in inaccurate adjustment for local LD patterns.

In this work, we present a novel polygenic prediction method, PRS-CS, which utilizes a Bayesian regression framework and places a conceptually different class of priors — the continuous shrinkage (CS) priors — on SNP effect sizes. Continuous shrinkage priors allow for marker-specific adaptive shrinkage (that is, the amount of shrinkage applied to each genetic marker is adaptive to the strength of its association signal in GWAS), and thus can accommodate diverse underlying genetic architectures. In addition, continuous shrinkage priors enable conjugate block update of the SNP effect sizes in posterior inference (that is, effect sizes for SNPs in each LD block are updated jointly, in a multivariate fashion, in contrast to updating the effect size for each marker separately and sequentially), and thus can accurately model local LD patterns and provide sub-

stantial computational improvements. Several special cases of continuous shrinkage priors have been applied to quantitative trait prediction or gene mapping [7–12]. However, all previous work required individual-level data and was limited to small-scale analyses (both in term of the sample size and number of genetic markers). PRS-CS only requires GWAS summary statistics and an external LD reference panel, and therefore can be applied in a broader range of settings.

We conduct simulation studies using the UK Biobank genetic data [13,14], and demonstrate that PRS-CS dramatically improves the predictive performance of PRS over existing methods across a wide range of genetic architectures, especially when the training sample size is large. We apply PRS-CS to predict six curated common complex diseases (breast cancer, coronary artery disease, depression, inflammatory bowel disease, rheumatoid arthritis, and type 2 diabetes mellitus) and six quantitative traits (height, body mass index, high-density lipoproteins, low-density lipoproteins, cholesterol, and triglycerides) in the Partners HealthCare Biobank [15], and further demonstrate the potential of PRS-CS for the clinical translation of polygenic prediction.

Results

Conceptual frameworks. We consider a Bayesian high-dimensional regression framework for polygenic modeling and prediction:

$$\boldsymbol{y}_{N\times 1} = \boldsymbol{X}_{N\times M}\boldsymbol{\beta}_{M\times 1} + \boldsymbol{\epsilon}_{N\times 1}, \tag{1}$$

where N and M denote the sample size and number of genetic markers, respectively, $\boldsymbol{y}$ is a vector of traits, $\boldsymbol{X}$ is the genotype matrix, $\boldsymbol{\beta}$ is a vector of effect sizes for the genetic markers, and $\boldsymbol{\epsilon}$ is a vector of residuals. By assigning appropriate priors on the regression coefficients $\boldsymbol{\beta}$ to impose regularization, additive PRS can be calculated using posterior mean effect sizes.

Essentially all widely used prior densities for $\boldsymbol{\beta}$ can be represented as scale mixtures of normals:

$$p(\beta_j) = \int \mathrm{N}(0, \Psi_j)\mathrm{d}G(\Psi_j), \qquad j = 1, 2, \cdots, M, \tag{2}$$

or equivalently, as the following hierarchical form:

$$\beta_j \mid \Psi_j \sim \mathrm{N}(0, \Psi_j), \qquad \Psi_j \sim G, \qquad j = 1, 2, \cdots, M, \tag{3}$$

where $\mathrm{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2, and G is a mixing distribution. For example, if G places all its mass at a single point, i.e., $G(\Psi_j) = \delta_{\sigma_\beta^2}$, where $\delta_\bullet$ is the Dirac delta measure, then marginally $\beta_j \sim \mathrm{N}(0, \sigma_\beta^2)$, and we have recovered the infinitesimal model [16]. To create a more flexible model of

the genetic architecture, a discrete mixture of two or more point masses or densities can be used, which allows for a wider effect size distribution than a normal prior can produce. For example, $G(\Psi_j) = (1-\pi)\delta_0 + \pi\delta_{\tau^2}$, where π is the mixing probability (the fraction of causal variants), produces the point-normal prior on effect sizes, $\beta_j \sim (1-\pi)\delta_0 + \pi \mathrm{N}(0, \tau^2)$, which was used in LDpred[4]. Although discrete mixture priors offer a natural and intuitive approach to model non-infinitesimal genetic architectures, posterior inference requires a stochastic search over an exponentially large discrete model space, and does not allow for multivariate block update of effect sizes, which limits computational efficiency and may result in inaccurate modeling of local LD patterns.

In this work, we investigate a conceptually different class of priors — the continuous shrinkage priors. In particular, we consider the following prior on SNP effect sizes, which can be represented as global-local scale mixtures of normals:

$$\beta_j \mid \psi_j \sim \mathrm{N}(0, \phi\psi_j), \qquad \psi_j \sim g, \tag{4}$$

where ϕ is a global scaling parameter that shares across genetic markers and controls the degree of sparseness of the model, and g is an absolutely continuous density function, in contrast to a discrete mixture of atoms or densities. By appropriately choosing the continuous mixing density g, this modeling framework can produce a variety of shapes of the prior distribution on β_j. In particular, g can be designed to introduce a prior distribution on the SNP effect sizes that has a sizable amount of mass near zero to impose strong shrinkage on noise, while at the same time has heavy tails to avoid over-shrinkage of truly non-zero effects. The marker-specific local shrinkage parameter ψ_j can then adaptively squelch small noisy estimates towards zero, while leaving data-supported large signals unshrunk. In this work, we investigate a specific g (known as the Strawderman-Berger prior[17,18]; see Methods), and present two versions of the algorithm, which differ in the way to learn the global scaling parameter ϕ. In PRS-CS, we search a small number of fixed ϕ, select the ϕ value that produces the best predictive performance in a validation data set, and evaluate the algorithm in an independent testing set. In the second version of the algorithm, which we call PRS-CS-auto, we use a fully Bayesian approach and place a standard half-Cauchy prior on the global shrinkage parameter[19,20]: $\phi^{1/2} \sim \mathrm{C}^{+}(0, 1)$, such that ϕ is automatically learnt from data and no validation set is needed.

Individual-level Bayesian regression models (1) with a prior on SNP effect sizes can often be approximated using an external LD reference panel and turned into summary statistics based methods[4,6,21,22]. Here we enable posterior inference of SNP effect sizes from GWAS summary statistics under continuous shrinkage priors using an efficient Gibbs sampler with multivariate block update of the effect sizes (see Methods).

Overview of polygenic prediction methods. We compare PRS-CS and PRS-CS-auto with four polygenic prediction methods that rely on GWAS summary statistics in both simulations and real data analyses: polygenic scoring based on all genetic markers (unadjusted PRS), informed LD-pruning (also known as LD-clumping) and P-value thresholding (P+T), LDpred and LDpred-inf[4]. Throughout the paper, we use the 1000 Genomes Project (1KG) European sample ($N = 503$) as the external LD reference panel, but also assess the impact of using an in-sample LD reference panel on prediction accuracy in supplementary materials.

Simulations. We first compared the predictive performance of six polygenic prediction methods across different genetic architectures and training sample sizes (i.e., GWAS sample sizes) in simulation studies (Fig. 1 and Supplementary Table 1). SNP effect sizes were simulated using (1) a point-normal model with different numbers of causal variants, and (2) a normal mixture model, as described in the Methods section. Tuning parameters (P-value threshold in P+T, fraction of causal SNPs in LDpred, and global shrinkage parameter in PRS-CS) were selected in a validation data set ($N = 3,000$). Prediction accuracy for all methods was quantified by R^2 between the observed and predicted traits in an independent testing set ($N = 3,000$).

Figure 1 shows that polygenic prediction methods that do not account for non-infinitesimal genetic architectures (unadjusted PRS and LDpred-inf) performed poorly when the number of causal variants is small, but became more comparable to other methods when the genetic architectures are highly polygenic. For all the methods, the prediction accuracy decreased as the number of causal variants increases with fixed heritability, because as more causal SNPs are in LD (as a result of more causal SNPs being randomly sampled across the genome) and their effect sizes decline, it becomes increasingly difficult to distinguish real signals from noise. Overall, methods that account for local LD patterns (LDpred, PRS-CS and PRS-CS-auto) outperformed P+T, which discards LD information. However, one unexpected observation is that the prediction accuracy of LDpred decreased dramatically as the training sample size grows when the genetic architecture is sparse. This is likely because when the number of causal variants is small and the training sample size is large, all markers in LD with the causal variant become highly statistically significant in association tests, and LDpred does not accurately adjust for the LD structure, resulting in a decrease in predictive performance. In contrast, PRS-CS and PRS-CS-auto were minimally affected in the combination of sparse genetic architectures and large training sample sizes, which demonstrates the advantage of multivariate modeling and block update of the effect sizes for genetic markers in LD. In a few scenarios where the training sample size is small, PRS-CS produced lower prediction accuracy than LDpred, but it outperformed LDpred as the sample size grows across all genetic architectures. PRS-CS-auto did not perform well when the training sample size is small and the genetic

architecture is sparse (e.g., in the case of 100 causal variants and 10,000 training samples), but approached the performance of PRS-CS as the sample size increases.

