Polygenic Prediction via Bayesian Regression and Continuous Shrinkage Priors

Tian Ge[a,b,c], Chia-Yen Chen[a,b,c,d], Yang Ni[e], Yen-Chen Anne Feng[a,b,c,d], Jordan W. Smoller[a,b,c]

[a]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; [b]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA; [c]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02138, USA; [d]Analytic and Translational Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; [e]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. To date, most applications calculate PRS using a subset of largely independent genetic markers, but this approach discards information and limits the predictive value of PRS. More sophisticated Bayesian genomic prediction methods that jointly model genetic markers across the genome are computationally challenging and do not accurately account for linkage disequilibrium (LD) structure. Here, we present PRS-CS, a novel polygenic prediction method that infers posterior SNP effect sizes using GWAS summary statistics and an external 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 common, complex diseases and six quantitative traits in the Partners HealthCare Biobank, for which external large-scale GWAS summary statistics are publicly available, 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 quantitative traits and identifying individuals at high risk for complex diseases, and may facilitate early detection, stratification, and prevention of heritable, common diseases in healthcare settings [Chatterjee et al., 2016; Khera et al., 2018].

To maximize the translational potential of PRS, statistical and computational methods are needed that can (1) jointly model all 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 [International Schizophrenia Consortium, 2009]. 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 predictive accuracy [Vilhjálmsson et al., 2015]. More sophisticated Bayesian polygenic prediction methods that rely on GWAS summary statistics, including LDpred [Vilhjálmsson et al., 2015] and the normal-mixture model recently developed by Zhang et al. [2018], can incorporate genome-wide markers and accommodate varying genetic architectures, and thus have enhanced performance and flexibility. However, the type of prior used in these methods on SNP effect sizes, known as discrete mixture priors, imposes daunting computational challenges and may result in insufficient 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 substantial computational improvements. Several special cases of continuous shrinkage priors have been applied to quantitative trait prediction or gene mapping [De Los Campos et al., 2009; Hoggart et al., 2008; Makowsky et al., 2011; Xu, 2003; Yi and Xu, 2008]. 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 genome-wide association 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 [Bycroft et al., 2017; Sudlow et al., 2015], 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 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 [Gainer et al., 2016; Karlson et al., 2016; Smoller et al., 2016], and further demonstrate the potential of PRS-CS for the clinical translation of polygenic prediction.

Material and Methods

Conceptual frameworks. We consider a Bayesian high-dimensional regression framework for polygenic modeling and prediction: $\boldsymbol{y}_{N\times 1} = \boldsymbol{X}_{N\times M}\boldsymbol{\beta}_{M\times 1} + \boldsymbol{\epsilon}_{N\times 1}$, 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 the posterior mean effect sizes.

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

$$p(\beta_j) = \int \mathrm{N}(0, \Psi_j)\mathrm{d}G(\Psi_j), \tag{1}$$

or equivalently, as the following hierarchical form:

$$\beta_j \mid \Psi_j \sim \mathrm{N}(0, \Psi_j), \qquad \Psi_j \sim G, \tag{2}$$

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 [Yang et al., 2010]. 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, 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 [Vilhjálmsson et al., 2015]. 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 the effect sizes, which limits sampling 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:

$$\beta_j \mid \psi_j \sim \mathrm{N}(0, \phi\psi_j), \qquad \psi_j \sim g, \tag{3}$$

where ϕ is a global scaling parameter that 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, and encompasses many well-known priors as special cases. For example, if ψ_j follows an exponential distribution, then marginally β_j has independent Laplace (i.e., double-exponential) priors. This model is known as Bayesian LASSO [Hans, 2009; Park and Casella, 2008], because the Bayesian posterior mode estimate corresponds to the frequentist LASSO estimate. Often g is designed such that the prior distribution on the SNP effect sizes 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. PRS-CS further extends this framework to enable posterior inference of SNP effect sizes under continuous shrinkage priors using genome-wide association summary statistics and an external LD reference panel.

Overview of polygenic prediction methods. We compare PRS-CS with four polygenic prediction methods that rely on GWAS summary statistics: polygenic scoring based on all genetic markers (unadjusted PRS), informed LD-pruning and P-value thresholding (P+T), LDpred and LDpred-inf [Vilhjálmsson et al., 2015]. Throughout the paper, we use the 1000 Genomes Project [1000 Genomes Project Consortium, 2015] European

samples ($N = 503$) as the external LD reference panel. Below, we first briefly describe each of the existing methods, and then present PRS-CS in detail.

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 [Euesden et al., 2014] 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 \{$1E-8, 1E-7, 1E-6, 1E-5, 3E-5, 1E-4, 3E-4, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1$\}$, and report the best predictive performance across these thresholds in this paper.

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 the LD information from an external reference panel [Vilhjálmsson et al., 2015]. Consider the linear model $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, 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), $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_M]^\top$ is a vector of true effect sizes, and $\boldsymbol{\epsilon}$ is a vector of independent environmental effects. LDpred places an independent point-normal prior on each β_j:

$$\beta_j \sim \begin{cases} \mathrm{N}\left(0, \dfrac{h_g^2}{\pi M}\right), & \text{with probability } \pi \\ 0, & \text{with probability } 1-\pi, \end{cases} \tag{4}$$

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 to the GWAS summary statistics [Bulik-Sullivan et al., 2015], LDpred employs a Markov Chain Monte Carlo (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 \{$1E-5, 3E-5, 1E-4, 3E-4, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1$\}$, and report the highest prediction accuracy across these fraction values.

