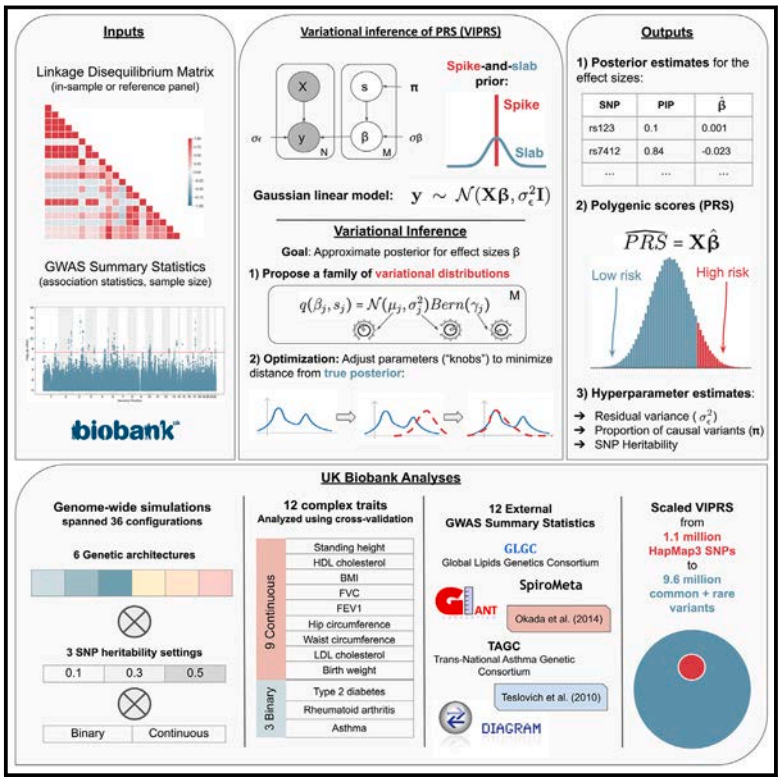
Fast and accurate Bayesian polygenic risk modeling with variational inference

Graphical abstract



Authors

Shadi Zabad, Simon Gravel, Yue Li

Correspondence

simon.gravel@mcgill.ca (S.G.),
yueli@cs.mcgill.ca (Y.L.)

We present VIPRS, a fast and accurate variational Bayesian method for estimating polygenic risk scores from genome-wide association study (GWAS) data. The method is shown to be robust and competitively accurate against popular baselines and scales well to dense genotype array data.

Zabad et al., 2023, The American Journal of Human Genetics *110*, 1–21
May 4, 2023

https://doi.org/10.1016/j.ajhg.2023.03.009

CellPress

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

ARTICLE

Fast and accurate Bayesian polygenic risk modeling with variational inference

Shadi Zabad,[1] Simon Gravel,[2,*] and Yue Li[1,*]

Summary

The advent of large-scale genome-wide association studies (GWASs) has motivated the development of statistical methods for phenotype prediction with single-nucleotide polymorphism (SNP) array data. These polygenic risk score (PRS) methods use a multiple linear regression framework to infer joint effect sizes of all genetic variants on the trait. Among the subset of PRS methods that operate on GWAS summary statistics, sparse Bayesian methods have shown competitive predictive ability. However, most existing Bayesian approaches employ Markov chain Monte Carlo (MCMC) algorithms, which are computationally inefficient and do not scale favorably to higher dimensions, for posterior inference. Here, we introduce variational inference of polygenic risk scores (`VIPRS`), a Bayesian summary statistics-based PRS method that utilizes variational inference techniques to approximate the posterior distribution for the effect sizes. Our experiments with 36 simulation configurations and 12 real phenotypes from the UK Biobank dataset demonstrated that `VIPRS` is consistently competitive with the state-of-the-art in prediction accuracy while being more than twice as fast as popular MCMC-based approaches. This performance advantage is robust across a variety of genetic architectures, SNP heritabilities, and independent GWAS cohorts. In addition to its competitive accuracy on the "White British" samples, `VIPRS` showed improved transferability when applied to other ethnic groups, with up to 1.7-fold increase in R^2 among individuals of Nigerian ancestry for low-density lipoprotein (LDL) cholesterol. To illustrate its scalability, we applied `VIPRS` to a dataset of 9.6 million genetic markers, which conferred further improvements in prediction accuracy for highly polygenic traits, such as height.

Introduction

In recent years, with the rapid growth of large-scale biobank data with comprehensive genotyping and phenotyping efforts,[1–3] there has been growing interest in developing statistical methods to quantify an individual's disease risk from their genotype data.[4–8] At the same time, these rich biobank data sources have powered many recent analyses of complex traits and diseases, revealing highly polygenic architectures[9–11] with a wide range of effect sizes across different genomic categories.[12–14] Linear models are an important framework for complex trait analysis that allow for the estimation of the additive genetic component of a phenotype, also known as a polygenic score (PGS) or polygenic risk score (PRS) in clinical contexts.[5,15] Even though many examples of genetic interactions have been documented, such additive effects capture much of the genetic variation underlying human complex traits.[16,17] Recent work has highlighted the clinical relevance of polygenic scores for some diseases and health conditions,[18,19] especially in applications related to disease risk stratification[20–22] and personalized medicine.[23]

Estimating PGSs from genome-wide association study (GWAS) data has a long and rich history in the field of quantitative genetics as well as the animal and plant breeding literature.[24,25] In human and medical genetics, it remains an active area of research, with numerous methods recently developed.[4,6,26–34] Standard PRS methods formulate the problem of polygenic risk estimation in terms of a multiple linear regression framework, where the goal is to infer the joint effect sizes of all genetic variants on the trait. The most common class of genetic variation considered in these analyses are single-nucleotide polymorphisms (SNPs), which are either measured by modern genotyping arrays or statistically imputed with reference haplotypes.[35,36]

Genotyping arrays combined with imputation can accurately capture the genotype of an individual at millions of genetic markers. When paired with modern GWAS sample sizes routinely exceeding hundreds of thousands of individuals, high dimensional data of this scale present several computational and statistical challenges. Furthermore, most individual-level GWAS data sources are protected for privacy concerns.[37] These two factors motivated the development of a number of PRS methods that estimate PRSs on the basis of GWAS summary statistics,[4,6,27–29,31–34] which are the marginal test statistic per SNP.

Within this class of summary statistics-based methods, Bayesian PRS models enable a principled way to incorporate prior knowledge as probability distributions over the genetic causal architecture of complex traits. In addition to providing meaningful estimates of parameter uncertainties,[38] Bayesian approaches have shown competitive predictive ability, exceeding the predictive performance of heuristic or penalized estimators in many settings.[6,28,32,33,39,40] However, a major limitation of some existing Bayesian methods is that their scalability is

[1]School of Computer Science, McGill University, Montreal, QC, Canada; [2]Department of Human Genetics, McGill University, Montreal, QC, Canada
*Correspondence: simon.gravel@mcgill.ca (S.G.), yueli@cs.mcgill.ca (Y.L.)
https://doi.org/10.1016/j.ajhg.2023.03.009.

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

hampered by slow and inefficient inference techniques. While heuristic methods such as clumping-and-thresholding (C+T) are routinely applied on millions of SNPs, Bayesian approaches are generally restricted to a smaller subset of approximately one million genetic markers. One of the main reasons for this limitation stems from computational considerations: most Bayesian PRS methods employ Markov chain Monte Carlo (MCMC) algorithms to approximate the posterior for the effect sizes.[4,6,28,32] MCMC algorithms are known to be asymptotically accurate but often slow to converge.[41,42] In practice, to obtain accurate posterior estimates, the MCMC chains need to be run for hundreds or thousands of iterations.[4,6] This challenge can be partially remedied with the help of efficient software implementation and enhanced linear algebra routines, which recently enabled scaling up two well-known MCMC-based Bayesian PRS methods to several million SNPs.[6,40] While this is an important advance, these variants still constitute a small fraction of the genetic variation that can be assayed by modern whole-genome-sequencing technologies.[3,43]

An alternative scheme for approximating the posterior density for the effect sizes is variational inference (VI), a fast and deterministic class of algorithms that recast the problem of posterior inference in the form of an optimization problem.[41,42,44,45] Variational methods have seen a surge of interest in the machine learning literature in recent years as a result of significant advances in stochastic optimization techniques.[46,47] Methods that utilize VI have been explored in a wide variety of statistical genetics applications, specifically in the context of linear mixed models (LMMs),[48] association mapping,[49,50] fine-mapping,[51,52] and enrichment analysis,[53,54] among others. More recently, a number of studies examined the properties and relative accuracy of certain variational approximations to PRS by using both individual level data and summary statistics.[34,55–57]

In this work, we present variational inference of polygenic risk scores (`VIPRS`), a Bayesian summary statistics-based PRS method that utilizes VI to approximate the posterior for the effect sizes. We conduct a comprehensive set of experiments by using simulated and real traits to assess the predictive ability of `VIPRS` in comparison with the some of the most popular Bayesian and non-Bayesian PRS methods. Overall, we show that `VIPRS` is a scalable and flexible method that enjoys the speed and efficiency of heuristic approaches, such as clumping-and-thresholding (C+T), while rivaling state-of-the-art Bayesian methods in terms of its predictive performance. We demonstrate the flexibility of the method by testing its predictions with different families of priors on the effect size, paired with four distinct strategies for tuning the hyperparameters of the model. To illustrate its scalability, we evaluate the predictive accuracy of `VIPRS` with approximately 9.6 million SNPs, almost an order of magnitude greater than the standard HapMap3 subset routinely used for this task. This allows us to examine the potential for phenotype prediction by using a more comprehensive set of genetic variants segregating in the human population.

Material and methods

Overview of the `VIPRS` model

Given a random sample of individuals from a general population with paired genotype and phenotype data, we model the dependence of the phenotype on the genotype via the standard linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \text{(Equation 1)}$$

where $\mathbf{y}$ is an $N\times1$ vector of phenotypic measurements for N individuals, $\mathbf{X}$ is the $N\times M$ genotype matrix that records the counts of alternative alleles for each individual at each genetic marker, $\boldsymbol{\beta}$ is a vector of effect sizes for each of the M markers, and $\boldsymbol{\epsilon}$ is an $N\times1$ vector that captures the residual effects on the trait for each individual. Our model derivation assumes that both the genotype matrix $\mathbf{X}$ and phenotype vector $\mathbf{y}$ are column-wise standardized to have zero mean and unit variance. For quantitative traits, we assume that the phenotypes follow a Gaussian likelihood, such that $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_\epsilon^2\mathbf{I})$, where σ_ϵ^2 is the residual variance. For case-control traits, we model the latent continuous liability underlying the disease,[58] which is assumed to follow the same Gaussian likelihood. Since our method operates on summary statistics, we followed the common practice of converting the marginal statistics of binary phenotypes from the log-odds to the liability scale.[4,59,60] Practically, a central goal of polygenic risk modeling is to arrive at a robust estimate for the effect sizes $\boldsymbol{\beta}$. In the Bayesian framework, this problem is tackled by imposing a prior distribution over the effect sizes and then deriving a solution for the posterior distribution given the data likelihood and the prior,

$$p(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}, \theta) \propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \theta)p(\boldsymbol{\beta}, \theta), \qquad \text{(Equation 2)}$$

where θ encapsulates all fixed hyperparameters in the model, i.e., parameters that we do not assign a prior. Here, the constant of proportionality is the marginal likelihood or the partition function for the posterior, $\int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \theta)p(\boldsymbol{\beta}, \theta)d\boldsymbol{\beta}$, also known as the model evidence.[41,42] In recent years, considerable work has been devoted to deriving Bayesian PRS models with flexible priors on the effect sizes, such as the continuous shrinkage[28] and mixture priors.[6] In this work, we follow the lead of earlier approaches, e.g., Vilhjálmsson et al.[4] and Carbonetto and Stephens,[51] and assign a spike-and-slab prior[61,62] on the effect sizes,

$$\beta_j \sim \pi\mathcal{N}\left(\beta_j; 0, \sigma_\beta^2\right) + (1 - \pi)\delta_0. \qquad \text{(Equation 3)}$$

Here, π is a parameter that denotes the prior probability that a variant is causal, σ_β^2 is the prior variance on the effect size of each SNP, and δ_0 is the Dirac delta function. In the simplest formulation of this model, we assume that π and σ_β^2 are shared across all SNPs. Thus, π may also be considered as the fraction of variants that are causal for the trait of interest, and the σ_β^2 parameter is related to the trait's per-SNP heritability.[4,63] The spike-and-slab prior is a special case of the more general mixture prior:

$$p\left(\beta_j|s_j\right)p(s_j) = \prod_{k=1}^{K}\mathcal{N}\left(\beta_j; 0, \sigma_k^2\right)^{s_{jk}}\prod_{k=1}^{K}\pi_k^{s_{jk}}, \qquad \text{(Equation 4)}$$

where s_{jk} is binary indicator for SNP j belonging to the k^{th} mixture component and π_k and σ_k^2 denote the the mixing proportion and prior variance for component k, respectively. It is well known that

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

Bayesian linear regression models with a spike-and-slab prior on the effect sizes result in an intractable posterior,[51,61,62] necessitating the use of approximate posterior inference schemes.