In addition to prediction accuracy, we assessed the calibration of polygenic prediction methods by regressing the true phenotype onto the PRS predictor and inspecting the regression slope. A slope close to one indicates that a predictor is correctly calibrated. Consistent with predictive performance, as the training sample size grows, our Bayesian approach provides the best calibration among all methods examined (Supplementary Table 7). PRS-CS-auto is particularly well calibrated for large training sample sizes, because it automatically learns the sparseness of the genetic architecture from data and adjusts for the LD structure accordingly.

Secondary simulation studies using (1) the point-normal model with different total heritability (0.2 and 0.8); (2) a point-t model with different numbers of causal variants; and (3) a point-gamma model with different numbers of causal variants produced similar patterns of prediction accuracy (Supplementary Figs. 1-4; Supplementary Tables 2-5) and calibration properties (Supplementary Tables 8-11). Using the combined UK Biobank validation and testing data sets ($N = 6,000$) as an in-sample LD reference panel in the point-normal simulations produced, in general, slightly higher prediction accuracy for methods making use of LD information (Supplementary Fig. 5; Supplementary Tables 6 and 12), suggesting that using a larger reference panel that better aligns with the LD structure of the target sample may increase predictive performance. However, as the improvement was marginal, it appears that the performance of PRS-CS(-auto) is not particularly sensitive to the LD reference panel, and 1KG can serve as a valid reference despite its relatively small sample size.

Polygenic prediction in the Partners Biobank. We applied PRS-CS, PRS-CS-auto, and alternative methods to predict six curated common complex diseases (breast cancer, coronary artery disease, depression, inflammatory bowel disease, rheumatoid arthritis, and type 2 diabetes mellitus), and six quantitative traits (height, body mass index, high-density lipoproteins, low-density lipoproteins, cholesterol, and triglycerides) in the Partners HealthCare Biobank. Large-scale GWAS summary statistics for each disease and trait were downloaded from public domains (Table 1 and Supplementary Table 13). SNP heritability for each disease (both on the observed scale and the liability scale) and trait estimated using GWAS summary statistics and LD score regression[23] are presented in Supplementary Table 14.

Predictive performance measured by Nagelkerke's R^2 (for disease phenotypes) and R^2 (for quantitative traits) is summarized in Fig. 2. Additional prediction accuracy metrics, including area under the receiver operating characteristic (ROC) curve (known as AUC), area under the precision-call curve, and the odds ratio (OR) comparing top 10% of the participants having high polygenic risk with the remaining 90% of the sample,

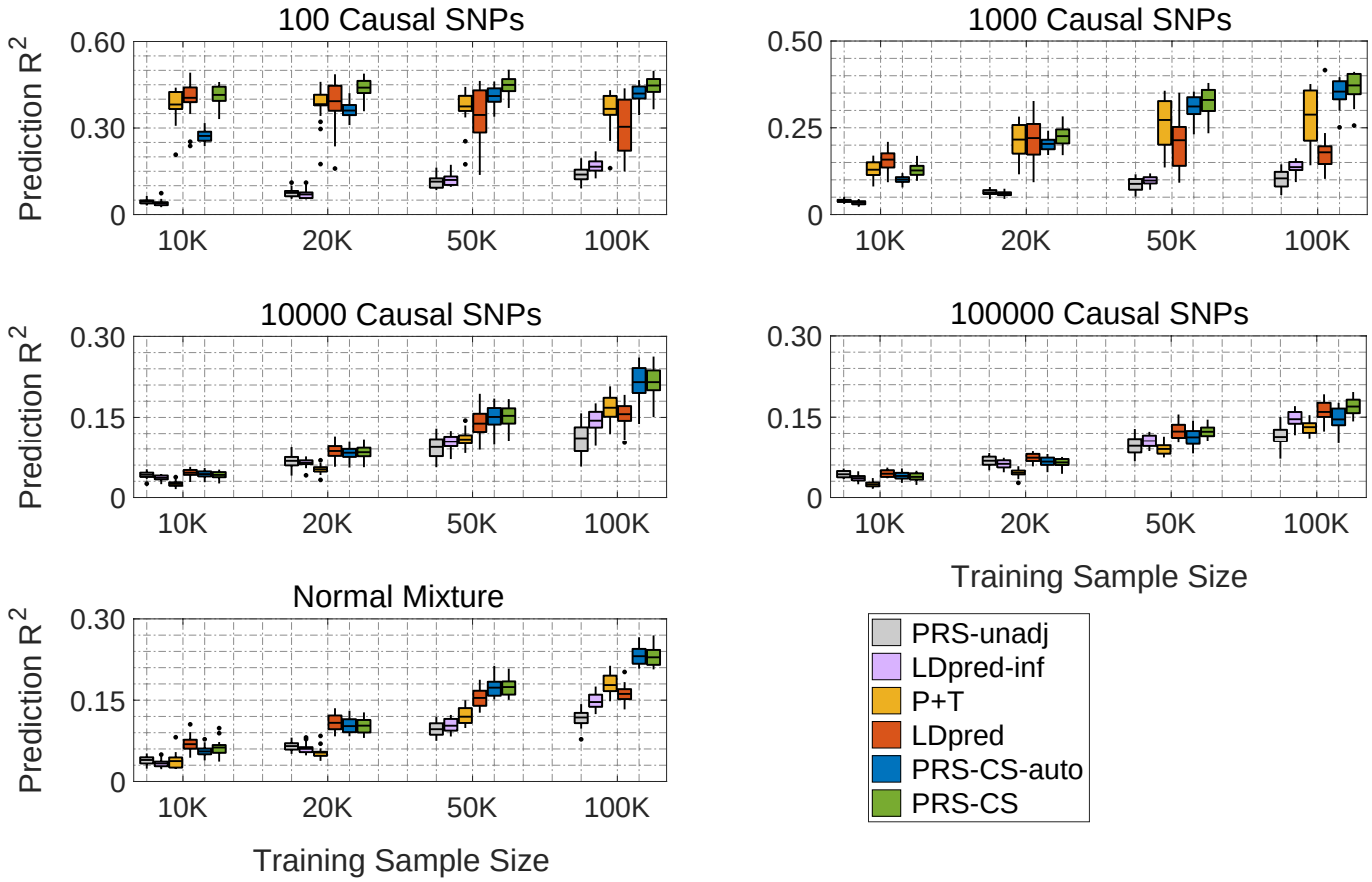

Figure 1: Predictive performance of six polygenic prediction methods in simulation studies using a point-normal model and a normal mixture model. Heritability was fixed at 0.5. The 1000 Genomes Project European sample was used as an external linkage disequilibrium (LD) reference panel. Tuning parameters (P-value threshold in P+T, fraction of causal SNPs in LDpred, and global shrinkage parameter in PRS-CS) were selected in a validation data set. Prediction accuracy was quantified by R^2 between the observed and predicted traits in an independent testing set. The upper four panels correspond to the four genetic architectures (100, 1,000, 10,000 and 100,000 causal variants) simulated using the point-normal model. The lower panel corresponds to the normal mixture model. Within each panel, results for four different training sample sizes (10,000, 20,000, 50,000 and 100,000) are shown. On each box, the central mark is the mean across 20 simulations, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points that are not considered outliers, and the outliers are plotted individually.

produced similar results in terms of the ranked performance of polygenic prediction methods and are reported in Supplementary Table 15.

Table 1: Information on six common complex diseases and six quantitative traits. The sample size for the external genome-wide association studies (GWAS), and the number of genetic markers included in the polygenic prediction are shown, along with the sample size for each disease and quantitative phenotype in the Partners HealthCare Biobank (PBK). For unadjusted PRS and P+T, all common genetic markers (minor allele frequency $\geqslant 1\%$) that passed quality control and are present in the summary statistics and 1000 Genomes Project (1KG) European sample were used in prediction. For LDpred(-inf) and PRS-CS(-auto), genetic markers were further restricted to the HapMap3 (HM3) panel.