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:

$$\mathbf{E}[\boldsymbol{\beta}_\ell \mid \widehat{\boldsymbol{\beta}}_\ell, \boldsymbol{D}_\ell] \approx \left(\boldsymbol{D}_\ell + \frac{M}{N h_g^2} \boldsymbol{I} \right)^{-1} \widehat{\boldsymbol{\beta}}_\ell, \tag{5}$$

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 the local LD pattern, as suggested in Vilhjálmsson et al. [2015].

PRS-CS. Consider the phenotype model:

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}), \qquad p(\sigma^2) \propto \sigma^{-2}, \tag{6}$$

where both the phenotype $\boldsymbol{y}$ and the genotype matrix $\boldsymbol{Z}$ have been standardized, and we have assigned a non-informative scale-invariant Jefferey's prior on the residual variance σ^2. In contrast to discrete mixture priors such as the point-normal prior used in LDpred, we consider a conceptually different class of priors:

$$\beta_j \sim \mathrm{N}\left(0, \frac{\sigma^2}{N} \phi \psi_j\right), \qquad \psi_j \sim g, \tag{7}$$

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

$$\mathbf{E}[\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}] = (\boldsymbol{D} + \boldsymbol{T}^{-1})^{-1} \widehat{\boldsymbol{\beta}}, \tag{8}$$

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

$$\mathbf{E}[\beta_j \mid \widehat{\beta}_j] = \frac{1}{1+\phi^{-1}\psi_j^{-1}}\widehat{\beta}_j = \left(1 - \frac{1}{1+\phi\psi_j}\right)\widehat{\beta}_j := (1-\tau_j)\widehat{\beta}_j, \tag{9}$$

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 penalty 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:

$$\psi_j \sim \mathrm{G}(a, \delta_j), \qquad \delta_j \sim \mathrm{G}(b, 1), \tag{10}$$

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 [Armagan et al., 2011]:

$$\tau_j \sim \mathrm{TPB}(a, b, \phi), \tag{11}$$

where the TPB distribution has the following density function:

$$f(x; a, b, \phi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\phi^b x^{b-1}(1-x)^{a-1}\{1+(\phi-1)x\}^{-(a+b)}, \tag{12}$$

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 1 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 1, upper), 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 1, middle), along with heavy, Cauchy-like tails (Figure 1, lower). This prior is known as the horseshoe prior [Carvalho et al., 2010], 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. (7), Gibbs samplers with block updating of the regression coefficients (i.e., SNP effect sizes) $\boldsymbol{\beta}$ can be easily derived. By using LD information from an external reference panel, the method can be applied to genome-wide association summary statistics and does not require individual-level data. We describe the Gibbs sampler in Appendix A. 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): $a = 1$ and $b = 1/2$. This particular specification is known as the Strawderman-Berger prior [Berger, 1980; Strawderman, 1971] or the quasi-Cauchy prior [Johnstone and Silverman, 2004], 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 1000 Genomes Project European samples [Berisa and Pickrell, 2016], and conduct multivariate updating of the effect sizes within each LD block (see Appendix A). 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 treat the global shrinkage parameter ϕ as fixed in this work, and find that setting $\phi^{1/2}$ roughly to the proportion of causal variants [Piironen and Vehtari, 2016] works well. The predictive performance of the model is not sensitive to ϕ, and thus if a prior guess of the sparsity of the genetic architecture is not available, testing a small number of ϕ would be enough. In this work, when predicting disease and quantitative phenotypes in the Partners HealthCare Biobank (see below), we report the best prediction accuracy across four different ϕ values: $\phi^{1/2} \in \{0.001, 0.01, 0.1, 1\}$. The Gibbs sampler usually attains reasonable convergence after 1,000 MCMC iterations and produces prediction accuracy close to what can be achieved by much longer MCMC runs. We use 1,000 MCMC iterations with the first 500 steps as burn-in in simulation studies, and report the predictive performance of PRS-CS in Partners Biobank based on longer MCMC runs with 10,000 iterations in total and 5,000 burn-in steps.

UK Biobank genetic data. UK Biobank is a prospective cohort study of ~500,000 individuals recruited across Great Britain during 2006-2010 [Sudlow et al., 2015]. The protocol and consent were approved by the UK Biobank's Research Ethics Committee. Details about the UK Biobank project are provided at `http://www.ukbiobank.ac.uk`. Data for the current analyses were obtained under an approved data request.