VIPRS model inference

In most of the previous Bayesian PRS formulations, the authors employ a Gibbs sampler, a MCMC technique that relies on conditional conjugacy between the prior and the likelihood, to approximate the posterior distribution of the effect sizes.[4,6,28,32] In this work, we instead leverage a technique known as Variational Inference (VI),[44] which approximates intractable densities by proposing a simple parametric distribution $q(\boldsymbol{\beta}, \mathbf{s})$ and optimizing its parameters to match the true posterior as closely as possible.[45] The closeness between the true posterior and the proposed distribution is measured by the Kullback-Leibler (KL) divergence,

$$KL[q\|p] = \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log q(\boldsymbol{\beta},\mathbf{s})] - \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log p(\boldsymbol{\beta},\mathbf{s}|\mathbf{X},\mathbf{y},\theta)] \quad \text{(Equation 5)}$$

$$= \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log q(\boldsymbol{\beta},\mathbf{s})] - \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log p(\mathbf{y},\boldsymbol{\beta},\mathbf{s}|\theta,\mathbf{X})] + \log p(\mathbf{y}|\mathbf{X},\theta), \quad \text{(Equation 6)}$$

where $\mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}$ is the expectation taken with respect to the proposed distribution.[41,42] However, the KL divergence includes the normalizing constant that made the posterior intractable in the first place. Thus, practitioners typically optimize a surrogate objective known as the evidence lower bound (ELBO) of the log marginal likelihood[41,42,45]:

$$\begin{aligned}\mathcal{L} &= \log p(\mathbf{y}|\mathbf{X},\theta) = \log \sum_{\mathbf{s}} \int \frac{p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\theta)p(\boldsymbol{\beta},\mathbf{s}|\theta)}{q(\boldsymbol{\beta},\mathbf{s})} q(\boldsymbol{\beta},\mathbf{s})d\boldsymbol{\beta} \\ &\geq \sum_{\mathbf{s}} \int q(\boldsymbol{\beta},\mathbf{s}) \log p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\mathbf{s},\boldsymbol{\theta})p(\boldsymbol{\beta},\mathbf{s}|\boldsymbol{\theta})d\boldsymbol{\beta} - \sum_{\mathbf{s}} \int q(\boldsymbol{\beta},\mathbf{s}) \log q(\boldsymbol{\beta},\mathbf{s})d\boldsymbol{\beta} \\ &= \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log p(\mathbf{y},\boldsymbol{\beta},\mathbf{s}|\theta,\mathbf{X})] - \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log q(\boldsymbol{\beta},\mathbf{s})] \equiv ELBO\end{aligned} \quad \text{(Equation 7)}$$

Here, the first term $\mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log p(\mathbf{y},\boldsymbol{\beta},\mathbf{s}|\theta,\mathbf{X})]$ in Equation 7 is the expectation of the log joint likelihood of the phenotypes and the effect sizes and the second term $-\mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log q(\boldsymbol{\beta},\mathbf{s})]$ corresponds to the entropy of the variational distribution. The ELBO in Equation 7 and the KL-Divergence in Equation 6 add up to the marginal likelihood: $ELBO + KL[q\|p] = \mathcal{L}$. Therefore, maximizing ELBO is equivalent to minimizing the KL-Divergence.[41,42,45]

The choice of approximating variational distribution $q(\boldsymbol{\beta}, \mathbf{s})$ is a central component in this setting. For simplicity and computational efficiency, we make use of the paired mean-field assumption,[41,42,64] whereby the density factorizes across the input coordinates, and model the effect size at each locus with a two-component Gaussian mixture density,[51,64,65]

$$q(\boldsymbol{\beta},\mathbf{s}) = \prod_j^M q\left(\beta_j, s_j\right) = \prod_j^M \mathcal{N}\left(\beta_j;\mu_j,\sigma_j^2\right)\text{Bern}\left(s_j;\gamma_j\right). \quad \text{(Equation 8)}$$

Here, $\mu_j, \sigma_j^2, \gamma_j$ are the variational parameters defined for each variant in the dataset and $\text{Bern}\left(s_j;\gamma_j\right) = \gamma_j^{s_j}\left(1-\gamma_j\right)^{1-s_j}$ denotes a Bernoulli distribution with probability γ_j for SNP j. Therefore, the Bernoulli indicator in the proposed distribution approximates the posterior probability that the variant is causal for the trait of interest and the Gaussian component approximates the posterior for the effect size.[51] In the supplemental methods, we provide detailed derivations that show that, under certain assumptions, this variational family leads to the following closed-form updates that only depend on GWAS summary statistics and the SNP-by-SNP correlation or linkage disequilibrium (LD) matrix, which can be derived from an appropriately matched reference panel:

$$\sigma_j^2 = \frac{\sigma_\epsilon^2}{N + \sigma_\epsilon^2 \big/ \sigma_{\boldsymbol{\beta}}^2} \quad \text{(Equation 9)}$$

$$\mu_j = \frac{N\sigma_j^2}{\sigma_\epsilon^2}\left(\widehat{\beta}_j - \sum_{k\neq j}\gamma_k\mu_k R_{jk}\right) \quad \text{(Equation 10)}$$

$$\gamma_j = \text{Sigmoid}\left(\log\left(\frac{\pi}{1-\pi}\right) + \frac{1}{2}\log\left(\frac{\sigma_j^2}{\sigma_\beta^2}\right) + \frac{1}{2\sigma_j^2}\mu_j^2\right). \quad \text{(Equation 11)}$$

Assuming standardized genotype and phenotype, $\widehat{\beta}_j = \mathbf{y}^\top\mathbf{x}_j/N$ is the marginal GWAS effect size of SNP j, $R_{jk} = \mathbf{x}_j^\top\mathbf{x}_k/N$ is the LD between a pair of SNPs j and k, and the sigmoid function is defined as $\text{Sigmoid}(x) = 1/(1+\exp(-x))$. This formulation enables us to employ a fast coordinate ascent optimization algorithm to approximate the posterior distributions of the effect sizes. Related variational approaches have been explored in the context of fine mapping[51,52] and polygenic modeling[34,55,57] but differ in both conceptual and technical details—we discuss these similarities and differences in the supplemental methods section S2.7.

In addition, to perform inference given some of the unknown fixed parameters, i.e., $\theta = (\pi, \sigma_\beta^2, \sigma_\epsilon^2)$, the default formulation of the VIPRS model uses the variational expectation maximization (VEM) algorithm,[50,51,65] where in an alternating fashion, in the E-step we update the variational parameters given the hyperparameters and in the M-step we update the hyperparameters of the model. In both the E- and M-steps, we update the free parameters of the model to maximize our objective, the ELBO. In supplemental methods section S2.3, we show that updating the hyperparameters to maximize the ELBO also results in closed-form solutions for those parameters.

Once the VEM optimization procedure converges to a local maximum, we output the posterior mean for the effect sizes and use this quantity to weigh the contribution of each genetic marker to the polygenic score of a given sample. Concretely, under the spike-and-slab prior outlined above, the posterior mean for effect size of variant j is given by $\eta_j := \gamma_j\mu_j$. Thus, the polygenic score for a new test sample i can be defined as $\widehat{PRS}_i = \sum_j X_{ij}\eta_j$. For

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

case-control traits, this polygenic score approximates the latent continuous liability underlying the disease and can be thresholded or used as is along with other clinical and demographic data for downstream clinical applications.[20–23]

Despite the conceptual simplicity of the framework described above, fitting such a model to biobank-scale data presents several computational challenges. For instance, the closed-form update equations for some of the variational parameters involve terms that relate to the LD between the focal variant and all other variants in the genome. This can be computationally prohibitive to compute for millions of variants and for hundreds of EM iterations. To overcome this, we follow the lead of other summary-statistics-based PRS methods and use a banded or shrunk LD matrix,[4,6,28,32] which results in substantial improvements in speed without substantially degrading predictive performance.

Hyperparameter tuning strategies

The standard VIPRS model employs a VEM framework to infer the hyperparameters $\boldsymbol{\theta} = (\pi, \sigma_\beta^2, \sigma_\epsilon^2)$, where in the M-step, we update each hyperparameter to maximize the surrogate objective, i.e., the ELBO.[51,65] This strategy works well in many settings, but it is prone to entrapment in local optima,[65] which may degrade overall predictive performance of the model. In this work, we explored three alternative strategies for tuning the hyperparameters of the model.

VIPRS-GS

In the first strategy, we performed grid search (GS) over the hyperparameters of the model, selecting the values that result in the best predictive performance on a held-out validation set.[27,29,32] We also explored a pseudo-validation variant of the model (supplemental methods) and showed that it results in almost identical prediction accuracy (Figures S6 and S7). The grid search was performed specifically over the proportion of causal variants π, with the remaining parameters updated according to their approximate maximum likelihood estimates. The grid for π ranged from $\frac{1}{M}$ to $\frac{M-1}{M}$ with 30 equidistant values on a $\log_{10}$ scale, where M is the number of variants included in the model.

VIPRS-BO

To search over the hyperparameter space without the constraint of a predefined and discrete grid, we experimented with a second hyperparameter tuning technique known as Bayesian optimization (BO).[66] In BO, we assume that there is an underlying unknown function $f(\boldsymbol{\theta})$ that takes the hyperparameters as input and outputs a certain score that we wish to optimize, such as the training ELBO or the validation R^2. This unknown function is modeled with a Gaussian Process (GP) prior, which allows us to explore the parameter space efficiently while accounting for uncertainty in a principled manner. The other component in this framework is the acquisition function, a heuristic that maps from the GP posterior to information about the most promising regions in hyperparameter space.[66,67] In our experiments, we used the `scikit-optimize` python package to perform this optimization, with `gp_hedge` as the default acquisition function. The optimizer was allowed to sequentially evaluate up to 20 points in a bounded $1D$ space for the hyperparameter π.

VIPRS-BMA

In the third strategy, we used a Bayesian model-averaging (BMA) framework, where we use importance sampling to integrate out some of the hyperparameters of the model, as outlined in Carbonetto and Stephens.[51] The main idea here is that instead of fixing π to a particular value, we fit the `VIPRS` model along a grid of π values, as in `VIPRS-GS`, and then take a weighted average of the effect size estimates for each SNP on the basis of each model's ELBO.[50,51]

Similar to previous work in this area, we note that these three strategies can be deployed in conjunction with the VEM framework,[50,51,68] where some of the hyperparameters are updated with their approximate maximum likelihood estimates while the remaining parameters are optimized via the user's strategy of choice. This is important in practice because, with the three hyperparameters of the model, an exhaustive search will require searching over a three-dimensional grid, which can be computationally expensive. Therefore, in our experiments and analyses, for all of the three strategies that we explored, we only implemented a search over the fraction of causal variants π and estimated the other two hyperparameters by using the closed-form updates in the M-step.