Disease/Trait	Abbreviation	GWAS Reference	GWAS sample size (case/control)	1KG ∩ PBK SNPs	1KG ∩ PBK ∩ HM3 SNPs	PBK sample size (case/control)
Breast Cancer	BRCA	Michailidou et al. [24]	228,951 (122,977/105,974)	5,022,127	857,616	10,220 (884/9,336)
Coronary Artery Disease	CAD	Nikpay et al. [25]	184,305 (60,801/123,504)	4,803,592	849,399	16,251 (2,759/13,492)
Depression	DEP	Wray et al. [26]	173,005 (59,851/113,154)	4,924,025	850,291	15,276 (2,361/12,915)
Inflammatory Bowel Disease	IBD	Liu et al. [27]	34,652 (12,882/21,770)	4,823,570	849,749	18,998 (750/18,248)
Rheumatoid Arthritis	RA	Okada et al. [28]	58,284 (14,361/43,923)	3,872,637	849,680	18,170 (753/17,417)
Type 2 Diabetes Mellitus	T2DM	Scott et al. [29]	159,208 (26,676/132,532)	4,901,848	856,912	18,823 (1,978/16,845)
Height	HGT	Yengo et al. [30]	693,529	1,578,533	750,888	3,957
Body mass index	BMI	Yengo et al. [30]	681,275	1,579,905	751,676	3,954
High-density lipoproteins	HDL	Willer et al. [31]	188,578	1,604,577	758,036	2,491
Low-density lipoproteins	LDL	Willer et al. [31]	188,578	1,600,625	756,724	1,713
Cholesterol	CHOL	Willer et al. [31]	188,578	1,604,391	757,970	2,561
Triglycerides	TRIG	Willer et al. [31]	188,578	1,601,270	756,913	2,505

Consistent with previous work, unadjusted PRS performed poorly regardless of the genetic architecture, and LDpred showed an overall improvement over P+T. Among the six curated disease phenotypes, PRS-CS produced substantially better predictions for breast cancer (40.06% relative increase in Nagelkerke's R^2 compared to LDpred) and rheumatoid arthritis (27.76% relative increase in Nagelkerke's R^2 compared to LDpred). For coronary artery disease, depression and type 2 diabetes mellitus, LDpred and PRS-CS had similar predictive performance, and both performed dramatically better than P+T. PRS-CS was only inferior to LDpred in the prediction of inflammatory bowel disease (9.35% relative decrease in Nagelkerke's R^2). However, we note that inflammatory bowel disease has the smallest training sample size among all diseases and traits (Table 1). The lower prediction accuracy of PRS-CS for this disease is thus consistent with our simulation studies, where we observed that when the training sample size is limited, LDpred can outperform PRS-CS. PRS-CS-auto produced lower prediction accuracy than LDpred except for breast cancer, indicating that the current GWAS sample sizes for most diseases may not be large enough to accurately learn the global

shrinkage parameter from GWAS summary statistics.

For the six quantitative traits, both PRS-CS and PRS-CS-auto consistently outperformed all alternative methods examined. The relative improvement in prediction accuracy for PRS-CS compared to LDpred ranged from 7.39% for LDL and 9.49% for BMI, to 28.62% for height and 31.66% for cholesterol, with an average improvement of 18.34%. The average improvement of PRS-CS-auto relative to LDpred across the six quantitative traits was 11.22%. The average improvements of PRS-CS and PRS-CS-auto relative to P+T were 47.74% and 37.74%, respectively. We note that LDpred was the best method after PRS-CS and PRS-CS-auto for all quantitative traits except height, for which its prediction accuracy was lower than LDpred-inf and P+T. This is theoretically expected and consistent with a recent study, which also observed that for highly polygenic traits, LDpred-inf often outperforms LDpred[32].

Overall, using the Partners HealthCare Biobank data as an in-sample LD reference ($N = 19,136$) instead of the 1KG reference panel slightly increased the prediction accuracy but the improvement was marginal (Supplementary Fig. 6 and Supplementary Table 16).

Discussion

Polygenic prediction, which exploits genome-wide genetic markers to estimate the genetic liability to a complex human disease or trait, is likely to become useful in clinical care and contribute to personalized medicine. As a high-dimensional regression problem that requires regularization, a majority of the existing methods that jointly model genetic markers across the genome employ Bayesian approaches and assign a discrete mixture prior on SNP effect sizes. Although intuitively appealing, this class of priors generates daunting computational challenges: the model space grows exponentially with the number of markers, which is difficult to fully explore, and more importantly, discrete mixture priors do not allow for block update of effect sizes and thus hinder accurate LD adjustment in polygenic prediction. LDpred[4] partially addressed this issue by making several simplifying assumptions to the posterior distribution and using marginal posterior without LD to approximate the true posterior. However, our simulation studies suggest that this approximation may be inaccurate.

We have presented a conceptually different class of priors — the continuous shrinkage priors — which can be represented as global-local scale mixtures of normals, for polygenic modeling. By using a continuous mixing density on the scales of the marker effects, continuous shrinkage priors enable a simple and efficient

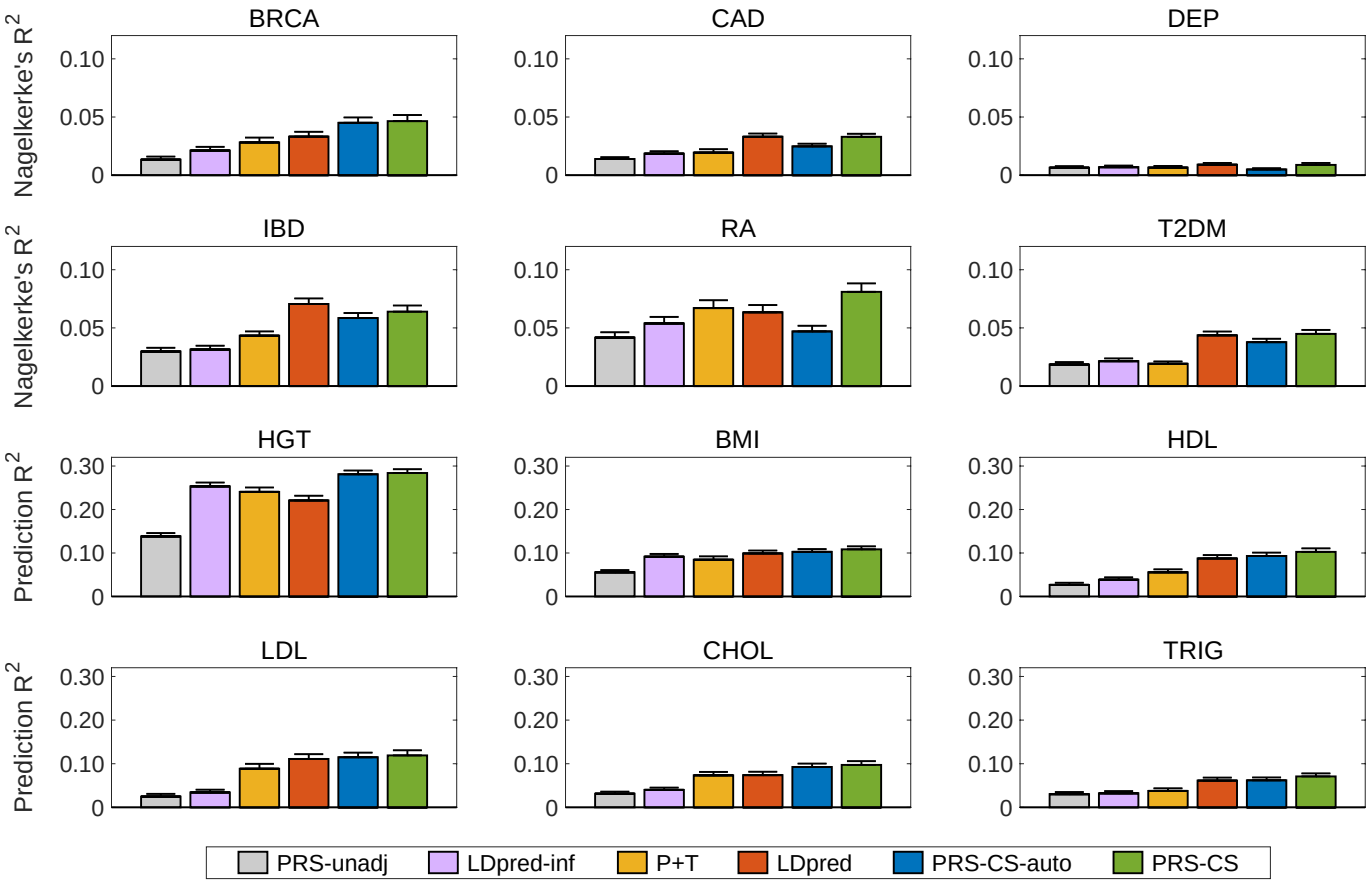

Figure 2: Prediction accuracy of six polygenic prediction methods in the Partners HealthCare Biobank. Posterior SNP effect sizes were trained with large-scale genome-wide association summary statistics, using the 1000 Genomes Project European sample as an external linkage disequilibrium (LD) reference panel. Polygenic scores were applied to predict six curated common complex diseases — breast cancer (BRCA), coronary artery disease (CAD), depression (DEP), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), and type 2 diabetes mellitus (T2DM), and six quantitative traits — height (HGT), body mass index (BMI), high-density lipoproteins (HDL), low-density lipoproteins (LDL), cholesterol (CHOL), and triglycerides (TRIG). The Partners HealthCare Biobank sample for each disease and quantitative phenotype was repeatedly and randomly split into a validation set comprising 1/3 of the data and a testing set comprising 2/3 of the data. Tuning parameters (P-value threshold in P+T, fraction of causal SNPs in LDpred, and global shrinkage parameter in PRS-CS) were selected in the validation data set, and the predictive performance was assessed in the testing set. For disease (case-control) phenotypes and quantitive traits, prediction accuracy was measured by the Nagelkerke's R^2 and R^2, respectively, averaged across 100 random splits. The error bar indicates the standard deviation of prediction accuracy across 100 random splits.