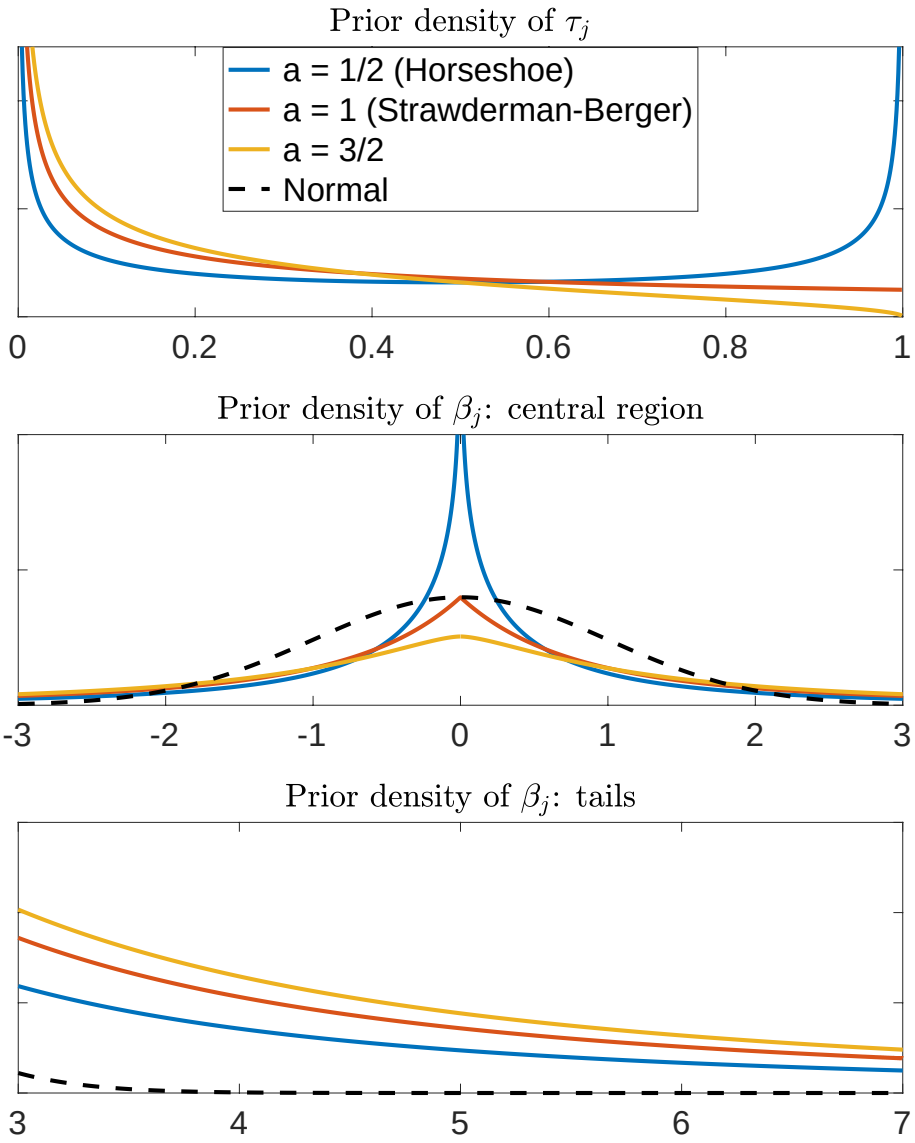

Figure 1: Densities of the priors. Upper panel: Density of the three-parameter beta prior on the shrinkage factor τ_j with $\phi = 1$, $b = 1/2$, and three different a values. Middle panel: Central region of the marginal prior density on the effect size β_j with $\phi = 1$, $b = 1/2$, and three different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

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+1000 Genomes reference panel. We leveraged the QC metrics provided by the UK Biobank [Bycroft et al., 2017] 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 [Chang et al., 2015].

Simulations. We performed simulation studies using real genetic data from the UK Biobank. We used the point-normal model specified in Eq. (4) to sample SNP effect sizes. The simulated trait was generated by the sum of all genetic markers, weighted by their simulated effect sizes, for each individual, and adding a normally distributed noise term which fixed the heritability at 0.5. We then conducted GWAS to produce marginal least squares effect size estimate for each SNP. The five polygenic prediction methods were applied to the GWAS summary statistics, and their predictive performance was evaluated in 3,000 individuals (the validation sample) that are unrelated to the training sample. R^2 between the observed and predicted traits was used to quantify the prediction accuracy. We considered 100, 1,000, 10,000 and 100,000 causal variants in the simulations, which represent extremely sparse to highly polygenic genetic architectures, and four different training sample sizes: 10,000, 20,000, 50,000 and 100,000. For each combination of the number of causal variants and the training sample size, the simulation was repeated 100 times.

We conducted secondary simulation studies using a point-t model (a mixture of a point mass at zero and a Student's t-distribution with 4 degrees of freedom), and a normal mixture model. The 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 effect sizes for the rest of the SNPs were set to zero. In all secondary simulations, the heritability was fixed at 0.5, and we considered four different training sample sizes: 10,000, 20,000, 50,000 and 100,000. For each combination of the genetic architecture (the point-t model with different number of causal variants and the normal mixture model) and the training sample size, the simulation was repeated 20 times.

Partners HealthCare Biobank genetic data. The Partners HealthCare 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 [Karlson et al., 2016]. To date, Partners Biobank has enrolled more than 88,800 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 1000 Genomes (1KG) Project samples as a population reference panel to infer the ancestry of Partners Biobank participants. Specifically, we computed principal components (PC) 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 [Loh et al., 2016] for pre-phasing and Minimac3 [Das et al., 2016; Howie et al., 2009] for imputation in the Partners Biobank European samples. 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 [Chang et al., 2015].

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 (Table S1). We removed ambiguous (A/T and C/G) SNPs and mapped the genetic markers to the Genome Reference Consortium human genome build 37. For unadjusted PRS and P+T, we used all the genetic markers that are present in the summary statistics, LD reference panel (1000 Genomes Project) and the Partners Biobank genetic data. For LDpred and PRS-CS, 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. We use R^2 between the observed and predicted phenotypes to assess the predictive performance for the quantitative traits, and report the Nagelkerke's R^2 metric for disease (case-control) phenotypes. For all the analyses, we adjusted for current age, sex and top 10 principal components of the genotype data.