Data preprocessing

To assess the performance of `VIPRS` on a biobank-scale dataset, we made use of the UK Biobank (UKB), a large database of genomic and phenotypic measurements from 488,377 participants from the United Kingdom.[1] In its latest release, the UKB database has genotype information from 488,377 individuals, from which, after applying standard quality control procedures, we retained data for a total of 337,205 samples. We accessed the UKB data under IRB Study Number A10-M48-19B. Briefly, the sample quality controls involved selecting unrelated individuals with White British ancestry, defined by the UKB on the basis of self-reported ethnic background as well as principal-component analysis (PCA) of the GRM, and who were also included in the PCA and phasing procedures outlined in Bycroft et al.[1] We restricted our main analyses to the White British cohort in order to maximize power to detect causal effects while reducing confounding. In addition this, we filtered data for individuals with detected sex chromosome aneuploidy, excess relatedness, or missing genotype rate exceeding 5% from this analysis.

The genetic variants or SNPs included in the study were selected on the basis of a number of quality control filters applied at various stages in the analysis. For the base dataset, we excluded variants with duplicate `rsIDs`, ambiguous strand, imputation quality score < 0.3, Hardy-Weinberg equilibrium p value $< 10^{-10}$, or genotype missingness rate > 0.05. We also removed multi-allelic variants as well as SNPs in long-range LD regions, as specified in supplemental table 13 of Bycroft et al. (2018).[1] In the GWAS analyses or LD matrix construction, we further filtered variants with minor allele count (MAC) < 5 or minor allele frequency (MAF) $< 0.1\%$. This resulted in a total of 9,590,026 bi-allelic variants that were used in the expanded SNP set analyses. Finally, following standard practice in PRS methodologies,[6,32] for the base analyses with the `VIPRS` model, we restricted to the set of variants in the HapMap3 reference panel,[36] resulting in a total of 1,093,308 SNPs. Most of these quality control procedures were carried out with the genetic analysis software tool `plink2`.[69]

Construction and efficient representation of LD matrices

An important quantity in the model is the LD or SNP-by-SNP correlation matrix $\mathbf{R}$. The matrix, or its columns, show up mainly in the update equations for the variational parameters μ_j of each SNP

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

j (Equation 10), the estimate of the residual variance σ_ϵ^2, as well as in the objective function (i.e., ELBO) (supplemental methods). In our model derivation, the LD matrix is assumed to be estimated in-sample from the same GWAS cohort. In practice, this information is generally not publicly available and working with dense LD matrices can be computationally inefficient. To get around these difficulties, we experimented with approximate and sparse LD estimators that were previously explored for in-sample or out-of-sample settings.[4,6,27] Our software supports a number of these approximate LD matrix estimators, including sample, block, shrinkage, and windowed estimators.

Sample estimator

In the sample estimator, we estimate the sample Pearson correlation coefficient between all SNPs on the same chromosome, which results in a dense matrix. For larger chromosomes and SNP sets, it is impractical to load dense matrices of this scale to memory. To handle data at that scale, we use compressed and chunked on-disk storage with `Zarr` arrays in `python` for fast, multi-threaded read and write access. Then, as we iterate through SNPs in the E-step, we load the matrix into memory one chunk at a time, thus allowing us to train `VIPRS` with extremely large LD matrices. In supplemental methods, we describe a procedure that allows us to load the LD matrix only once per iteration, resulting in improved speed and efficiency.

Block estimator

In the block LD estimator, we only estimate the sample LD between SNPs that are within the same LD block, as defined by, e.g., LDetect.[70] This is similar to what is done in the `Lassosum` and `PRScs` frameworks.[27,28]

Shrinkage estimator

In the shrinkage estimator, we shrink and threshold the entries of the sample LD matrix according to procedure outlined by Lloyd-Jones et al.[6] and Wen and Stephens[71] and implemented in the `gctb` software. Briefly, for the shrinkage estimator, we shrink each element of the LD matrix by a quantity proportional to the distance between pairs of variants j and k in along the chromosome: $\widehat{R}_{jk} = R_{jk} \cdot e^{c \cdot d(j,k)}$. In this context, $d(j,k)$ is the distance in centimorgans (cM) between variants j and k and the constant c is related to sample size used to infer the genetic map as well as effective population size.[6,71]

Windowed estimator

For the windowed LD estimator, we only consider the correlation between a focal variant with variants that are at most 3 cM away from it along the chromosome.[32,63] This estimator results in compact and banded LD matrices that can easily fit in memory on modern compute nodes.

To construct LD matrices for the main analyses of this paper, we selected a random subset of 50,000 individuals from the White British cohort described above. Within that group of individuals, we filtered SNPs with MAC $<$ 5 or MAF $<$ 0.1%, and again restricted to variants in the HapMap3 reference panel. For the analyses with the expanded set of variants, we only removed the HapMap3 filter. Unless explicitly stated otherwise, the analyses with the `VIPRS` model employed the windowed estimator for LD, with the distance cutoff set to 3 cM. The matrices are stored in compressed `Zarr` array format and are publicly available for download (see web resources).

Simulation study

To assess the predictive performance of `VIPRS` on large-scale datasets and for varying genetic architectures, we conducted a GWAS by using the pre-processed genotype data from the UK Biobank cohort. We simulated quantitative and binary traits according to six different genetic architectures and three settings for the additive genetic variance, $h^2_{SNP} = \{0.1, 0.3, 0.5\}$, for a total of 18 simulation configurations for each trait category (continuous and case-control). For the first three genetic architectures, we simulated the effect size for each variant according to the generative model outlined previously in Equation 3, with three settings for the proportion of causal variants, $\pi = \{10^{-4}, 10^{-3}, 10^{-2}\}$. The next two simulation scenarios involved sampling the effect size for each variant from a scale mixture of Gaussians (supplemental methods), $p(\beta_j) = \sum_{k=1}^{4} \pi_k \mathcal{N}(\beta_j; 0, d_k \sigma_\beta^2)$, with the mixing proportions set to $\pi = \{0.95, 0.02, 0.02, 0.01\}$. The variance multipliers d_k were set to $d = \{0.0, 0.01, 0.1, 1\}$ for the sparse mixture model and $d = \{0.001, 0.01, 0.1, 1\}$ for the infinitesimal mixture model. Finally, the last genetic architecture tested was the standard infinitesimal model, with the effect size drawn from a zero-centered Gaussian density $p(\beta_j) = \mathcal{N}(\beta_j; 0, \sigma_\beta^2)$. For each configuration, we generated ten independent phenotypes for a total of 180 simulated traits. For the binary traits, we followed the same procedure but used the liability threshold model[58] to obtain case-control status, with prevalence set to 15%.

After we generated simulated phenotypes for all individuals in the study ($N = 337,205$), we excluded the 50,000 samples used to generate the LD matrices and randomly split the remaining samples into 70% training ($N = 201,043$), 15% validation, and 15% testing ($N = 43,081$ each). We then used the genotype and simulated phenotype data of the training samples to generate GWAS summary statistics with `plink2`.[69]

Application to real traits from the UKB

To assess the predictive performance of `VIPRS` on real phenotypes, we extracted phenotypic measurements for nine quantitative and three case-control traits for the UKB cohort described previously. The quantitative phenotypes included log-transformed waist circumference (WC), log-transformed hip circumference (HC), standing height (HEIGHT), birth weight (BW), log-transformed body mass index (BMI), log-transformed high-density lipoprotein (HDL), low-density lipoprotein (LDL), forced vital capacity (FVC), and forced expiratory volume in the first second (FEV1). For each trait, we excluded samples with outlier or extreme values for the trait. For the remaining samples, within each sex separately, we corrected for age and the top ten principal components (PCs) of the genetic relationship matrix (GRM) and then applied a rank-based inverse normal transform (RINT) on the residuals.[72] To assess the predictive performance on held-out test sets, we performed 5-fold cross-validation. For each split, we further split the training data into 90% training and 10% validation to facilitate running PRS methods that require a validation set to tune their hyperparameters.

The case-control phenotypes included in the analysis are asthma (prevalence 12.7%), type 2 diabetes (T2D) (prevalence 2.3%), and rheumatoid arthritis (RA) (prevalence 1.7%). To assess the predictive performance on held-out test sets, we performed stratified 5-fold cross-validation, followed by splitting the training data into 90% training and 10% validation in a stratified manner to keep the prevalence approximately the same for all subsets of the data.

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

Table 1. The list of real phenotypes and GWAS data sources analyzed in this study

Phenotype	Description	GWAS source	GWAS sample size	Validation sample size	Test sample size
HEIGHT	standing height	UKB	242,213	26,913	67,282
HDL	high-density lipoprotein	UKB	211,856	23,540	58,849
BMI	body mass index	UKB	241,959	26,885	67,211
FVC	forced vital capacity	UKB	221,249	24,584	61,459
FEV1	forced expiratory volume in 1 s	UKB	221,265	24,586	61,463
HC	hip circumference	UKB	242,311	26,924	67,309
WC	waist circumference	UKB	242,340	26,927	67,317
LDL	low-density lipoprotein	UKB	230,995	25,667	64,166
BW	birth weight	UKB	138,300	15,367	38,417
T2D	type 2 diabetes	UKB	235,937	26,216	65,538
RA	rheumatoid arthritis	UKB	186,239	20,694	51,734
ASTHMA	asthma	UKB	229,031	25,448	63,620
LangoAllen2010_HEIGHT	standing height	Allen et al.[74]	131,547	26,913	67,282
Speliotes2010_BMI	body mass index	Speliotes et al.[75]	122,033	26,885	67,211
GLGC2021_HDL	high-density lipoprotein cholesterol	Graham et al.[76]	888,227	23,540	58,849
GLGC2021_LDL	low-density lipoprotein cholesterol	Graham et al.[76]	842,660	25,667	64,166
Teslovich2010_HDL	high-density lipoprotein cholesterol	Teslovich et al.[77]	97,749	23,540	58,849
Teslovich2010_LDL	low-density lipoprotein cholesterol	Teslovich et al.[77]	93,354	25,667	64,166
SpiroMeta2019_FVC	forced vital capacity	Shrine et al.[78]	79,005	24,584	61,459
SpiroMeta2019_FEV1	forced expiratory volume in 1 s	Shrine et al.[78]	79,005	24,586	61,463
Morris2012_T2D	type 2 diabetes	Morris et al.[79]	60,786	26,216	65,538
Scott2017_T2D	type 2 diabetes	Scott et al.[80]	159,208	26,216	65,538
Okada2014_RA	rheumatoid arthritis	Okada et al.[81]	37,681	20,694	51,734
Demenais2018_ASTHMA	asthma	Demenais et al.[82]	142,486	25,448	63,620

With each phenotype code, we provide the full phenotype name and description and the GWAS data source or cohort (UKB or external study) as well as the sample sizes for the training, validation, and test sets. The sample sizes for each subset may vary slightly across the five folds. For the external summary statistics, we prepended the phenotype codes with either the consortium name or the name of the first author as well as the year in which the GWAS was published. For analyses with the external GWAS summary statistics, the validation and test sets come from the UK Biobank.

The phenotypes and associated sample sizes in the UK Biobank are listed in Table 1. The detailed scripts with the extraction and transformation procedure for each phenotype are included in the public repository associated with this publication (web resources). The 5-fold cross-validation procedure was performed with the `scikit-learn` package in `python`.[73]

Validation in minority populations in the UKB

To validate the relative predictive ability of `VIPRS` in individuals of different backgrounds, we used the approach of Privé et al.[83] to identify subgroups of relatively uniform ancestry and ethnicity. Using self-reported ethnic background as well as PCA medoids from Privé et al.,[83] we extracted genotype data for individuals of Italian ($N = 6,177$), Indian ($N = 6,011$), Chinese ($N = 1,769$), and Nigerian ($N = 3,825$) ancestry. In genetic analyses, those ancestry groups show various levels of allele frequency differentiation (Fst) when compared to the White British cohort.[83] The samples were selected after applying the same quality control filters as before. Mainly, we retained individuals who were used in the PCA and phasing procedures and filtered samples with detected sex chromosome aneuploidy or excess relatedness from this analysis.