Gibbs sampler with multivariate block update of the effect sizes, and thus resolve a major technical hurdle of discrete mixture priors. A second feature of the continuous shrinkage prior is its ability to shrink adaptively. By constructing a prior density on SNP effect sizes that is both peaked at zero and heavy-tailed, the method imposes strong shrinkage on small effects that are likely to be noise, while applying practically no shrinkage to data-supported truly non-zero signals. Simulated and real data analyses showed that PRS-CS consistently outperforms existing methods across a wide range of genetic architectures, especially when the training sample size is large. We note that previous work often extrapolated prediction accuracy for larger effective sample sizes by restricting the analysis to a subset of the genetic markers[4,32]. However, our simulations suggest that this approach may not fully capture the behavior of a polygenic prediction algorithm when the training sample size grows, and underscore the need for actually scaling up the sample size in future studies.

PRS-CS has a tuning parameter, i.e., the global shrinkage parameter ϕ, which needs to be fixed based on prior beliefs about the sparseness of the genetic architecture, or selected by testing a small number of values. If a grid search is used, like other polygenic prediction methods that have tuning parameters such as P+T and LDpred, the optimal value of ϕ should be selected using a validation data set that is independent of the testing set where predictive performance is assessed to avoid overfitting. In this work, we also presented PRS-CS-auto, a fully Bayesian approach that enables automatic learning of ϕ from GWAS summary statistics. Although analyses in the Partners Biobank indicate that, for many disease phenotypes, the current GWAS sample sizes may not be large enough to accurately learn ϕ and the prediction accuracy of PRS-CS-auto may be lower than PRS-CS and LDpred, simulation studies and quantitative trait analyses suggest that PRS-CS-auto can be useful when the training sample size is large or when an independent validation set is difficult to acquire.

Although continuous shrinkage priors enable multivariate modeling of the LD structure, simultaneous updating of the effect sizes for genome-wide markers remains computationally infeasible. In this work, we used a genome partition computed and validated by prior work[33], which divides the genome into 1,703 largely independent genomic regions, and has been successfully used in local heritability and genetic correlation analyses[34,35]. Block update of posterior SNP effect sizes can thus be performed within each LD block, assuming no LD between blocks. Using a sliding window approach as implemented in LDpred[4] may more accurately capture LD across blocks, but is more memory intensive and computationally expensive. By restricting the analysis to HapMap3 variants, the partition we employed gives a moderate number of SNPs within each block (on average $\sim$500 SNPs per block), and the Bayesian computation with 1,000 MCMC iterations on the longest chromosome can be completed within an hour using one Intel(R) Xeon(R) CPU core and 2GB of memory.

Expanding the size of LD blocks may improve prediction accuracy but increases computational cost (as each MCMC iteration requires inverting an $L \times L$ matrix where L is the block size), while reducing the size of LD blocks has the potential risk of missing long-range LD. Therefore, the partition we chose represents a balance between modeling accuracy and computational burden. Including multi-million SNP predictors may increase prediction accuracy[36] but requires further work.

We note that the prior we investigated in this work, i.e., the Strawderman-Berger prior on the local marker-specific shrinkage parameter, is only one of the possible choices within the class of continuous shrinkage priors, which includes the normal-gamma prior[37,38], the normal-inverse-gaussian prior[37], the generalized t (generalized double Pareto) prior[39,40], and the normal-exponential-gamma prior[41,42], among others. In addition, most frequentist regularization procedures, such as LASSO, elastic net and bridge regression, have a Bayesian counterpart that can be represented as global-local scale mixtures priors in combination with posterior mode inferences. Each of these priors uses a different continuous mixing density to produce a different marginal prior on the SNP effect sizes. These alternatives may perform equally well or better than the Strawderman-Berger prior for certain genetic architectures. However, we found that as long as the prior on the effect sizes places a sizable amount of mass around zero and has heavier-than-exponential tails, variation in the shape of the prior does not seem to have a large impact on prediction accuracy. Therefore, we believe that the primary gain of PRS-CS over existing methods lies in its more accurate multivariate modeling of local LD patterns and its block-updated Gibbs sampling that can improve the mixing and convergence rate of the Markov chain. We thus recommend using the Strawderman-Berger prior as a default choice. A systematic investigation and comparison of different continuous shrinkage priors is a direction of future work.

We note several additional directions for further technical developments that may be useful. First, although this paper is focused on polygenic prediction methods that only require GWAS summary statistics, PRS-CS and PRS-CS-auto can be straightforwardly applied to individual-level data. Given that a majority of the existing Bayesian genomic prediction models, including Bayes alphabetic methods[10,43–48], BayesR[49,50], BVSR[51], BSLMM[52], and DPR[53], have used discrete mixture priors on SNP effect sizes, we expect that PRS-CS can provide substantial improvements in computational efficiency and prediction accuracy for genomic prediction that leverages individual-level data. Second, jointly modeling multiple genetically correlated traits and including functional annotations in polygenic modeling are expected to increase the predictive performance of PRS, as shown by recent studies[32,54,55]. Lastly, current research on polygenic prediction has largely been restricted to European samples. Heterogeneity between the GWAS, LD reference and testing samples may reduce prediction accuracy as recently demonstrated in genetic correlation analysis and fine-mapping[56,57].

Expanding genomic prediction methods to handle unknown ancestry of the target sample (e.g., applications in forensic science) and enable cross-ethnic risk prediction is critical to maximize the value of PRS in a diverse population.

Although PRS-CS provides a substantial improvement over existing methods for polygenic prediction, current prediction accuracy of PRS is still lower than what can be considered clinically useful, and much work is needed to further improve the predictive performance and translational value of PRS. In theory, the utility of PRS depends on multiple factors, including the GWAS sample size, and the heritability and genetic architecture of the disease. For example, among the six complex diseases we analyzed, depression had the lowest prediction accuracy (Nagelkerke's R^2 less than 1%), likely due to a combination of its relatively low heritability, extremely polygenic genetic architecture, and the heterogeneous nature of the disorder. A recent study projected that a GWAS with multi-million subjects is needed to identify genetic variants that explain 80% of the SNP heritability for major depressive disorder[5]. In contrast, it may be easier to produce a clinically useful prediction for some autoimmune diseases or late-onset chronic diseases (e.g., coronary artery disease and type 2 diabetes), due to the existence of SNPs with moderate to larger effect sizes. With these being said, as the GWAS sample size continues to grow, we believe that the predictive value of PRS will keep increasing, and PRS-CS(-auto) will demonstrate bigger advantages over existing methods with larger training sample sizes.

URLs.

PRS-CS: https://github.com/getian107/PRScs

Eagle2: https://data.broadinstitute.org/alkesgroup/Eagle

Genome partition: http://bitbucket.org/nygcresearch/ldetect-data

LDpred: https://github.com/bvilhjal/ldpred

Minimac3: https://genome.sph.umich.edu/wiki/Minimac3

Partners HealthCare Biobank: https://biobank.partners.org

PLINK 1.9: https://www.cog-genomics.org/plink/1.9

PRSice-2: https://choishingwan.github.io/PRSice

UK Biobank: http://www.ukbiobank.ac.uk

Methods

PRS-CS and PRS-CS-auto. We consider the following phenotype model:

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \qquad p(\sigma^2) \propto \sigma^{-2}, \tag{5}$$

where $\boldsymbol{y}$ is a vector of standardized phenotypes from N individuals, $\boldsymbol{Z}$ is an $N \times M$ matrix of standardized genotypes (each column is mean centered and has unit variance), $\boldsymbol{\beta}$ is a vector of effect sizes, $\boldsymbol{\epsilon}$ is a vector of independent environmental effects, and we have assigned a non-informative scale-invariant Jeffreys prior on the residual variance σ^2. In contrast to discrete mixture priors, we consider a conceptually different class of priors:

$$\beta_j \sim \mathrm{N}\left(0, \frac{\sigma^2}{N}\phi\psi_j\right), \qquad \psi_j \sim g, \tag{6}$$

where the variance of β_j scales with the residual variance and the sample size, ϕ is a global scaling parameter that is shared across all effect sizes, ψ_j is a local, marker-specific parameter, and g is an absolutely continuous mixing density function. This type of prior is known as global-local scale mixtures of normals.