Results

Simulations. We compared the predictive performance of five polygenic prediction methods across different genetic architectures and training sample sizes in the simulation studies. Results are shown in Figure 2 and the corresponding numerical values can be found in Table S2.

We first note that 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, because as more causal SNPs fall in the same LD block and their effect sizes decline, it becomes increasingly difficult to distinguish real signals from noise.

Table 1: Information on the six 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 of the validation data set in the Partners HealthCare Biobank (PBK).

Disease/Trait	Abbreviation	GWAS Reference	GWAS sample size (case/control)	1KG & PBK SNPs	1KG & HM3 & PBK SNPs	PBK sample size (case/control)
Breast Cancer	BRCA	Michailidou et al. [2017]	228,951 (122,977/105,974)	5,022,127	857,616	10,220 (884/9,336)
Coronary Artery Disease	CAD	Nikpay et al. [2015]	184,305 (60,801/123,504)	4,803,592	849,399	16,251 (2,759/13,492)
Depression	DEP	Wray et al. [2018]	173,005 (59,851/113,154)	4,924,025	850,291	15,276 (2,361/12,915)
Inflammatory Bowel Disease	IBD	Liu et al. [2015]	34,652 (12,882/21,770)	4,823,570	849,749	18,998 (750/18,248)
Rheumatoid Arthritis	RA	Okada et al. [2014]	58,284 (14,361/43,923)	3,872,637	849,680	18,170 (753/17,417)
Type 2 Diabetes Mellitus	T2DM	Scott et al. [2017]	159,208 (26,676/132,532)	4,901,848	856,912	18,823 (1,978/16,845)
Height	HGT	Yengo et al. [2018]	693,529	1,578,533	750,888	3,957
Body mass index	BMI	Yengo et al. [2018]	681,275	1,579,905	751,676	3,954
High-density lipoproteins	HDL	Willer et al. [2013]	188,578	1,604,577	758,036	2,491
Low-density lipoproteins	LDL	Willer et al. [2013]	188,578	1,600,625	756,724	1,713
Cholesterol	CHOL	Willer et al. [2013]	188,578	1,604,391	757,970	2,561
Triglycerides	TRIG	Willer et al. [2013]	188,578	1,601,270	756,913	2,505

Overall, methods that account for the local LD pattern (LDpred and PRS-CS) outperformed P+T, which discards LD information. However, one unexpected observation is that the prediction accuracy of LDpred decreased quite 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 adequately adjust for the LD structure, resulting in a decrease in predictive performance. In contrast, PRS-CS was 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 the 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. Secondary simulations using the point-t model and the normal mixture model produced similar results (Figure S1 and Table S4).

Polygenic prediction in the Partners Biobank. We applied PRS-CS and alternative methods to predict six curated 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, for which external large-scale GWAS summary statistics are publicly available. Predictive perfor-