For each individual in those target populations, we extracted phenotype data for the traits analyzed previously (Table 1). Then, we used effect size estimates derived from the 5-fold analyses on the White British cohort to generate polygenic scores for individuals in those minority populations. Given these polygenic score estimates, we computed the relative prediction R^2 as the incremental R^2 in the target population divided by the R^2 of the best performing PRS model on the test set in the White British cohort. This metric is designed to highlight the transferability of PRS estimates across different population and ancestry groups.

PRS method comparisons and specifications

To compare the predictive performance of VIPRS to state-of-the-art methods for polygenic risk prediction with summary statistics, we included a diverse collection of methods with different assumptions and implementations, including three stochastic Bayesian

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

methods SBayesR (gctb 2.03),[6] PRScs,[28] and LDPred2 (bigsnpr 1.9.11);[32] a variational Bayesian PRS method MegaPRS (LDAK 5.2);[34] a penalized regression method (Lassosum 0.4.5[27]); and finally, as a simple baseline, we included a C+T method (PRSice2 2.3.5[29]). In addition to being widely used in practice, most of these methods were selected because they have been shown to be competitively accurate in recent comprehensive surveys.[39,40,84]

For the main analyses presented in the text, our VIPRS method used the windowed LD estimator with 3 cM distance cutoff as well as the spike-and-slab prior family for the effect size. The hyperparameters of this model, $\theta = \{\pi, \sigma_\epsilon^2, \sigma_\beta^2\}$, were updated in the M-step of the VEM algorithm. For VIPRS-GS, we performed grid search over the proportion of causal variants π, where the grid spanned 30 points from $\frac{1}{M}$ to $\frac{M-1}{M}$ on a $\log_{10}$ scale, where M is the number of genetic variants included in the model. Out of these 30 models, we selected the one that maximized the prediction accuracy in a held-out validation set. The remaining two hyperparameters, the residual variance σ_ϵ^2 and prior variance σ_β^2, were updated in the M-step via closed-form solutions as before.

For each external method, we provided the GWAS summary statistics for the simulated and real phenotypes and ran the model with default or recommended settings. Specifically, for SBayesR, we ran the MCMC chain for 10,000 iterations, with the first 2,000 taken as burn-in, and specified the default four component Gaussian mixture prior, with mixing proportions $\pi = \{0.95, 0.02, 0.02, 0.01\}$ and corresponding γ parameters set to $\gamma = \{0.0, 0.01, 0.1, 1\}$.[6] For PRScs, we used the PRScs-auto variant of the model, in which the φ hyperparameter is inferred automatically in a fully Bayesian fashion.[28] For the LDPred2 model, we ran the three variations of the method (LDPred2-inf, LDPred2-grid, and LDPred2-auto) and reported the performance for the grid model because it performed the best on average for the simulations and real traits. For the LDPred2-grid and LDPred2-auto models, we used the sparse model setting and the recommended grid over the two hyperparameters: (1) the SNP heritability $h^2 \in \{0.7h^2_{LDSC}, h^2_{LDSC}, 1.4h^2_{LDSC}\}$, where h^2_{LDSC} is the LD score regression SNP heritability estimate[32,63] and (2) the proportion of causal variants π ranging along 21 points on a log-scale from $\pi = 1e-5$ to $\pi = 1$. For MegaPRS, we used the BayesR variant of the model with the genome-wide complex trait analysis (GCTA) heritability model used to infer the prior variance for the effect size of each genetic marker.[34] As per the recommendations of the authors of that method, the inference in MegaPRS was performed within overlapping windows of 1 cM. For Lassosum, we used the default grid for the LASSO penalty parameter λ, which covers a 20-point grid on a log-scale from 0.001 to 0.1. Finally, for PRSice2, we used the default clumping and thresholding parameters.[29] For LD clumping, we used a window size of 250 kb and r^2 values greater than 0.1. For p value thresholding, we used a grid from 5e−8 to 0.5 with step size of 5e−5. For most of these external methods (SBayesR, LDPred2, PRScs, and Lassosum) we used pre-computed, publicly available LD matrices published by the authors and maintainers of each software. All of these LD matrices were pre-computed from large random samples in the UKB.[6,27,28,32] The LD matrices used by each method employed different LD estimators (e.g., block, windowed, or shrinkage) as well as different sample sizes within the UKB. For MegaPRS and PRSice2, we computed the LD matrices and associated files by using recommended or default parameters.

Some of the PRS methods included in our analyses require a held-out validation set to tune some of their hyperparameters while others only use data from the training set (i.e., the GWAS summary statistics). The former category of methods includes VIPRS-GS, Lassosum, LDPred2-grid, MegaPRS, and PRSice2 and the latter includes VIPRS, SBayesR, and PRScs. Thus, the methods that do not use cross-validation were effectively trained with a smaller subset of the data. In principle, the hyperparameters of PRScs and SBayesR can be tuned via cross-validation, but since this can be computationally expensive, we only include either the automated or default version of these two methods. Once the models converge, we output the effect size estimates and then generated polygenic scores for the samples in the test set. Given these polygenic scores, the models were then evaluated for the quality of their predictions. For quantitative traits, we reported the incremental prediction R^2, defined as the R^2 of a linear model with the PRS and covariates (age, sex, and top ten PCs) minus the R^2 obtained from a linear model with the covariates alone. For case-control phenotypes, we reported the area under the precision-recall curve (AUPRC) between the polygenic score and the binary phenotype.

In addition to these prediction metrics, we also compared the run-time (wall-clock time) of the different methods to gauge their scalability and computational efficiency. In all the experiments and analyses, each method was allocated eight cores and 16 GB of memory, and thus PRS methods that support parallel processing will have shorter wall-clock run-time but may have higher CPU utilization. The LDPred2 model was assigned an entire compute node (40 cores and >200 GB of memory) because we found that the sparse on-disk LD matrices were not working optimally in a shared computing environment. However, for fair comparisons, the method was restricted to using only eight cores.

The detailed specification of priors, grid values, hyperparameters, and computational resources for each PRS method is shown in the repository accompanying this manuscript (see web resources).

VIPRS software implementation

The data structures and inference algorithms for the VIPRS model are implemented in two python packages that are open source and publicly available on GitHub (see web resources). The first software package, magenpy, implements scripts and routines for computing LD matrices and transforming them to Zarr array format, simulating complex traits from genotype data and harmonizing multiple genetic data sources, such as GWAS summary statistics, LD reference panels, functional annotations, etc. The second software package, viprs, implements the optimized VI algorithms to obtain posterior estimates for the effect sizes. For optimal speed and efficiency, the coordinate ascent routine is written in cython, a compiled programming language that produces python-compatible modules with minimal overhead.[85] Both software packages follow object-oriented design principles to allow for streamlined user extensions and experimentation by experienced developers. We also provide runner scripts that allow users to perform inference with commandline interfaces.

Results

Genome-wide simulation results

To examine the predictive performance of the VIPRS model compared to existing PRS methods, we simulated quantitative and case-control traits with varying genetic

architectures and heritability values. To align our simulations with the real trait analyses in terms of cohort size and composition, we used genotype data for a subset of $\approx 340,000$ unrelated White British individuals from the UKB (material and methods) and a HapMap3 subset of $\approx$ 1.1 million genotyped and imputed SNPs. The simulations followed the generative models outlined in the material and methods section and supplemental methods, with the effect size of each variant drawn from different architectures and residuals for each individual sampled from an isotropic Gaussian density. For binary traits, we simulated case-control status following the liability threshold model,[58] with the prevalence set to 15%.

The simulations spanned six different genetic architectures, including both sparse and infinitesimal scenarios. In the first three scenarios, we simulated under the spike-and-slab model (material and methods), where the effect size for a given variant j was drawn from $p(\beta_j) = \pi\mathcal{N}(\beta_j; 0, \sigma_\beta^2) + (1 - \pi)\delta_0$, and the proportion of SNPs contribute to the variation in the trait ranging along a pre-specified grid $\pi \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. In the next two simulation scenarios, the effect size is drawn from a scale mixture of Gaussians, $p(\beta_j) = \sum_{k=1}^{4} \pi_k \mathcal{N}(\beta_j; 0, d_k\sigma_\beta^2)$, with the mixing proportions set to $\pi = \{0.95, 0.02, 0.02, 0.01\}$. The variance multipliers d_k were set to $d = \{0.0, 0.01, 0.1, 1\}$ for the sparse mixture model and $d = \{0.001, 0.01, 0.1, 1\}$ for the infinitesimal mixture model. Finally, for the infinitesimal model, we assumed that the effect size for all variants is drawn from a zero-centered Gaussian density, $p(\beta_j) = \mathcal{N}(\beta_j; 0, \sigma_\beta^2)$. For each genetic architecture, we varied the proportion of additive genetic variance captured by all causal SNPs, $h^2_{SNP} \in \{0.1, 0.3, 0.5\}$, such that the simulated traits range from the mildly to the highly heritable. For each unique configuration, we simulated ten independent phenotypes, for a total of 180 traits for each class (binary and continuous). Once the traits were simulated for all individuals in the dataset, we randomly split the sample into 70% training, 15% validation, and 15% testing; we used the training set to generate GWAS summary statistics.

Next, we fit the `VIPRS` model to the summary statistics from the training data, along with other commonly used PRS methods. Given the many existing PRS approaches, we selected for comparison five methods that performed favorably in recent comprehensive surveys,[39,40,84] namely `SBayesR`,[6] `LDPred2`,[32] `PRScs`,[28] `MegaPRS`,[34] and `Lassosum`.[27] The first three methods use the Bayesian framework outlined above for approximate posterior inference, all employing a Gibbs sampling algorithm. They are mainly distinguished by the families of prior density they assign to the effect sizes, among many other algorithmic choices. The fourth model, `MegaPRS`, is similar to `VIPRS` in that it uses VI for polygenic risk estimation, though there are some fundamental differences in the details of the optimization algorithm (see supplemental methods). `Lassosum` is a penalized regression method that derived and implemented the LASSO estimator for PRS inference from GWAS summary statistics.[27] Finally, we included `PRSice2`,[29] which implements clumping and thresholding, a commonly used baseline method. After fitting each method on the summary statistics from the training data, we used the effect size estimates to generate polygenic scores for individuals in the held-out test set and evaluated their predictive performance. For quantitative traits, we computed the incremental prediction R^2 for each model, while for binary traits we show the AUPRC, a preferable metric in the presence of class imbalance.[86] In addition to the six external PRS models, we also examined the predictive performance of the basic `VIPRS` model trained with the VEM framework (material and methods) as well as a version of the `VIPRS` model, dubbed `VIPRS-GS`, in which we perform grid search and tune the hyperparameters on the basis of predictive performance on a held-out validation set.

The predictive performance results for this simulation study are summarized in Figures 1 and S1, which show that `VIPRS-GS` outperforms or is on-par with state-of-the-art PRS methods in most of the scenarios tested. In particular, our analyses indicate that `VIPRS` provides the most benefit for more sparse architectures and highly heritable traits (leftmost panels in Figures 1A and S1A). Notably, in this particular setting, `VIPRS` is able to capture most of the additive genetic variance (as measured by the R^2 metric, which is upper-bounded by the heritability), while other Bayesian and non-Bayesian methods often lag behind. For infinitesimal and mixture-based architectures, `VIPRS` shows competitive predictive ability across the range, only lagging slightly behind `SBayesR` in those settings. For highly polygenic traits with the proportion of causal variants equal or greater than 1%, all models conferred lower prediction accuracy relative to the heritability values that we simulated with. This is because, under our simplified simulation scenario, the larger the number of causal variants, the smaller the effect size per SNP. Consequently, this makes it more difficult for PRS methods to pick up the true causal signals, at least given the training sample sizes available. Nonetheless, the `VIPRS` models conferred higher predictive performance relative to most competing methods in many of those scenarios. This pattern holds for both quantitative (Figure 1) as well as binary case-control phenotypes (Figure S1). This improvement in prediction accuracy comes also with improved computational efficiency, with the run-time of the standard `VIPRS` model rivaling other heuristic and deterministic methods, such as `Lassosum`, `MegaPRS`, and `PRSice2` (Figure 2).

The prediction accuracy of PRS methods on simulated phenotypes may be over-optimistic as a result of the similarity between the generative process for the simulations and their model assumptions. Additionally, our simulations assume that all the causal SNPs are genotyped or imputed and thus present in the dataset, which is certainly not the case for real traits. Therefore, it is important to systematically evaluate these methods on real phenotypes as shown next.