We first note that, given variance parameters σ^2, ϕ and ψ_j, $j = 1, 2, \cdots, M$, and the marginal least squares effect size estimates of the regression coefficients $\widehat{\boldsymbol{\beta}} = \boldsymbol{Z}^\top\boldsymbol{y}/N$, the posterior mean of $\boldsymbol{\beta}$ is

$$\mathbf{E}[\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}] = (\boldsymbol{D} + \boldsymbol{T}^{-1})^{-1}\widehat{\boldsymbol{\beta}}, \tag{7}$$

where $\boldsymbol{T} = \mathrm{diag}\{\phi\psi_1, \phi\psi_2, \cdots, \phi\psi_M\}$ is a diagonal matrix, and $\boldsymbol{D} = \boldsymbol{Z}^\top\boldsymbol{Z}/N$ is the LD matrix. It can be seen that the posterior mean is a matrix shrinkage version of the least squares estimate. In the degenerative special case where $\psi_j \equiv 1$, the model becomes Ridge regression and all effect sizes are shrunk towards zero at the same constant rate controlled by the overall shrinkage parameter ϕ. The introduction of the local shrinkage parameter ψ_j thus allows heterogeneity in the scales of effect sizes.

To provide further intuitions, assuming that all genetic markers are unlinked (i.e., no LD), we have $\boldsymbol{D} = \boldsymbol{I}$ and thus

$$\mathbf{E}[\beta_j \mid \widehat{\beta}_j] = \frac{1}{1 + \phi^{-1}\psi_j^{-1}}\widehat{\beta}_j = \left(1 - \frac{1}{1 + \phi\psi_j}\right)\widehat{\beta}_j := (1 - \tau_j)\widehat{\beta}_j, \tag{8}$$

where $\tau_j = 1/(1 + \phi\psi_j)$ is the shrinkage factor for the j-th marker, which relies on both ϕ and ψ_j, and describes the amount of shrinkage from the marginal least squares solution towards zero; $\tau_j = 0$ indicates no shrinkage while $\tau_j = 1$ yields total shrinkage. Therefore, ϕ controls the overall sparsity level of the model and plays a similar role as the regularization parameter in penalized regression, while ψ_j adaptively modifies the amount of shrinkage for each marker. By assigning a prior on ψ_j, which can produce a marginal prior density

on β_j that has both a sharp peak at zero and heavy tails, the model can pull small effects towards zero, while asserting little influence on larger effects.

In this work, we investigate a specific continuous shrinkage prior. We assign an independent gamma-gamma prior on the local shrinkage parameter ψ_j:

$$\psi_j \sim \mathrm{G}(a, \delta_j), \qquad \delta_j \sim \mathrm{G}(b, 1), \tag{9}$$

where $\mathrm{G}(\alpha, \beta)$ denotes the gamma distribution with shape parameter α and scale parameter β. By using change of variables, it can be verified that placing a gamma-gamma prior on ψ_j is equivalent to placing a three-parameter beta (TPB) prior on the shrinkage factor τ_j[41]:

$$\tau_j \sim \mathrm{TPB}(a, b, \phi), \tag{10}$$

where the TPB distribution has the following density function:

$$f(x; a, b, \phi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \phi^b x^{b-1} (1-x)^{a-1} \{1 + (\phi - 1)x\}^{-(a+b)}, \tag{11}$$

with $0 < x < 1$, $a > 0$, $b > 0$ and $\phi > 0$. When $\phi = 1$, the TPB distribution becomes a standard Beta distribution. For a fixed value of ϕ, a controls the behavior of the TPB prior near one, and thus the behavior of the prior on β_j around zero; b controls the behavior of the TPB prior near zero, and thus affects the tails of the prior on β_j. Figure 3 shows the prior densities on τ_j (upper panel) and β_j (middle and lower panels) with $\phi = 1$, $b = 1/2$, and three different values of a: $a = 1/2$, $a = 1$ and $a = 3/2$. It can be seen that when $a = 1/2$ and $b = 1/2$, the TPB prior has substantial mass near zero and one (Figure 3, upper panel), and thus the corresponding prior density on β_j has a very sharp peak around the origin, with zero being a pole (singular point; Figure 3, middle panel), along with heavy, Cauchy-like tails (Figure 3, lower panel). This prior is known as the horseshoe prior[58], due to the horseshoe-shaped prior density on the shrinkage factor τ_j. As a increases, the prior on β_j becomes less peaked at zero but the tails remain heavy. Finally, for fixed a and b, decreasing the global shrinkage parameter ϕ shifts the TPB prior from left to right, which imposes stronger shrinkage on the regression coefficients β_j.

For all continuous shrinkage priors that take the general form in Eq. (6), Gibbs samplers with block update of the regression coefficients $\boldsymbol{\beta}$ (i.e., SNP effect sizes) can be easily derived. By using LD information from an external reference panel, the method can be applied to GWAS summary statistics and does not require individual-level data. We describe the Gibbs sampler in Supplementary Note. In this study, we focus on a specific set of parameter values of the gamma-gamma prior on ψ_j (or equivalently, the TPB prior on τ_j): $a = 1$

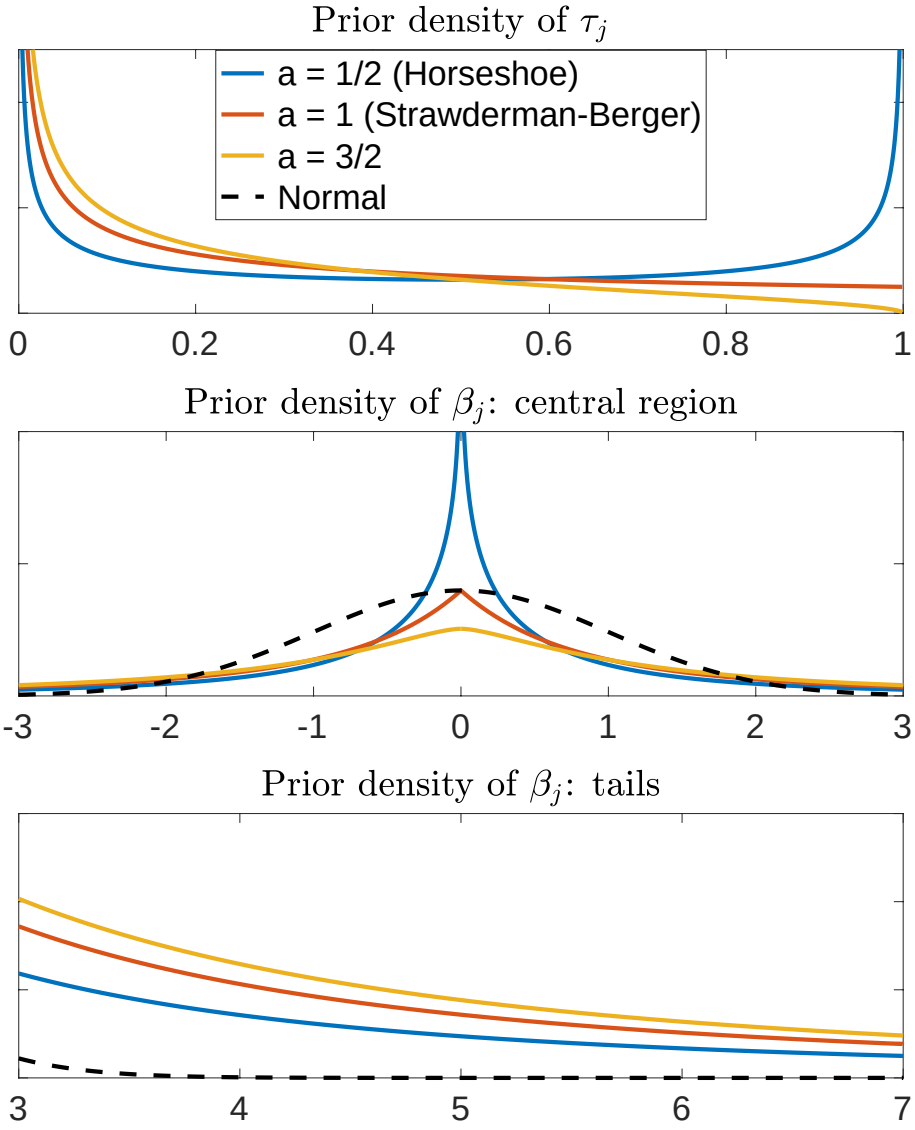

Figure 3: Densities of the priors. Upper panel: Density of the three-parameter beta prior on the shrinkage factor τ_j with $\phi = 1$, $b = 1/2$, and three different a values. Middle panel: Central region of the marginal prior density on the effect size β_j with $\phi = 1$, $b = 1/2$, and three different a values, in comparison with the standard normal density. Lower panel: Tails of the marginal prior density on the effect size β_j with $\phi = 1$, $b = 1/2$, and three different a values, in comparison with the standard normal density.

and $b = 1/2$. This particular specification is known as the Strawderman-Berger prior[17,18] or the quasi-Cauchy prior[59], and appears to work well across a range of simulated and real genetic architectures.