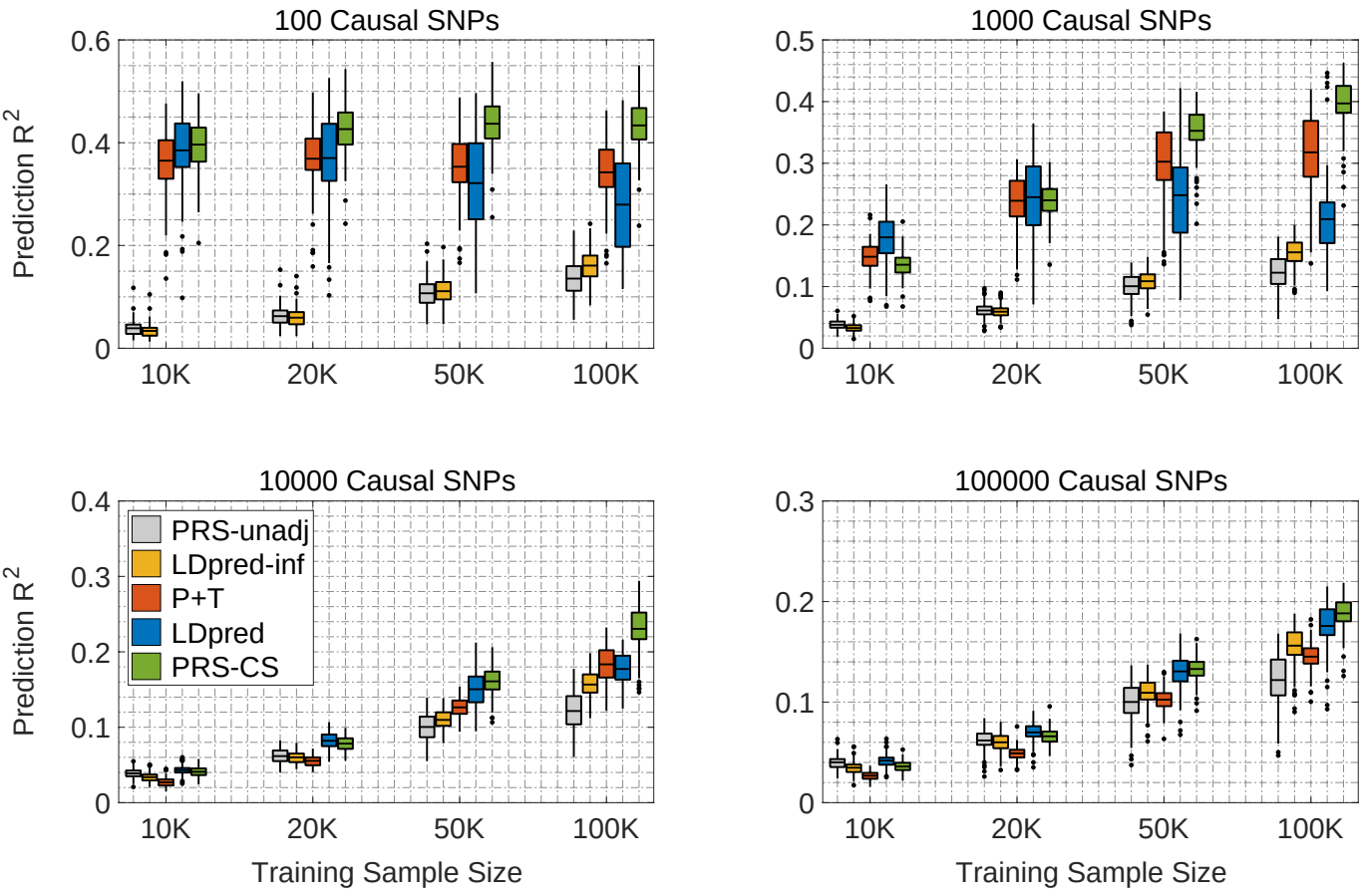

Figure 2: Prediction accuracy, quantified by R^2 between the observed and predicted traits, of five polygenic prediction methods in simulation studies. The four panels correspond to the four genetic architectures (100, 1,000, 10,000 and 100,000 causal variants) simulated using the point-normal model. 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 100 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.

mance of the five methods are summarized in Figure 3, and the corresponding numerical values can be found in Table S3.

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 (43.58% relative increase in Nagelkerke's R^2 compared to LDpred) and rheumatoid arthritis (30.11% 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 other methods. PRS-CS was only inferior to LDpred in the prediction of inflammatory bowel disease (8.48% 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.

For the six quantitative traits, PRS-CS consistently outperformed all alternative methods. The relative improvement in prediction accuracy compared to LDpred ranged from 8.70% for LDL and 8.80% for BMI, to 25.60% for height and 32.75% for cholesterol, with an average improvement of 18.21%. The average improvement of PRS-CS relative to P+T across the six quantitive traits was 40.47%. We note that LDpred was the second best method for most quantitative traits, but its predictive accuracy for height 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 [Marquez-Luna et al., 2018].

Discussion

Polygenic prediction, which exploits genome-wide genetic markers to estimate the genetic liability to a common disease or complex 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 all genetic markers across the genome employ Bayesian approaches and assign a discrete mixture prior on the 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 updating of