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

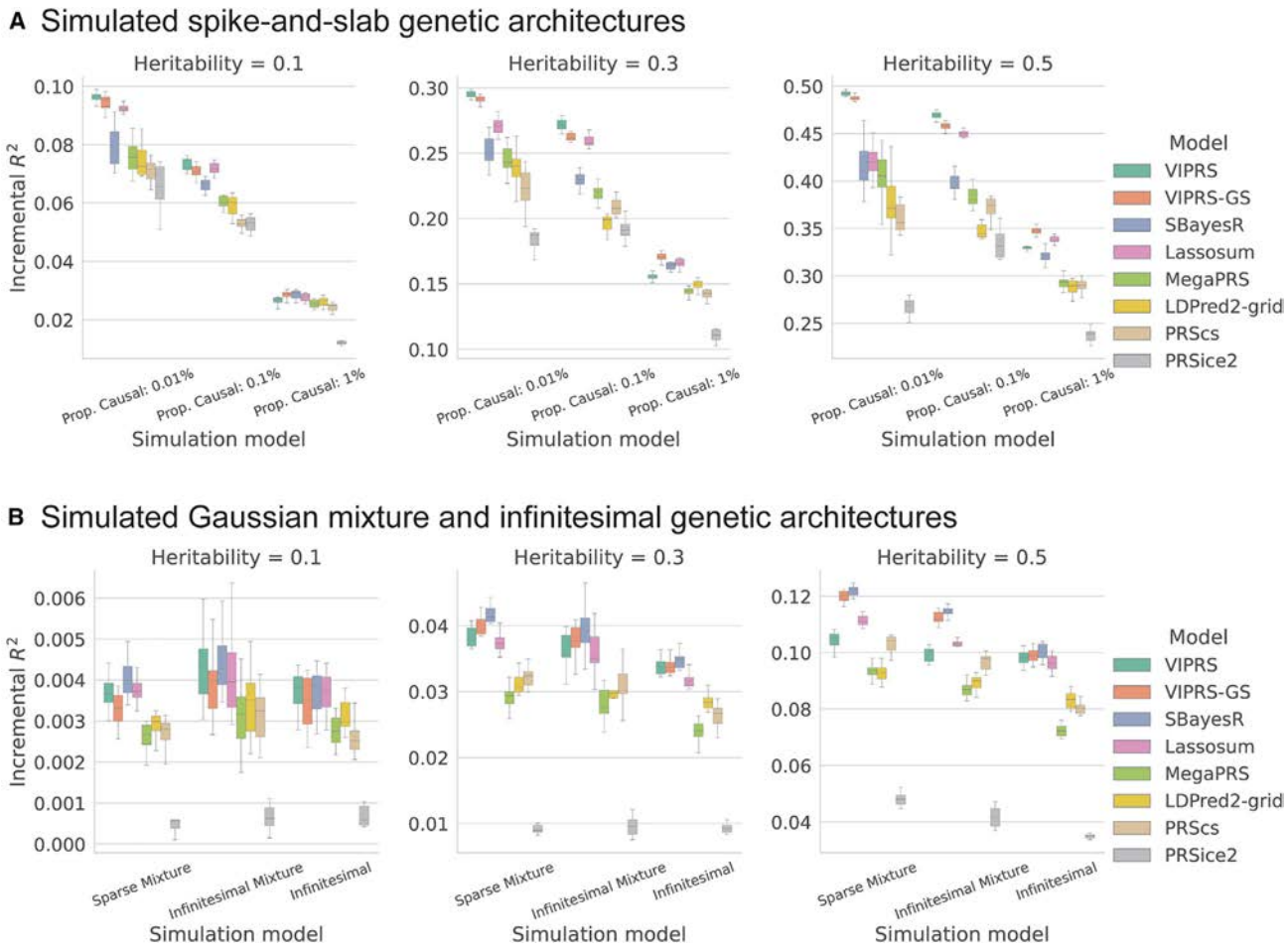


Figure 1. Predictive performance of summary statistics-based PRS methods on simulated quantitative traits following spike-and-slab and Gaussian mixture or infinitesimal genetic architectures

The phenotypes were simulated with real genotype data from the White British cohort in the UK Biobank ($N = 337,225$), leveraging a subset of 1.1 million HapMap3 variants. The simulation scenarios encompass a total of 18 configurations, spanning six genetic architectures and three values for SNP heritability. (A) shows prediction accuracy for traits simulated under the spike-and-slab model and (B) shows predictive performance for traits simulated under Gaussian mixture and infinitesimal genetic architectures. For each configuration, we simulated ten independent phenotypes. Each panel shows results for phenotypes simulated with the pre-specified SNP heritability and each column within a panel shows performance metrics for phenotypes simulated with a pre-specified genetic architecture. The performance metric shown is the incremental prediction R^2. The boxplot for each method and simulation configuration shows the quartiles of the R^2 scores for the ten simulated phenotypes. The PRS methods shown are our proposed `VIPRS`and `VIPRS-GS` (using grid search to tune model hyperparameters) as well as six other baseline models: `SBayesR`, `Lassosum`, `MegaPRS`, `LDPred2` (grid), `PRScs`, and `PRSice2` (C+T).

Application on real phenotypes in the UK Biobank

Given its competitive performance on simulated traits, we next sought to assess the relative predictive ability of the `VIPRS` model on real phenotypes measured for a subset of $\approx 340,000$ unrelated White British individuals in the UKB. This focus on a large sub-cohort of relatively uniform ancestry helps us achieve sufficient power while reducing confounding due to population structure. A downside of this approach is that it is expected to yield PRS estimates that perform more poorly for individuals of other ancestries,[83,87] a limitation that we examine in more detail in the next section.

For this analysis, we extracted and processed phenotypic measurements for nine quantitative traits and three binary traits that are commonly used to benchmark PRS methods (Table 1). The traits considered have varying (inferred) genetic architectures and SNP heritabilities. To make full use of the data, we followed a 5-fold cross-validation study design, where in each iteration, 80% of the samples with trait measurements were used to generate the GWAS summary statistics and training the PRS models and 20% were used as an independent test set.

Our experiments show that, across a variety of different phenotypes, `VIPRS` is competitive with commonly used Bayesian PRS methods (Figure 3). Within the category of Bayesian PRS methods, the predictive performance of `VIPRS` is especially distinguished for anthropometric and blood lipid traits (Figure 3A). For instance, when compared to the `LDPred2` model, which imposes the same spike-and-slab prior on the effect sizes, `VIPRS` shows an average of 4.6% improvement in prediction R^2 on continuous traits. However, in many cases the basic `VIPRS` model

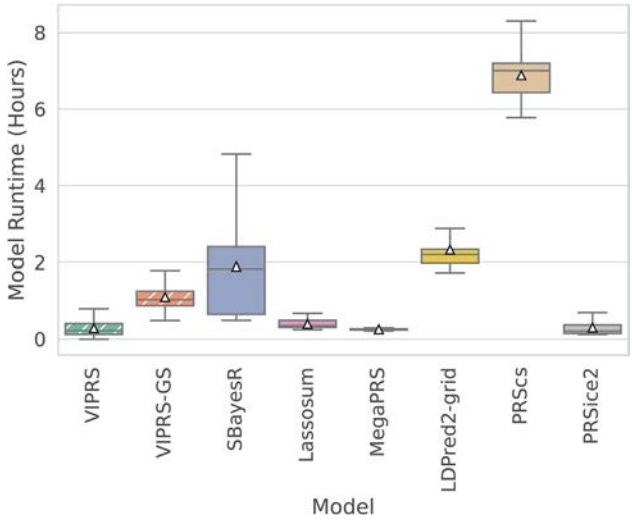


Figure 2. The total runtime (in hours) of the summary statistics-based PRS methods included in the study

The boxplot for each method shows the quartiles of the runtime from a total of 420 independent experiments (360 simulated traits plus 60 experiments on real measured traits, comprising the 12 phenotypes analyzed multiplied by the five training folds). The white triangles indicate the mean runtime for each method. The PRS methods shown are our proposed `VIPRS` and `VIPRS-GS` (using grid search to tune model hyperparameters) as well as six other baseline models: `SBayesR`, `Lassosum`, `MegaPRS`, `LDPred2` (grid), `PRScs`, and `PRSice2` (C+T). Dashed lines highlight the models contributed in this work.

lags behind the `SBayesR`[6] and `Lassosum`[27] models (Figure 3). In addition to the difference in posterior approximation strategy (VI versus Gibbs sampling), the `SBayesR` model differs from the `VIPRS` model in three other important respects: (1) the prior on the effect size, (2) the estimator for the linkage disequilibrium (LD) between variants, and finally (3) the approach for estimating the hyperparameters of the model. We sought to understand the effect of each of these modeling choices on the predictive performance of our model.

To address the first point, we derived and implemented a version of `VIPRS` called `VIPRSMix`, where we replaced the spike-and-slab prior on the effect sizes with a sparse Gaussian mixture prior with four mixture components (supplemental methods).[6,26,50] Our experiments show that the more expressive mixture prior improves the performance of the standard `VIPRS` model on some traits, especially highly heritable and polygenic traits such as standing height and HDL (Figure S5), with an average of 2.4% increase in prediction R^2 on continuous traits. However, the improvement is not consistent across all traits and use of this prior does not fully bridge the gap between `VIPRS` and `SBayesR`.

Secondly, we assessed the impact of the LD estimator on the predictive performance of the `VIPRS` model by refitting the model with three commonly used estimators for LD: windowed,[32,63] shrinkage,[6,71] and block[28,70] (material and methods). Our experiments indicate that, on many of the traits tested, using the shrinkage estimator for LD results in slight improvements in prediction accuracy, though as in the case of the windowed LD estimator, it still slightly lags behind the `SBayesR` model (Figures S2 and S3). Notably, however, the shrinkage estimator tends to be more robust when the sample size of the LD reference panel is small (Figures S2 and S3).

Finally, and most importantly, the basic `VIPRS` model differs from the `SBayesR` model in terms of its hyperparameter estimation strategy. Most PRS methods have global hyperparameters, such as the residual variance σ_ϵ^2 or proportion of causal variants π, that need to be estimated or fixed to reasonable values. `SBayesR` follows a fully Bayesian approach for learning the hyperparameters of the model, assigning them priors and inferring their posterior distributions.[6] By contrast, `VIPRS` follows a VEM framework where in the M-step we set the hyperparameters to their approximate maximum-likelihood estimates.[50,51] This latter strategy is known to be prone to overfitting or entrapment in local maxima.[50,51,65,88,89] As an alternative to the VEM framework, we tested three other strategies for tuning the hyperparameters of the model, including grid search,[90] Bayesian optimization,[66] and Bayesian model averaging[51] (see material and methods, Figures S6 and S7). In this context, similar to the `Lassosum`, `MegaPRS`, and `LDPred2` methods, we found that by setting some of the hyperparameters of the model via grid search with an independent validation set, `VIPRS-GS` provides a powerful remedy in most settings (Figures 3, S6, and S7), resulting in a balanced trade-off between computational speed and predictive accuracy (Figures 2 and 3). Indeed, our results show that the `VIPRS-GS` model conferred the highest or second highest predictive performance on all traits tested (Figure 3), consistently exceeding the performance of the VEM-based `VIPRS`. At the same time, the main drawback of the grid search approach is that, despite the parallel software implementation, it results in a significant slowdown compared to the VEM approach (Figure 2). In terms of predictive performance, the advantage of the grid search is most prominent for highly heritable traits, such as standing height and HDL (Figure 3A). For the other traits, `SBayesR` is on-par or only marginally better. This indicates that the gap in predictive performance between `SBayesR` and the basic `VIPRS` model is mostly due to differences in hyperparameter estimation strategy, i.e., fully Bayesian inference of hyperparameters versus VEM approach.