In practice, we partition the genome into 1,703 largely independent genomic regions estimated using data from the 1KG European sample[33–35], and conduct multivariate update of the effect sizes within each LD block (see Supplementary Note). To avoid numerical issues caused by collinearity between SNPs, we set a lower bound on the amount of regularization applied to the genetic markers (i.e., restricting $\phi^{-1}\psi_j^{-1} \geqslant \rho$, where ρ is a small constant). We use $\rho = 1$ throughout this paper.

We find that the predictive performance of the model is not sensitive to the global shrinkage parameter ϕ, and setting $\phi^{1/2}$ roughly to the proportion of causal variants[60] works well. If a prior guess of the sparseness of the genetic architecture is not available, we provide two ways to learn ϕ. In PRS-CS, we search a small number of ϕ values: $\phi^{1/2} \in \{0.0001, 0.001, 0.01, 0.1, 1\}$, and select the ϕ that produces the best predictive performance in a validation data set, which is independent of the testing set where prediction accuracy of the algorithm is evaluated. In PRS-CS-auto, we use a fully Bayesian approach and assign a standard half-Cauchy prior on $\phi^{1/2}$ [19,20], such that ϕ is automatically learnt from GWAS summary statistics and no validation data set is needed. See Supplementary Note for the Gibbs updates of ϕ.

For both PRS-CS and PRS-CS-auto, the Gibbs sampler usually attains reasonable convergence after 1,000 Markov Chain Monte Carlo (MCMC) iterations and produces prediction accuracy close to what can be achieved by much longer MCMC runs. We thus use 1,000 MCMC iterations with the first 500 steps as burn-in in simulation studies to reduce computational cost. In practice, we recommend using longer MCMC runs when time and computational resources permit. In the Partners Biobank, we report the predictive performance of PRS-CS and PRS-CS-auto based on 10,000 MCMC iterations in total and 5,000 burn-in steps.

Unadjusted PRS. The unadjusted PRS is the sum of all genetic markers across the genome, weighted by their marginal effect size estimates. More specifically, the unadjusted polygenic score for the i-th individual is $\mathrm{PRS}_i = \sum_{j=1}^{M} X_{ij}\widehat{b}_j$, where M is the total number of genetic markers, X_{ij} is the genotype for the i-th individual and the j-th SNP, and $\widehat{b}_j$ is the estimated marginal per-allele effect size of the j-th SNP.

P+T. The P+T method refers to the calculation of PRS using informed LD-pruning (also known as LD-clumping) and P-value thresholding. In this study, we use the implementation of the P+T method in the software package PRSice-2[61] and its default parameter settings. Specifically, for any pair of SNPs that have a physical distance smaller than 250 kb and an R^2 greater than 0.1, the less significant SNP is removed. The polygenic score is then calculated as the sum of the remaining, largely independent SNPs with a GWAS

association P-value below a threshold P_T, weighted by their marginal effect size estimates. We consider $P_T \in \{$1E-8, 1E-7, 1E-6, 1E-5, 3E-5, 1E-4, 3E-4, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1$\}$ in this paper. The P_T value that produces the highest prediction accuracy in a validation data set is selected, and the predictive performance is assessed in an independent testing set.

LDpred and LDpred-inf. LDpred is a method that infers the posterior mean effect size of each genetic marker from GWAS summary statistics while accounting for LD, using a point-normal prior on the SNP effect sizes and LD information from an external reference panel[4]. Consider the linear model $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where both the phenotype $\boldsymbol{y}$ and the genotype matrix $\boldsymbol{Z}$ have been standardized. LDpred places an independent point-normal prior on each regression coefficient β_j:

$$\beta_j \sim \begin{cases} \mathrm{N}\left(0, \dfrac{h_g^2}{\pi M}\right), & \text{with probability } \pi, \\ 0, & \text{with probability } 1-\pi, \end{cases} \tag{12}$$

where h_g^2 is the heritability explained by genome-wide genetic markers (known as SNP heritability), and π is the fraction of causal variants. Given π and an estimate of h_g^2, which can be obtained, for example, by applying LD score regression[23] to the GWAS summary statistics, LDpred employs an MCMC sampler to approximate the posterior mean of β_j, conditioning on marginal least squares effect size estimates and LD information from a reference panel. In this paper, we consider $\pi \in \{$1E-5, 3E-5, 1E-4, 3E-4, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1$\}$. The π value with the highest prediction accuracy in a validation data set is selected, and the predictive performance is assessed in an independent testing set.

LDpred-inf is a special case of LDpred when all variants are assumed to be causal (i.e., $\pi = 1$). Under this infinitesimal model, the posterior mean effect sizes in the ℓ-th LD window have a closed-form approximation:

$$\mathbf{E}[\boldsymbol{\beta}_\ell \mid \widehat{\boldsymbol{\beta}}_\ell, \boldsymbol{D}_\ell] \approx \left(\boldsymbol{D}_\ell + \frac{M}{N h_g^2}\boldsymbol{I}\right)^{-1} \widehat{\boldsymbol{\beta}}_\ell, \tag{13}$$

where $\widehat{\boldsymbol{\beta}}_\ell$ is a vector of marginal least squares effect size estimates, $\boldsymbol{D}_\ell$ is the LD matrix that can be estimated from an external reference panel, $\boldsymbol{I}$ is an identity matrix, and it has been assumed that h_ℓ^2, the heritability explained by SNPs in the ℓ-th LD window, is small such that $1 - h_\ell^2 \approx 1$. In this work, we use an LD radius of $M/3000$ to approximate local LD patterns, as suggested in Vilhjálmsson et al.[4].

UK Biobank genetic data. UK Biobank is a prospective cohort study of ~500,000 individuals recruited across Great Britain during 2006-2010[13]. The protocol and consent were approved by the UK Biobank's Research Ethics Committee. Data for the current analyses were obtained under an approved data request.

The genetic data for the UK Biobank comprises 488,377 samples and was phased and imputed to ∼96 million variants with the Haplotype Reference Consortium (HRC) haplotype resource and the UK10K+1KG reference panel. We leveraged the QC metrics provided by the UK Biobank[14] and removed samples that had mismatch between genetically inferred sex and self-reported sex, high genotype missingness or extreme heterozygosity, sex chromosome aneuploidy, and samples that were excluded from kinship inference and autosomal phasing. We further restricted the analysis to unrelated white British participants. We conducted simulation studies using 819,941 HapMap3 SNPs after removing ambiguous (A/T and C/G) SNPs and markers with minor allele frequency (MAF) $< 1\%$, missing rate $> 1\%$, imputation quality INFO score < 0.8, and significant deviation from Hardy-Weinberg equilibrium (HWE) with $P < 1 \times 10^{-10}$. All genetic analyses in the UK Biobank were conducted using PLINK 1.9[62].

Simulations. We performed simulation studies using real genetic data from the UK Biobank and the 1KG European sample as an external LD reference panel. SNP effect sizes were simulated using (1) a point-normal model as specified in Eq. (12) with different numbers of causal variants (100, 1,000, 10,000 and 100,000), which represent extremely sparse to highly polygenic genetic architectures; and (2) a normal mixture model comprised 10 group-one SNPs, 1,000 group-two SNPs and 10,000 group-three SNPs, and the three effect size groups explained 10%, 20% and 70% of the total heritability, respectively. The simulated trait was generated by the sum of all genetic markers, weighted by their simulated effect sizes, and adding a normally distributed noise term which fixed the heritability at 0.5. We then conducted GWAS to produce a marginal least squares effect size estimate for each SNP, and applied each polygenic prediction method to the GWAS summary statistics. For P+T, LDpred and PRS-CS, tuning parameters were selected in a validation data set of 3,000 individuals that are unrelated to the training sample. The predictive performance of all the six methods was evaluated in 3,000 individuals (the testing set) that are unrelated to both the training sample and the validation set. R^2 between the observed and predicted traits was used to quantify the prediction accuracy. We regressed the true phenotype onto the PRS predictor, and used the regression slope as a measure of calibration. A slope close to one indicates that a predictor is well calibrated. For each combination of the genetic architecture and the training sample size (10,000, 20,000, 50,000 and 100,000), the simulation was repeated 20 times.