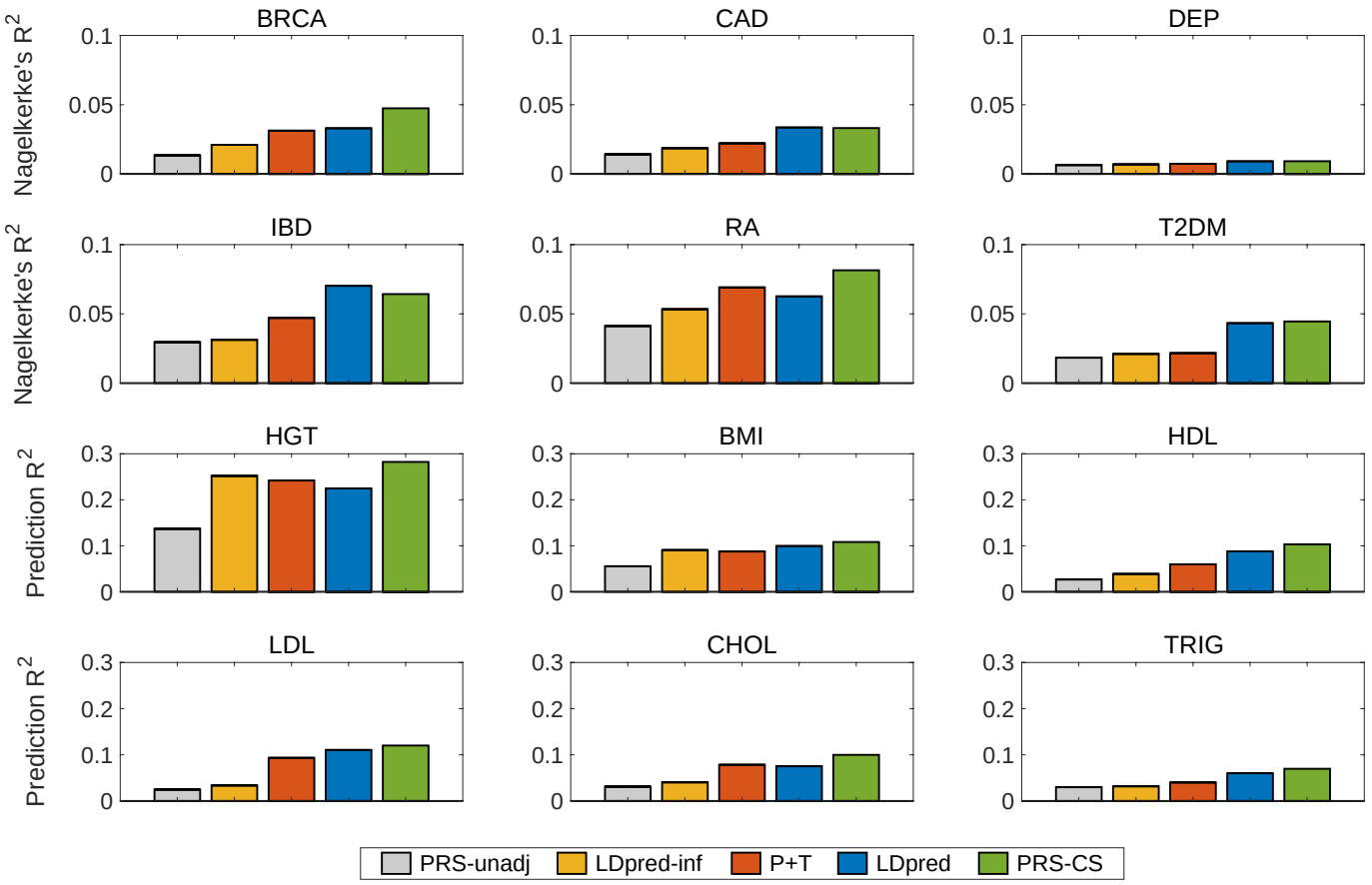

Figure 3: Prediction accuracy of five polygenic prediction methods in the Partners HealthCare Biobank. Polygenic scores were trained with external large-scale genome-wide association summary statistics, and applied to predict six curated complex diseases — breast cancer (BRCA), coronary artery disease (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). For disease (case-control) phenotypes, prediction accuracy is measured by the Nagelkerke's R^2. For quantitive traits, prediction accuracy is quantified by R^2.

effect sizes and thus hinder accurate LD adjustment in polygenic prediction. LDpred [Vilhjálmsson et al., 2015] 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 and the adjustment for LD may be inadequate.

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 Gibbs sampler with multivariate block updating 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 the 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 [see e.g., Marquez-Luna et al., 2018; Vilhjálmsson et al., 2015]. 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.

Although continuous shrinkage priors enable multivariate modeling of the LD structure, simultaneous updating of the effect sizes for genome-wide markers remains computationally infeasible and, in fact, unnecessary. In this work, we used a genome partition computed and validated by prior work [Berisa and Pickrell, 2016], which divides the genome into 1,703 largely independent genomic regions, and has been successfully used in local heritability and genetic correlation analyses [Shi et al., 2016a, 2017]. Expanding the size of LD blocks may improve prediction accuracy but also increases computational cost, 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. Using a pre-computed genome partition to model local LD patterns is also more memory and computationally efficient relative to a sliding window approach as implemented in LDpred [Vilhjálmsson et al., 2015].

We note that the prior we investigated in this work, i.e., the gamma-gamma prior on the local shrinkage parameter (also known as the generalized beta mixture model) [Armagan et al., 2011], or more specifically, the Strawderman-Berger prior [Berger, 1980; Strawderman, 1971], is only one of the possible choices within

the class of continuous shrinkage priors, which includes the normal-gamma prior [Caron and Doucet, 2008; Griffin and Brown, 2010], the normal-inverse-gaussian prior [Caron and Doucet, 2008], the generalized t (generalized double Pareto) prior [Armagan et al., 2013; Lee et al., 2012], and the normal-exponential-gamma prior [Armagan et al., 2011; Griffin and Brown, 2011], 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, i.e., a different g in Eq. (7), 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 the local LD pattern 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, in contrast to fixing the global shrinkage parameter ϕ in the model based on prior beliefs about the sparsity of the genetic architecture, or searching a small number of grid values, the global parameter could be learnt from data using empirical Bayes or a full Bayesian approach by placing, for example, a half-Cauchy prior on it [Gelman, 2006; Polson and Scott, 2010]. This would make PRS calculation fully automatic and reduce the potential risk of overfitting. Second, although this paper is focused on polygenic prediction methods that only require GWAS summary statistics, PRS-CS can be straightforwardly applied to individual-level data. Given that a majority of the existing Bayesian genomic prediction models, including Bayes alphabetic methods [Habier et al., 2011; Hayes et al., 2010; Meuwissen and Goddard, 2004; Meuwissen et al., 2001; Verbyla et al., 2009, 2010; Yi et al., 2003], BayesR [Erbe et al., 2012; Moser et al., 2015], BVSR [Guan and Stephens, 2011], BSLMM [Zhou et al., 2013], and DPR [Zeng and Zhou, 2017], 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. Third, 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 [Marquez-Luna et al., 2018; Shi et al., 2016b; Turley et al., 2018]. Lastly, current research on polygenic prediction has largely been restricted

to European samples. Expanding genomic prediction methods to 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, all curated disease phenotypes we predicted had variance explained less than 10%, which is considerably lower than their heritability. Therefore, much work is needed to further improve the performance of PRS. In theory, the utility of PRS depends on multiple factors, including the training 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 [Zhang et al., 2018]. In contrast, it may be easier to produce a clinically useful prediction for some autoimmune diseases or late-onset chronic diseases, 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 will demonstrate bigger advantages over existing methods with larger training sample sizes.

Appendix A

The Bayesian regression model for PRS-CS is:

$$\begin{aligned} &\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \qquad p(\sigma^2) \propto \sigma^{-2}, \\ &\beta_j \sim \mathrm{N}\left(0, \frac{\sigma^2}{N}\psi_j\right), \qquad \psi_j \sim \mathrm{G}(a, \delta_j), \qquad \delta_j \sim \mathrm{G}(b, \phi), \end{aligned} \tag{13}$$

where $\boldsymbol{y}$ and $\boldsymbol{Z}$ have been standardized. The full conditional distributions for all the parameters in this model are analytically tractable, and thus an efficient Gibbs sampler can be derived.