PRS validation in minority populations in the UK Biobank

When trained on GWAS data from a single source population, transferability of PRS estimates across populations is limited,[83,87] and the degradation in prediction accuracy increases with the increase in allele frequency differentiation (Fst) between populations.[83] At the same time, recent studies of cross-population genetic correlations have demonstrated strong correlations in the genetic

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

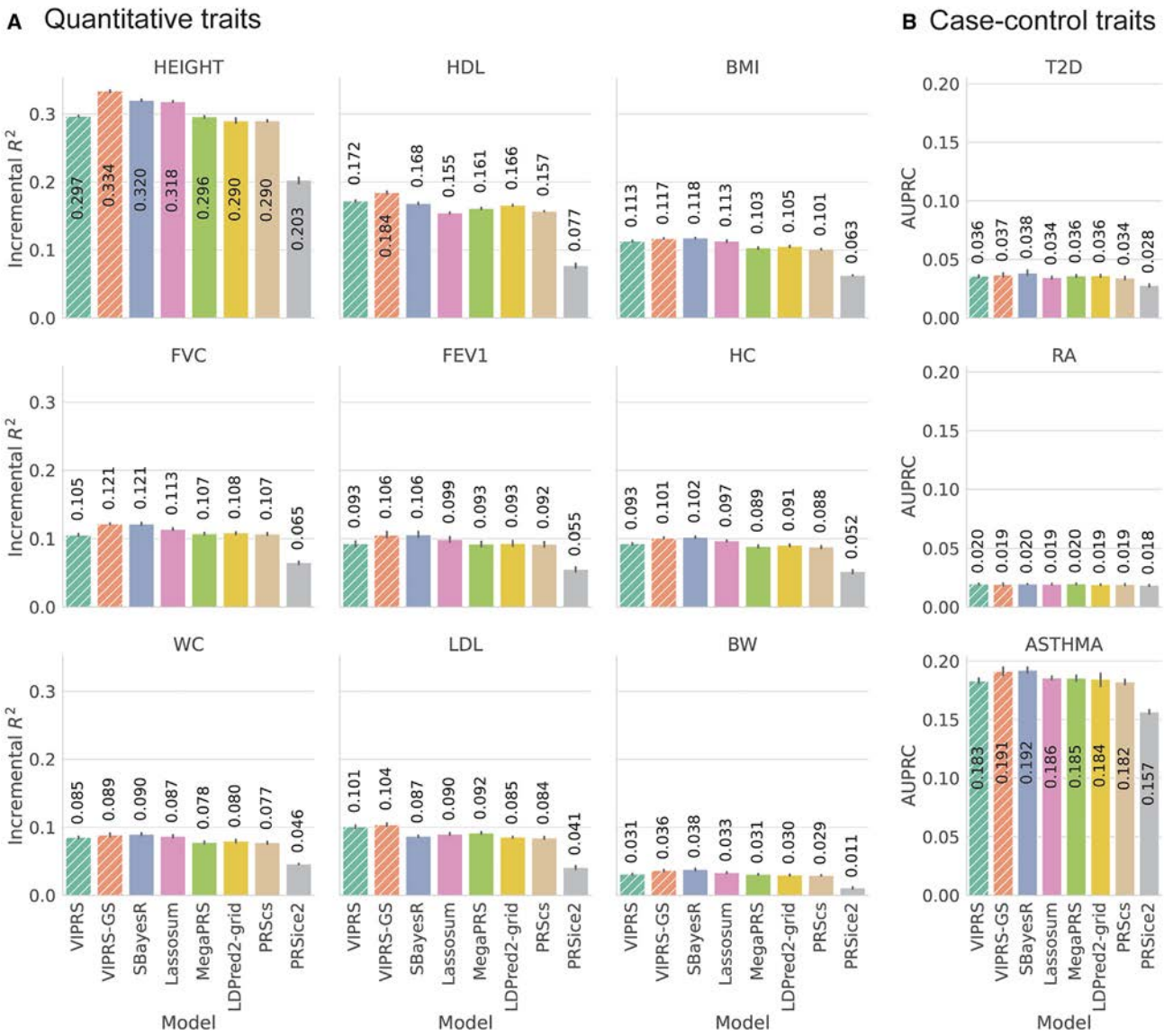


Figure 3. Predictive performance of summary statistics-based PRS methods on real quantitative and case-control phenotypes in the UK Biobank

(A and B) The measured phenotypes were pre-processed and analyzed in a 5-fold cross-validation study design and the prediction metrics show the performance of each PRS method in predicting the phenotype in a held-out test set. Each panel shows the predictive performance, in terms of (A) incremental R^2 and (B) area under the precision recall curve (AUPRC), of various PRS methods when applied to a given phenotype. The bars show the mean of the prediction metrics across the five folds and the black lines show the corresponding standard errors. The quantitative phenotypes analyzed are standing height (HEIGHT), high-density lipoprotein (HDL), body mass index (BMI), forced vital capacity (FVC), forced expiratory volume in 1 s (FEV1), hip circumference (HC), waist circumference (WC), low-density lipoprotein (LDL), and birth weight (BW). The binary phenotypes analyzed are asthma (ASTHMA), type 2 diabetes (T2D), and rheumatoid arthritis (RA). The PRS methods shown are our proposed `VIPRS` and `VIPRS-GS` (using grid search to tune model hyperparameters) as well as six other baseline models: `SBayesR`, `Lassosum`, `MegaPRS`, `LDPred2` (grid), `PRScs`, and `PRSice2` (C+T). Dashed lines highlight the models contributed in this work.

architectures of complex traits between various ancestry groups.[91,92] These correlations imply that PRS models that perform better in the source population will also tend to perform more favorably when applied to the target populations.

To assess this, we extracted genotype and phenotype data for individuals who self-identified as Italian ($N = 6,177$), Indian ($N = 6,011$), Chinese ($N = 1,769$), and Nigerian ($N = 3,825$). The self-reported ethnic backgrounds were further validated on the basis of the PCs of the GRM[83] (material and methods). Using the effect size estimates derived from training the PRS models on summary statistics from the White British cohort across the five training folds, we computed a PRS for each individual in the target population. Given the real phenotype measurements for these individuals, we evaluated the predictive performance by using relative incremental prediction R^2, where the R^2 in the target population was divided by the

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

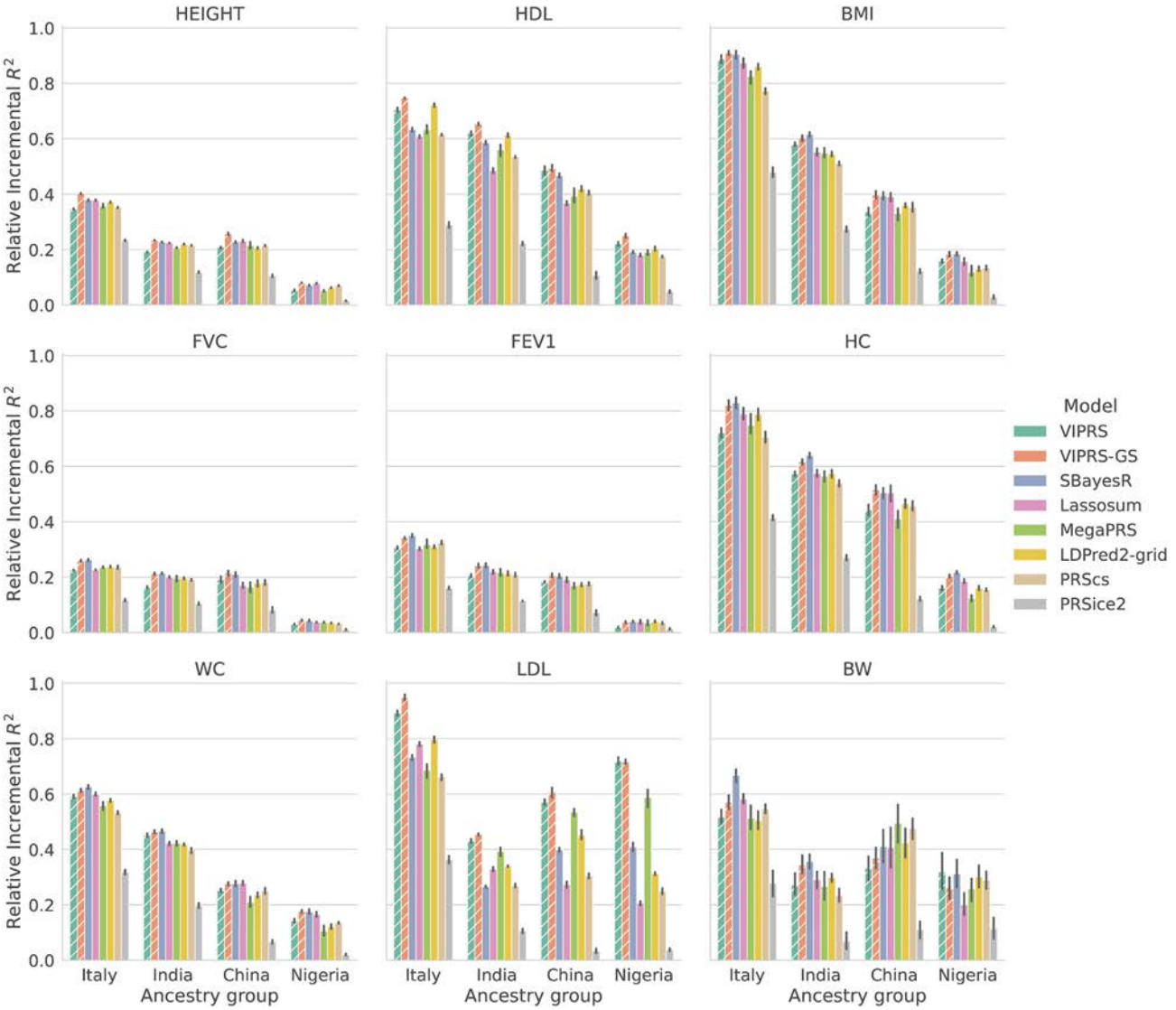


Figure 4. Relative predictive performance of summary statistics-based PRS methods on real quantitative phenotypes in minority populations in the UK Biobank

The PRS models were trained on summary statistics from the White British cohort in the UK Biobank using a 5-fold cross-validation design. Then, the effect size estimates from the five training folds were used to perform predictions in individuals of Italian, Indian, Chinese, and Nigerian ancestry. Each panel shows the incremental prediction R^2 in a given ancestry group relative to the prediction R^2 of the best performing model on the White British cohort. The bars show the mean of the relative prediction metric across the five training folds and the black lines show the corresponding standard errors. The quantitative phenotypes analyzed are standing height (HEIGHT), high-density lipoprotein (HDL), body mass index (BMI), forced vital capacity (FVC), forced expiratory volume in 1 s (FEV1), hip circumference (HC), waist circumference (WC), low-density lipoprotein (LDL), and birth weight (BW). The PRS methods shown are our proposed `VIPRS` and `VIPRS-GS` (using grid search to tune model hyperparameters) as well as six other baseline models: `SBayesR`, `Lassosum`, `MegaPRS`, `LDPred2` (grid), `PRScs`, and `PRSice2` (C+T). Dashed lines highlight the models contributed in this work.

R^2 of the best performing model on the test set in the White British cohort.