In order to systematically compare polygenic prediction methods across a wide range of settings, we conducted a number of secondary simulation studies: (1) sampling SNP effect sizes using a point-normal model with heritability fixed at 0.2 or 0.8; (2) sampling SNP effect sizes using a point-t model with heavy tails (a mixture of a point mass at zero and a Student's t-distribution with 4 degrees of freedom); (3) sampling

SNP effect sizes using a point-gamma model (a mixture of a point mass at zero and a gamma distribution with the shape parameter set to 2), which produces an effect size distribution that is asymmetric about zero and positively skewed with the right tail being long and thin and the left tail being short and fat; (4) using the combined UK Biobank validation and testing data sets ($N = 6,000$) as an in-sample LD reference panel in the point-normal simulations. For each setting and training sample size considered (10,000, 20,000, 50,000 and 100,000), and the simulation was repeated 20 times.

Partners HealthCare Biobank genetic data. The Partners Biobank is a collection of plasma, serum, DNA and buffy coats samples collected from consented subjects, which are linked to their electronic health records (EHR) and survey data on lifestyle, environment, and family history[63]. To date, Partners Biobank has enrolled more than 96,000 participants, and released genome-wide genetic data for 25,482 subjects.

We performed QC on each genotyping batch separately with the following steps: (1) SNPs with genotype missing rate > 0.05 were removed; (2) samples with genotype missing rate > 0.02 or absolute value of heterozygosity > 0.2, or samples that failed sex checks were excluded; (3) SNPs with missing rate > 0.02, or HWE test $P < 1 \times 10^{-6}$ were discarded. We then removed SNPs that showed significant batch associations with $P < 1 \times 10^{-6}$, and merged genotyping batches for subsequent processing and analyses.

The Partners HealthCare Biobank included individuals from diverse populations. We used the 1KG samples as a population reference panel to infer the ancestry of Partners Biobank participants. Specifically, we computed principal components (PCs) of the genotype data in all the 1KG samples, and trained a random forest model using the top 4 PCs on the super population labels (African [AFR], American [AMR], East Asian [EAS], European [EUR] and South Asian [SAS]), in which EUR ($N = 503$) included TSI, IBS, GBR, CEU, and FIN subpopulations. The random forest model was then applied to the Partners Biobank participants, and identified 19,136 unrelated subjects ($\widehat{\pi} < 0.2$) with European ancestry.

We used the Eagle2 software[64] for pre-phasing and Minimac3[65] for imputation in the Partners Biobank European sample. Lastly, we removed markers with MAF $< 1\%$, missing rate $> 2\%$, imputation quality INFO score < 0.8, and significant deviation from HWE with $P < 1 \times 10^{-10}$. All genetic analyses in the Partners Biobank were conducted using PLINK 1.9[62].

Partners Biobank curated disease populations and quantitative traits. For a number of common complex diseases, the Partners Biobank trained and validated a classification algorithm, which leverages both structured and unstructured EHR data, and combines natural language processing and statistical methods, in a gold standard training set created by expert chart review. The algorithm was then applied to all the participants in

the Biobank to identify cases and controls, and create curated disease populations. We selected six curated diseases — breast cancer (BRCA), coronary artery disease (CAD), depression (DEP), inflammatory bowel disease (IBD) (Crohn's disease or ulcerative colitis), rheumatoid arthritis (RA), and type 2 diabetes mellitus (T2DM) — for which there are more than 500 cases in the Biobank that have been genotyped, and external large-scale GWAS summary statistics are publicly available. For all the diseases, cases have an algorithm-based positive predictive value (PPV) of having current or past history of the disease greater than 0.90, and controls have a negative predictive value (NPV) of having no history of the disease greater than 0.99.

In addition, we selected six quantitative traits — height (HGT), body mass index (BMI), high-density lipoproteins (HDL), low-density lipoproteins (LDL), cholesterol (CHOL), and triglycerides (TRIG) — that have been measured in the Partners Biobank healthy control population with a Charlson age-comorbidity index 0-2 and the predicted 10-year survival probability greater than 90%. We predicted these quantitative traits in a relatively heathy population to avoid measurements affected by severe diseases or medications. For participants that have multiple measurements of a trait of interest, we used the median value. Table 1 presents the sample size for each curated disease and quantitative trait in the Partners Biobank.

Summary statistics and polygenic prediction. GWAS summary statistics for all the diseases and quantitative traits are publicly available (Supplementary Table 13). We removed ambiguous (A/T and C/G) SNPs and mapped the genetic markers to the Genome Reference Consortium human genome build 37. SNP heritability for each disease and trait was estimated using GWAS summary statistics and LD score regression[23]. Heritability estimates for diseases on the observed scale were transformed to the liability scale as described in Lee et al.[66], using the assumed population and sample prevalences shown in Supplementary Table 14. For unadjusted PRS and P+T, we used all the genetic markers that are present in the summary statistics, LD reference panel and the Partners Biobank genetic data. For LDpred(-inf) and PRS-CS(-auto), we further restricted the genetic markers to the HapMap3 panel to reduce memory and computational cost. Table 1 shows the total number of markers included in the analysis for each disease and quantitative phenotype.

For each curated disease and quantitative trait, the Partners HealthCare Biobank sample was repeatedly and randomly split into a validation set comprising 1/3 of the data and a testing set comprising 2/3 of the data. Tuning parameters (P-value threshold in P+T, fraction of causal SNPs in LDpred, and global shrinkage parameter in PRS-CS) were selected in the validation set, and the predictive performance was evaluated in the testing set. We use the average R^2 between the observed and predicted phenotypes across 100 random splits to assess the predictive performance for the quantitative traits, and report the average Nagelkerke's

R^2 metric across 100 random splits for disease (case-control) phenotypes. Nagelkerke's R^2 is defined as $R^2_{\text{nag}} = R^2/R^2_{\text{max}}$, where $R^2 = 1 - (\mathcal{L}_{\text{res}}/\mathcal{L}_{\text{full}})^{2/N}$, $R^2_{\text{max}} = 1 - \mathcal{L}_{\text{res}}^{2/N}$, $\mathcal{L}_{\text{res}}$ is the likelihood of a restricted logistic regression model with covariates only (an intercept, current age, sex and top 10 PCs of the genotype data), $\mathcal{L}_{\text{full}}$ is the likelihood of the full logistic regression model (covariates and the PRS predictor), and N is the sample size. We define the relative increase/decrease in R^2 of a polygenic prediction method A compared to method B as $(R^2_{\text{A}} - R^2_{\text{B}})/R^2_{\text{B}}$. In addition to R^2 or Nagelkerke's R^2, we also report area under the receiver operating characteristic (ROC) curve (known as AUC), area under the precision-call curve, and the odds ratio (OR) comparing top 10% of the participants having high polygenic risk with the remaining 90% of the sample. We adjusted for current age, sex and top 10 PCs of the genotype data in the calculation of all predictive performance metrics.

Acknowledgements

This work involved the use of the Enterprise Research Infrastructure & Services (ERIS) at Partners HealthCare. We thank the Partners HealthCare Biobank for providing genomic and health information data. This research was funded in part by National Institutes of Health (NIH) U01HG008685 supporting the eMERGE Network, and K99AG054573 (TG). JWS is a Tepper Family MGH Research Scholar and was supported in part by a gift from the Demarest Lloyd, Jr. Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This research has been conducted using the UK Biobank resource under an approved data request (ref: 32568).

The breast cancer genome-wide association analyses were supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the 'Ministère de l'Économie, de la Science et de l'Innovation du Québec' through Genome Qubec and grant PSR-SIIRI-701, The National Institutes of Health (U19CA148065, X01HG007492), Cancer Research UK (C1287/A10118, C1287/A16563, C1287/A10710) and The European Union (HEALTH-F2-2009-223175 and H2020 633784 and 634935). All studies and funders are listed in Michailidou et al. (2017).

Data on coronary artery disease have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from http://www.cardiogramplusc4d.org.