Let $\mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$; $\mathrm{G}(\alpha, \beta)$ and $\mathrm{iG}(\alpha, \beta)$ denote the gamma distribution and inverse-gamma distribution with shape parameter α and scale parameter β, respectively; and $\mathrm{giG}(p, \rho, \chi)$ denote the three-parameter generalized inverse Gaussian distribution with probability density function

$$f(x; \lambda, \rho, \chi) = \frac{(\rho/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\rho\chi})} x^{\lambda-1} \mathrm{e}^{-(\rho x + \chi/x)/2}, \qquad x > 0, \quad \rho > 0, \quad \chi > 0, \tag{14}$$

where K_λ is the modified Bessel function of the second kind. Let N and M denote the sample size and the total number of genetic markers, respectively. In addition, let $\widehat{\boldsymbol{\beta}} = \boldsymbol{Z}^\top \boldsymbol{y}/N$ denote the marginal least squares effect size estimates from the genome-wide association study, $\boldsymbol{\Psi} = \text{diag}\{\psi_1, \psi_2, \cdots, \psi_M\}$, and $\boldsymbol{D} = \boldsymbol{Z}^\top \boldsymbol{Z}/N$ denote the LD matrix. The Gibbs sampler then involves the following steps in each MCMC iteration:

- update $\boldsymbol{\beta}$: $\quad [\, \boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\Psi}, \widehat{\boldsymbol{\beta}}, \boldsymbol{D} \,] \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \dfrac{N}{\sigma^2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}, \quad \boldsymbol{\Sigma} = \dfrac{\sigma^2}{N}(\boldsymbol{D} + \boldsymbol{\Psi}^{-1})^{-1},$
- update σ^2: $\quad [\, \sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \widehat{\boldsymbol{\beta}}, \boldsymbol{D} \,] \sim \text{iG}\left(\dfrac{N+M}{2}, \dfrac{N}{2}\left[1 - 2\boldsymbol{\beta}^\top\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta}^\top(\boldsymbol{D} + \boldsymbol{\Psi}^{-1})\boldsymbol{\beta}\right]\right),$
- update ψ_j: $\quad [\, \psi_j \mid \beta_j, \sigma^2, \delta_j \,] \sim \text{giG}\left(a - \dfrac{1}{2}, 2\delta_j, \dfrac{N\beta_j^2}{\sigma^2}\right),$
- update δ_j: $\quad [\, \delta_j \mid \psi_j \,] \sim \text{G}(a + b, \psi_j + \phi).$

We generate random variates from the generalized inverse Gaussian distribution using the algorithm described in Devroye [2014]. We note that $\boldsymbol{y}$ and $\boldsymbol{Z}$ did not appear in any of the updates, and thus individual-level data is not required for model fitting. In practice, $\boldsymbol{D}$ and $\boldsymbol{\Psi}$ are $M \times M$ matrices, and the calculation of $(\boldsymbol{D} + \boldsymbol{\Psi}^{-1})^{-1}$ becomes computationally infeasible when M is large. We thus partition the genome into 1,703 largely independent genomic regions estimated using data from the 1000 Genomes Project European samples [Berisa and Pickrell, 2016], and in each MCMC iteration sequentially update the SNP effect sizes within each LD block ℓ:

$$[\, \boldsymbol{\beta}_\ell \mid \sigma^2, \boldsymbol{\Psi}_\ell, \widehat{\boldsymbol{\beta}}_\ell, \boldsymbol{D}_\ell \,] \sim \text{MVN}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell), \quad \boldsymbol{\mu}_\ell = \frac{N}{\sigma^2}\boldsymbol{\Sigma}_\ell\widehat{\boldsymbol{\beta}}_\ell, \quad \boldsymbol{\Sigma}_\ell = \frac{\sigma^2}{N}(\boldsymbol{D}_\ell + \boldsymbol{\Psi}_\ell^{-1})^{-1}. \tag{15}$$

The LD matrix $\boldsymbol{D}_\ell$ for each LD block can be estimated using an external LD reference panel.

Acknowledgements

This work involved the use of the Enterprise Research Infrastructure & Services (ERIS) at Partners HealthCare. We thank the Partners HealthCare Biobank for providing samples, genomic data, and health information data. This research was funded in part by NIH grants K99AG054573 (TG) and K24MH094614 (JWS). JWS is a Tepper Family MGH Research Scholar and was also 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.

Web Resources

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

References

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571): 68–74, 2015.

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.

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

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

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

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.

C. Bycroft, C. Freeman, D. Petkova, G. Band, L.T. Elliott, et al. Genome-wide genetic data on ∼500,000 UK Biobank participants. *bioRxiv*, 166298, 2017.

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

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

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.

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.

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.

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.

L. Devroye. Random variate generation for the generalized inverse Gaussian distribution. *Statistics and Computing*, 24(2):239–246, 2014.

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.

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

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.

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

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

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.

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.

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.

C. Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

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.

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.

B.N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, 2009.

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

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.

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.

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*, in press, 2018.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

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

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

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.

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, 2016a.

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.

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, 2016b.

J.W. Smoller, E.W. Karlson, R.C. Green, S. Kathiresan, D.G. MacArthur, et al. An eMERGE clinical center at partners personalized medicine. *Journal of Personalized Medicine*, 6(1):5, 2016.

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

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.

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.

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.

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.

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.

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.

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.

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

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.