Our results confirm that for most of the traits analyzed, the models with the best predictive performance on the source population (White British) tend to transfer better to the target populations (Figure 4). Furthermore, consistent with other analyses in this space,[83,87] the drop in prediction accuracy generally tracks with the Euclidean distance between the White British and the target populations in PC space. Interestingly, deviations from this general pattern were observed for LDL and birth weight, which may be due to gene-by-environment interactions.[92] For LDL specifically, we observed strong differentiation in transferability between PRS methods, with models employing VI techniques, `VIPRS` and `MegaPRS`, attaining upwards of 1.5 times the prediction accuracy of the next competing PRS method in individuals of Nigerian and Chinese ancestry (Figure 4). We hypothesize that the high variance in accuracy and transferability for LDL is due, in part, to differences in effect size estimates for large effect *APOE* variants. For instance,

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

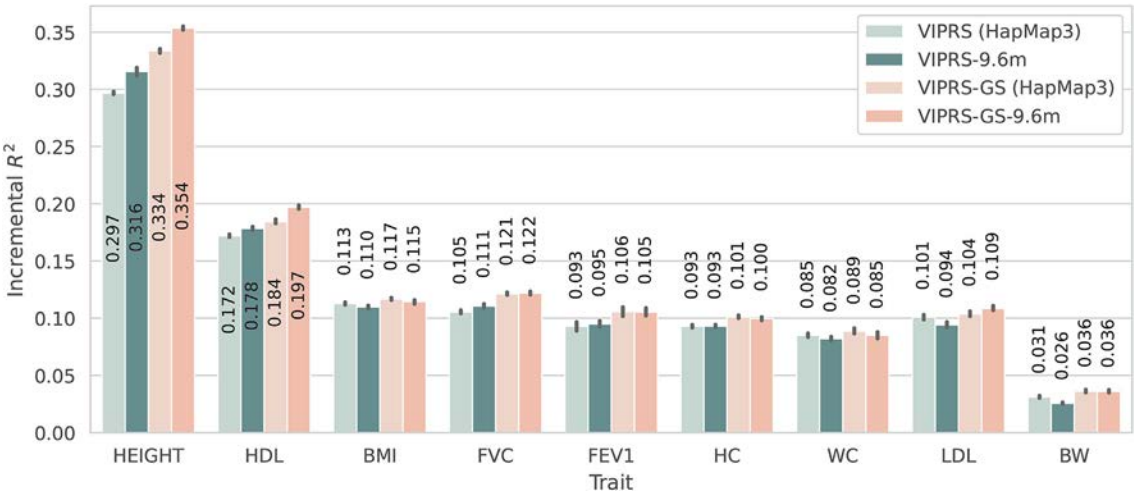


Figure 5. Comparing the predictive performance of the VIPRS method on real quantitative traits in the UK Biobank using the HapMap3 SNP set as well as an expanded set of 9.6 million genotyped and imputed variants

Each bar shows the predictive performance, in terms of incremental prediction R^2, of four different versions of the `VIPRS` model. From left to right, we have the standard `VIPRS` model trained on the HapMap3 subset (comprised of ≈1.1 million variants), `VIPRS-9.6m` is the `VIPRS` model trained on 9.6 million variants, `VIPRS-GS` is `VIPRS` with grid search trained on the HapMap3 subset, and finally `VIPRS-GS-9.6m` is the `VIPRS` with grid search trained on the 9.6 million SNPs. The black vertical lines show the standard errors across the five folds from the 5-fold cross-validation scheme. The quantitative phenotypes analyzed are standing height (HEIGHT), high-density lipoprotein (HDL), body mass index (BMI), forced vital capacity (FVC), forced expiratory volume in 1 s (FEV1), hip circumference (HC), waist circumference (WC), low-density lipoprotein (LDL), and birth weight (BW).

rs7412, a known major determinant of LDL,[76] is assigned a large effect by VIPRS, while other methods either excluded this variant *a priori* (e.g., `LDpred2`, `SBayesR`) or assigned it a small effect (e.g., `MegaPRS`, `PRScs`).

Scaling up `VIPRS` to 9.6 million variants

In recent years, with the advent of biobank-scale whole-genome sequencing efforts[3,43] and improved variant imputation pipelines,[35] there has been increasing interest in understanding the extent to which larger and larger sets of genetic variants enable us to better capture the genetic diversity underlying complex traits.[93] This is especially important in light of recent results that showed that a substantial proportion of the missing heritability is due to imperfect tagging of rare causal variants by common SNPs.[93] In this context, the main advantage of `VIPRS` is its speed and scalability (Figure 2); thus we wanted to understand the extent to which our method could benefit from modeling an expanded set of SNPs. Here, following recent efforts in this space,[34,40] we examine the predictive performance of `VIPRS` with approximately 9.6 million measured or imputed genetic variants, almost an order of magnitude greater than the HapMap3 subset analyzed previously. This includes all bi-allelic variants with MAF greater than 0.1% and MAC greater than 5 in the White British cohort in the UKB.

Following the same 5-fold cross-validation study design described earlier, we observed that modeling an expanded set of SNPs results in substantial improvements in prediction accuracy for highly heritable and polygenic traits, such as standing height and HDL, in comparison with the best performing PRS model using a subset of 1.1 million HapMap3 SNPs (Figure 3), which is consistent with previous studies[6] (Figures 5 and S5). Concretely, for standing height and HDL in particular, including almost an order of magnitude more variants resulted in 3%–7% relative improvement in the predictive performance of `VIPRS` and `VIPRS-GS`. However, the improvement is not consistent across all traits, especially for the VEM-based `VIPRS`. Similar to previous studies,[6] we saw that, in some cases, including more variants led to modest drop in prediction accuracy, perhaps because of increased noise in the PRS estimate. Presumably, imputation errors for rare variants could potentially degrade the performance of the model in this setting. Therefore, we believe that this analysis presents a lower-bound on what could be achieved with more accurate and complete whole-genome-sequencing data[3,43] (discussion).

PRS analysis with external GWAS summary statistics

A common use case in the inference of polygenic scores involves settings where the GWAS summary statistics and the LD reference panel are estimated from two different cohorts.[8,37] In other cases, the GWAS summary statistics may be derived from a meta-analysis that combines data from a number of different studies. These settings may present potential mismatches and heterogeneities between the LD reference panel and GWAS summary statistics and are thus challenging to model, often leading to substantial loss in predictive power.[32,33,94,95] In the case of meta-analyzed GWAS summary statistics, previous studies have also cautioned that various sources of heterogeneity between the source cohorts may introduce estimation errors and biases for summary statistics-based methods.[96,97]

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

To systematically assess the robustness of VIPRS to potential heterogeneities and mismatches between the GWAS cohort and the LD reference panel, we conducted an analysis where we downloaded a number of publicly available GWAS summary statistics for some of the traits analyzed previously, including standing height,[74] BMI,[75] HDL and LDL cholesterol,[76] FVC and FEV1,[78] type 2 diabetes,[80] rheumatoid arthritis,[81] and Asthma[82] (Table 1, see web resources). These studies combined data from individuals of general European ancestry (excluding UKB participants), most in the form of meta-analysis. Therefore, we would expect some degree of differences between the LD reference panels derived exclusively from the White British cohort in the UKB and the in-sample LD from these GWAS cohorts, which are not available.

We fit VIPRS as well as other PRS methods to the external GWAS summary statistics, providing the same 5-fold validation and testing cohorts within the UK Biobank for the purposes of hyperparameter tuning and evaluation as in the previous analysis. Our results indicate that, for most of the studies analyzed, the VIPRS model is robust in the out-of-sample LD setting and achieves competitive prediction accuracy compared to popular baseline methods (Figure 6). In particular, VIPRS benefited substantially from the increased sample sizes ($N > 800,000$) in the GLGC meta-analysis of blood lipid traits,[76] outperforming all competing methods by large margins and significantly improving its accuracy relative to the within-UKB analysis pipeline. For instance, when using the GLGC summary statistics, polygenic scores estimated with VIPRS can now explain up to 21.1% of the variance in HDL compared to 18.4% when using UKB summary statistics. However, there are some notable cases of older GWAS meta-analyses (ca. 2010) where VIPRS was sensitive to mismatches between the GWAS cohort and LD reference panel (Figures 6 and S14), though not to the same extent as SBayesR, which failed to converge for many of the quantitative traits analyzed, consistent with earlier work in this area.[33] The weak performance of both VIPRS and SBayesR in this setting is notable, since both models iteratively estimate the residual variance σ_ϵ^2, whereas recent work recommended fixing this hyperparameter in the out-of-sample LD setting.[97] In our experiments, we did not see significant differences between VIPRS with fixed versus estimated residual variance. However, when using a validation set to tune the hyperparameters of the model, VIPRS-GS recovered most of the drop in performance relative to other PRS methods and showed competitive predictive ability (Figures 6 and S14). This suggests that the VIPRS model trained to maximize the ELBO (material and methods) of the external GWAS data may not generalize well to the UKB individuals. Indeed, our experiments show a partial reversal in the correspondence between the training ELBO and the validation R^2 (the metric that VIPRS-GS is optimizing) in the analysis of some of the external summary statistics (Figure S13), which explains the poor predictive performance of the basic VIPRS model in those settings.

Given this observation, if an independent validation set is not available, we recommend that users of the VIPRS software run principled tests of LD mismatch and heterogeneity, such as the recently published DENTIST or SLALOM methods[95,96] before fitting the model to GWAS summary data. In the supplemental methods, we also derived a stochastic estimator of the DENTIST test statistic that can be computed efficiently and we provided that as a utility function in our software (see web resources).

Discussion

In this paper, we introduced VIPRS, a fast and flexible Bayesian PRS method that approximates the posterior for the effect sizes of genetic variants on the phenotype by using VI techniques. Our genome-wide simulation analyses using genotype data from the White British cohort in the UK Biobank demonstrated that variational approximations to the posterior are not only computationally efficient but they also provide highly accurate polygenic score estimates across diverse genetic architectures. Indeed, in some simulation scenarios, VIPRS exceeded the predictive performance of competing Bayesian and non-Bayesian methods by large margins. The competitive prediction accuracy of the VIPRS method replicated in our analyses of real quantitative and binary phenotypes measured for the same UKB participants, though the differences between the methods in this setup were more modest. Similar systematic but mostly modest benefits were observed when PRS methods were applied to individuals from ancestries not included in the training dataset, emphasizing the robustness of the approach. For example, the effect size estimates by VIPRS for LDL cholesterol showed a large enough improvement in performance across ancestries to have potential clinical relevance[98] and make a significant dent in the transferability problem for that trait.

As highlighted throughout the text, we found that many implementation and modeling choices can have a substantial impact on the performance of the VIPRS model in analyses with GWAS summary statistics for real measured traits: hyperparameter tuning strategies, LD estimators, and the prior on the effect size all influenced the predictive performance in ways that varied across phenotypes and experimental setups. Overall, in most of the setups and experimental conditions that we tested, the grid search approach for hyperparameter tuning combined with the spike-and-slab prior and windowed estimator of LD reliably outperformed or rivaled all the other variations of the model as well as previously described PRS methods. Notably, many of the individual modeling choices underpinning the VIPRS-GS model have been tried and tested in at least one other publication. Even the variational approximation that we derive bears some similarities to some existing methods that we compare against in our experiments, e.g.,