Author Contributions

T.G. conceived the study. T.G. and C.-Y.C. designed the experiments. T.G. developed the statistical methods with contributions from Y.N. C.-Y.C. preprocessed the Partners HealthCare Biobank genetic data. T.G. performed the simulations and real data analyses, with contributions from C.-Y.C. and Y.-C.A.F. T.G. developed the software, with input from C.-Y.C. and Y.-C.A.F. T.G. wrote the paper. C.-Y.C., Y.N., Y.-C.A.F., and J.W.S. provided critical revision for the manuscript. All authors reviewed and approved the final manuscript.

Competing Interests

The authors declare no competing financial interests.

References

1. N. Chatterjee, J. Shi, and M. García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17(7):392–406, 2016.

2. A.V. Khera, M. Chaffin, K.G. Aragam, M.E. Haas, C. Roselli, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50:1219–1224, 2018.

3. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

4. B.J. Vilhjálmsson, J. Yang, H.K. Finucane, A. Gusev, S. Lindström, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97(4): 576–592, 2015.

5. Y. Zhang, G. Qi, J.H. Park, and N. Chatterjee. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature Genetics*, 50:1318–1326, 2018.

6. L.R. Lloyd-Jones, J. Zeng, J. Sidorenko, L. Yengo, G. Moser, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *bioRxiv*, page 522961, 2019.

7. C.J. Hoggart, J.C. Whittaker, M. De Iorio, and D.J. Balding. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4(7):e1000130, 2008.

8. G. De Los Campos, H. Naya, D. Gianola, J. Crossa, A. Legarra, et al. Predicting quantitative traits with regression models for dense molecular markers and pedigrees. *Genetics*, 182(1):375–385, 2009.

9. R. Makowsky, N.M. Pajewski, Y.C. Klimentidis, A.I. Vazquez, C.W. Duarte, et al. Beyond missing heritability: prediction of complex traits. *PLoS Genetics*, 7(4):e1002051, 2011.

10. T.H.E. Meuwissen, B.J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

11. S. Xu. Estimating polygenic effects using markers of the entire genome. *Genetics*, 163(2):789–801, 2003.

12. N. Yi and S. Xu. Bayesian LASSO for QTL mapping. *Genetics*, 179(2):1045–1055, 2008.

13. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12 (3):e1001779, 2015.

14. C. Bycroft, C. Freeman, D. Petkova, G. Band, L.T. Elliott, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

15. V.S. Gainer, A. Cagan, V.M. Castro, S. Duey, B. Ghosh, et al. The Biobank Portal for Partners personalized medicine: a query tool for working with consented biobank samples, genotypes, and phenotypes using i2b2. *Journal of Personalized Medicine*, 6(1):11, 2016.

16. J. Yang, B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.

17. W.E. Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42(1):385–388, 1971.

18. J. Berger. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, 8:716–761, 1980.

19. A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1 (3):515–534, 2006.

20. N.G. Polson and J.G. Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538, 2010.

21. J. Yang, T. Ferreira, A.P. Morris, S.E. Medland, P.A.F. Madden, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, 44(4):369–375, 2012.

22. B. Pasaniuc and A.L. Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 18(2):117–127, 2017.

23. B.K. Bulik-Sullivan, P.R. Loh, H.K. Finucane, S. Ripke, J. Yang, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3): 291–295, 2015.

24. K. Michailidou, S. Lindström, J. Dennis, J. Beesley, S. Hui, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94, 2017.

25. M. Nikpay, A. Goel, H.H. Won, L.M. Hall, C. Willenborg, et al. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47(10): 1121–1130, 2015.

26. N.R. Wray, S. Ripke, M. Mattheisen, M. Trzaskowski, E.M. Byrne, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, 50(5):668–681, 2018.

27. J.Z. Liu, S. van Sommeren, H. Huang, S.C. Ng, R. Alberts, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, 47(9):979–986, 2015.

28. Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

29. R.A. Scott, L.J. Scott, R. Mägi, L. Marullo, K.J. Gaulton, et al. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes*, db161253, 2017.

30. L. Yengo, J. Sidorenko, K.E. Kemper, Z. Zheng, A.R. Wood, et al. Meta-analysis of genome-wide association studies for height and body mass index in ∼700,000 individuals of european ancestry. *Human Molecular Genetics*, 27(20):3641–3649, 2018.

31. C.J. Willer, E.M. Schmidt, S. Sengupta, G.M. Peloso, S. Gustafsson, et al. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45(11):1274–1283, 2013.

32. C. Marquez-Luna, S. Gazal, P.R. Loh, N. Furlotte, A. Auton, et al. Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *bioRxiv*, 375337, 2018.

33. T. Berisa and J.K. Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283–285, 2016.

34. H. Shi, G. Kichaev, and B. Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99(1):139–153, 2016.

35. H. Shi, N. Mancuso, S. Spendlove, and B. Pasaniuc. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *The American Journal of Human Genetics*, 101(5): 737–751, 2017.

36. S.H. Lee, S. Clark, and J.H.J. van der Werf. Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. *PLoS ONE*, 12(12):e0189775, 2017.

37. F. Caron and A. Doucet. Sparse bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine learning*, pages 88–95, 2008.

38. J.E. Griffin and P.J. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

39. A. Lee, F. Caron, A. Doucet, and C. Holmes. Bayesian sparsity-path-analysis of genetic association signal using generalized t priors. *Statistical Applications in Genetics and Molecular Biology*, 11(2), 2012.

40. A. Armagan, D.B. Dunson, and J. Lee. Generalized double pareto shrinkage. *Statistica Sinica*, 23(1): 119–143, 2013.

41. A. Armagan, M. Clyde, and D.B. Dunson. Generalized beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems*, volume 24, pages 523–531, 2011.

42. J.E. Griffin and P.J. Brown. Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, 53(4):423–442, 2011.

43. N. Yi, V. George, and D.B. Allison. Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics*, 164(3):1129–1138, 2003.

44. T.H.E. Meuwissen and M.E. Goddard. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution*, 36(3):261–279, 2004.

45. K.L. Verbyla, B.J. Hayes, P.J. Bowman, and M.E. Goddard. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genetics Research*, 91 (5):307–311, 2009.

46. B.J. Hayes, J. Pryce, A.J. Chamberlain, P.J. Bowman, and M.E. Goddard. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics*, 6(9):e1001139, 2010.

47. K.L. Verbyla, P.J. Bowman, B.J. Hayes, and M.E. Goddard. Sensitivity of genomic selection to using different prior distributions. *BMC Proceedings*, 4(1):S5, 2010.

48. R.L. Habier, D.and Fernando, K. Kizilkaya, and D.J. Garrick. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12(1):186, 2011.

49. M. Erbe, B.J. Hayes, L.K. Matukumalli, S. Goswami, P.J. Bowman, and other. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, 95(7):4114–4129, 2012.

50. G. Moser, S.H. Lee, B.J. Hayes, M.E. Goddard, N.R. Wray, et al. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genetics*, 11(4): e1004969, 2015.

51. Y. Guan and M. Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815, 2011.

52. X. Zhou, P. Carbonetto, and M. Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*, 9(2):e1003264, 2013.

53. P. Zeng and X. Zhou. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nature Communications*, 8(1):456, 2017.

54. J. Shi, J.H. Park, J. Duan, S.T. Berndt, W. Moy, et al. Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS Genetics*, 12(12):e1006493, 2016.

55. P. Turley, R.K. Walters, O. Maghzian, A. Okbay, J.J. Lee, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, 50(2):229–237, 2018.

56. C. Benner, A.S. Havulinna, M.R. Järvelin, V. Salomaa, S. Ripatti, et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *The American Journal of Human Genetics*, 101(4):539–551, 2017.

57. G. Ni, G. Moser, S. Ripke, B.M. Neale, A. Corvin, et al. Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *The American Journal of Human Genetics*, 102(6):1185–1194, 2018.

58. C.M. Carvalho, N.G. Polson, and J.G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

59. I.M. Johnstone and B.W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

60. J. Piironen and A. Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *arXiv*, 1610.05559, 2016.

61. J. Euesden, C.M. Lewis, and P.F. O'reilly. PRSice: polygenic risk score software. *Bioinformatics*, 31 (9):1466–1468, 2014.

62. C.C. Chang, C.C. Chow, L.C.A.M. Tellier, S. Vattikuti, S.M. Purcell, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):7, 2015.

63. E.W. Karlson, N.T. Boutin, A.G. Hoffnagle, and N.L. Allen. Building the partners healthcare biobank at partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations. *Journal of Personalized Medicine*, 6(1):2, 2016.

64. P.R. Loh, P. Danecek, P.F. Palamara, C. Fuchsberger, Y.A. Reshef, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448, 2016.

65. S. Das, L. Forer, S. Schönherr, C. Sidore, A.E. Locke, et al. Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10):1284–1287, 2016.

66. S.H. Lee, N.R. Wray, M.E. Goddard, and P.M. Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.