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. *bioRxiv*, 274654, 2018.

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

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

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

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*, in press, 2018.

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

Supplementary Tables

Table S1: Information on the genome-wide association summary statistics of the six 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).

Table S2: Numerical results of the simulation studies shown in Figure 2. For each combination of the number of causal variants (100, 1,000, 10,000 and 100,000) and the training sample size (10,000, 20,000, 50,000 and 100,000), the mean and standard error of the prediction accuracy for each polygenic prediction method across 100 simulations are reported.

Table S3: Numerical values of the prediction accuracy shown in Figure 3. For each of the curated diseases (breast cancer, coronary artery disease, depression, inflammatory bowel disease, rheumatoid arthritis, and type 2 diabetes mellitus), and quantitative traits (height, body mass index, high-density lipoproteins, low-density lipoproteins, cholesterol, and triglycerides), the prediction accuracy for each of the polygenic prediction methods is reported.

Table S4: Numerical results of the simulation studies shown in Figure S1. For each combination of the genetic architecture (the point-t model with different numbers of causal variants and the normal mixture model) and the training sample size (10,000, 20,000, 50,000 and 100,000), the mean and standard error of the prediction accuracy for each polygenic prediction method across 20 simulations are reported.

Supplementary Figures

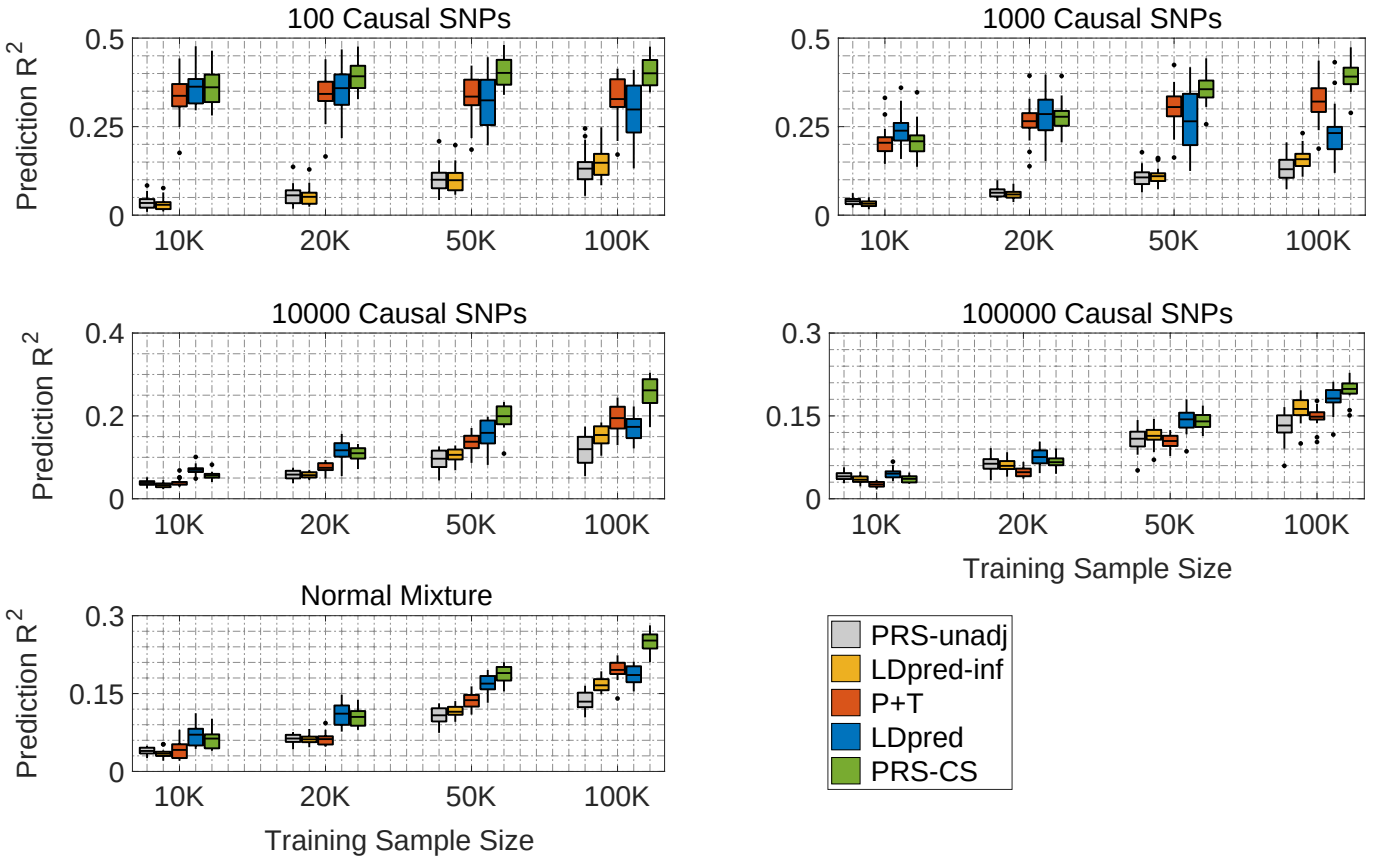

Figure S1: Prediction accuracy, quantified by R^2 between the observed and predicted traits, of five polygenic prediction methods in secondary simulation studies using the point-t model and the normal mixture model. The upper four panels correspond to the four genetic architectures simulated using the point-t model and different numbers of causal variants. The lower panel corresponds to the genetic architecture simulated using 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.