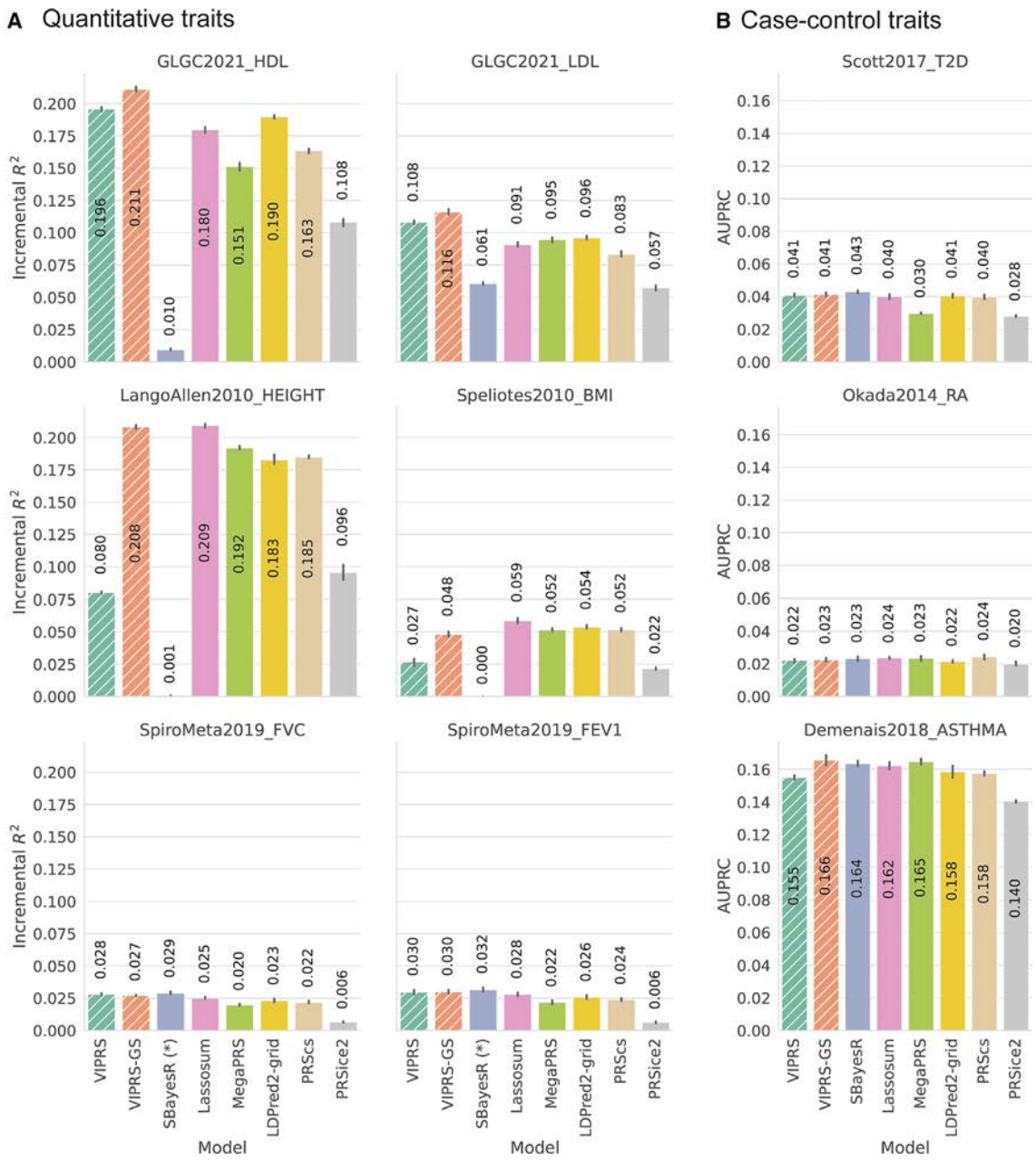


Figure 6. Predictive performance of summary statistics-based PRS methods on real quantitative and case-control phenotypes using external GWAS summary statistics

(A and B) Each panel shows the predictive performance, in terms of (A) incremental R^2 and (B) area under the precision recall curve (AUPRC), of various PRS methods when applied to an independent test cohort in the UK Biobank. The bars show the mean and standard error of the prediction metrics across the five folds and the black lines show the corresponding standard errors. The quantitative phenotypes analyzed are high-density lipoprotein (GLGC2021_HDL[76]), low-density lipoprotein (GLGC2021_LDL[76]), standing height (LangoAllen2010_HEIGHT[74]), body mass index (Speliotes2010_BMI[75]), forced vital capacity (SpiroMeta2019_FVC[78]), and forced expiratory volume in 1 s (SpitoMeta2019_FEV1[78]). The binary phenotypes analyzed are type 2 diabetes (Scott2017_T2D[80]), Rheumatoid arthritis (Okada2014_RA[81]), and asthma (Demenais2018_ASTHMA[82]). The PRS methods shown are our proposed `VIPRS` and `VIPRS-GS` (using grid search to tune model hyperparameters) as well as six other baseline models: `SBayesR`, `Lassosum`, `MegaPRS`, `LDPred2` (grid), `PRScs`, and `PRSice2` (C+T). The asterisk (*) next to the `SBayesR` method in (A) indicates that it did not converge on some of those traits. Dashed lines highlight the models contributed in this work.

`MegaPRS`[34] (supplemental methods). However, crucial details in the variational algorithm and its implementation and how they are joined together can still have significant impact on the overall performance, as illustrated by our experimental results.

One of the main strengths of the `VIPRS` model is its computational efficiency, which we exploited to test the predictive performance of the model with approximately 9.6 million SNPs, almost an order of magnitude greater than the standard HapMap3 subset used to train PRS

methods.[4,6,28,32] At this finer scale, we showed that modeling an expanded set of variants results in significant improvements in prediction accuracy for highly polygenic traits, such as standing height and HDL. This is consistent with recent whole-genome-sequencing analyses that showed that a considerable proportion of rare causal variants are not well tagged by common SNPs.[93] There are a number of reasons that lead us to believe that the performance metrics that we report here are a lower bound on what could be achieved in modeling large-scale SNP array data. First, the vast majority of the variants that we added beyond the HapMap3 subset are rare and statistically imputed. Rare variant imputation is still a challenging problem and existing algorithms are known to have elevated error rates.[99,100] We expect that these imputation errors can introduce substantial noise into the PRS estimate and thus result in decreased prediction accuracy, as we observed for a number of the traits that we analyzed. This difficulty can potentially be addressed by using whole-genome-sequencing data for GWASs, which may soon be enabled by recent large-scale initiatives by the UKB[43] and TOPMed.[3] Second, residual confounding due to population structure may affect effect size estimation for rare variants.[93,101,102] In our GWAS pipeline, we corrected for population stratification by using only the top ten PCs of the GRM, which may not adequately capture the more recent demographic history reflected by rare variants.[102,103] This residual confounding effect may be addressed by increasing the number of PCs used in the GWAS analysis[93] or utilizing more genealogically informed estimates of the GRM.[104]

Despite its competitive predictive ability, we believe that there are a number of modeling choices underlying `VIPRS` that can potentially be improved in future work. Firstly, compared to simulated phenotypes, the generative process for real traits is unknown and most likely involves complex and heterogeneous genetic architectures that are not well described by a two-component Gaussian mixture prior. The spike-and-slab prior assumes that all genetic variants have a uniform prior probability of being causal and that the causal SNPs have equal expected contribution to the heritability, which is a simplistic assumption given what is known about the genetic architectures of complex traits.[12–14] This motivated us to explore a more general and flexible Gaussian mixture prior with four mixture components.[6,26] Our experimental results show that adding mixture components improves accuracy for highly heritable and polygenic traits, such as standing height, but did not systematically improve accuracy for less heritable traits, perhaps because of reduced power to identify the larger number of parameters. Future work using priors informed by functional annotations (e.g., Zhang et al.[34] and Márquez-Luna et al.[105]) is a promising avenue to improve accuracy in these cases. Second, our validation analyses in the UKB confirmed that, in general, `VIPRS` and other PRS methods do not transfer well across populations or ancestry groups, despite some notable differences between the methods. Recent work has highlighted that transferability in the context of summary statistics-based PRS methods is best achieved when we jointly model the effect sizes of multiple ancestrally homogeneous populations within the same framework.[57,106,107] This formulation has proved successful for some Bayesian PRS methods[57,107] and we believe that fast variational approximations to the posterior under such models will increasingly be shown to be effective and highly competitive.

Finally, while our results showed that variational approximations to the posterior are a promising alternative to MCMC techniques in predictive settings, it is important to highlight that mean-field variational approaches are known to underestimate the posterior variances and covariances in some cases.[45,108,109] In practice, this may result in miscalibrated PRS confidence intervals, if such a quantity is sought for some downstream applications.[38] This limitation can be addressed with more expressive variational families,[110] such as those derived with variational boosting,[111] or alternatively, with the help of modern Bayesian inference techniques that combine variational methods and MCMC.[112]

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.03.009.

Acknowledgments

We thank Doug Speed, Tianyuan Lu, Lino Fonseca Ferreira, members of the Li and Gravel labs, and anonymous reviewers for useful feedback and discussions on earlier drafts of this manuscript. We are particularly grateful to Doruk Cakmakci and Goodarz Koli Farhood for testing various aspects of the software. We also thank Hans Markus Munter, Stephen Sawcer, and Mark Lathrop for help with data access and Celia Greenwood and Mathieu Blanchette for feedback on the project at various stages. The UK Biobank analyses in this work were conducted under application 4408. Y.L. is supported by Natural Sciences and Engineering Research Council (NSERC) discovery grant (RGPIN-2019-0621), Fonds de recherche Nature et technologies (FRQNT) New Career (NC-268592), and Canada First Research Excellence Fund Healthy Brains for Healthy Life (HBHL) initiative new investigator start-up award (G249591). S.G. was also supported by the Canadian Institute for Health Research (CIHR) project grant (437576) and the Canada Research Chair program.

Author contributions

S.G. and Y.L. jointly supervised this work. Y.L. conceived the study. Y.L. and S.Z. developed the methodology. S.Z. implemented the computational software and performed all the experiments. S.Z., Y.L., and S.G. analyzed the results. S.Z. wrote the initial manuscript. All of the authors wrote the final version.

Declaration of interests

The authors declare no competing interests.

Please cite this article in press as: Zabad et al., Fast and accurate Bayesian polygenic risk modeling with variational inference, The American Journal of Human Genetics (2023), https://doi.org/10.1016/j.ajhg.2023.03.009

Received: September 21, 2022
Accepted: March 13, 2023
Published: April 7, 2023

Web resources

External GWAS summary statistics for asthma, http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST006001-GCST007000/GCST006862/

External GWAS summary statistics for GLGC lipid traits, http://csg.sph.umich.edu/willer/public/glgc-lipids2021/

External GWAS summary statistics for SpiroMeta lung function traits, https://www.ebi.ac.uk/gwas/publications/30804560

External GWAS summary statistics for standing height, BMI, and Rheumatoid arthritis, https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/

External GWAS summary statistics for type 2 diabetes, http://diagram-consortium.org/downloads.html

LD matrices from the White British cohort in the UK Biobank in Zarr format, https://doi.org/10.5281/zenodo.7036625

Modeling and Analysis of (Statistical) Genetics data in python, https://github.com/shz9/magenpy

Variational Inference of Polygenic Risk Scores, https://github.com/shz9/viprs

Reproducible code for running the analyses in the paper "Fast and Accurate Bayesian Polygenic Risk Modeling with Variational Inference" (2022), https://github.com/shz9/viprs-paper

References

1. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data". Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.
2. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400. https://doi.org/10.1038/s41588-018-0047-6.
3. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299. https://doi.org/10.1038/s41586-021-03205-y.
4. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001.
5. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. *19*, 581–590. https://doi.org/10.1038/s41576-018-0018-x.
6. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Wray, N.R., Goddard, M.E., Yang, J., Visscher, P.M., and Metspalu, A. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. *10*, 5086. https://doi.org/10.1038/s41467-019-12653-0.
7. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. Genome Med. *12*, 44. https://doi.org/10.1186/s13073-020-00742-5.
8. Choi, S.W., Mak, T.S.H., and O'Reilly, P.F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nat. Protoc. *15*, 2759–2772. https://doi.org/10.1038/s41596-020-0353-1.
9. O'Connor, L.J., Schoech, A.P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A.L. (2019). Extreme polygenicity of complex traits is explained by negative selection. Am. J. Hum. Genet. *105*, 456–476. https://doi.org/10.1016/j.ajhg.2019.07.003.
10. Zeng, J., Xue, A., Jiang, L., Lloyd-Jones, L.R., Wu, Y., Wang, H., Zheng, Z., Yengo, L., Kemper, K.E., Goddard, M.E., et al. (2021). Widespread signatures of natural selection across human complex traits and functional genomic categories. Nat. Commun. *12*, 1164. https://doi.org/10.1038/s41467-021-21446-3.
11. Johnson, R., Burch, K.S., Hou, K., Paciuc, M., Pasaniuc, B., and Sankararaman, S. (2021). Estimation of regional polygenicity from GWAS provides insights into the genetic architecture of complex traits. PLoS Comput. Biol. *17*, e1009483. https://doi.org/10.1371/journal.pcbi.1009483.
12. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235. https://doi.org/10.1038/ng.3404.
13. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., and Price, A.L. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nat. Genet. *49*, 1421–1427. https://doi.org/10.1038/ng.3954.
14. Speed, D., Holmes, J., and Balding, D.J. (2020). Evaluating and improving heritability models using summary statistics". Nat. Genet. *52*, 458–462. https://doi.org/10.1038/s41588-020-0600-y.
15. Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat. Rev. Genet. *17*, 392–406. https://doi.org/10.1038/nrg.2016.27.
16. Hivert, V., Sidorenko, J., Rohart, F., Goddard, M.E., Yang, J., Wray, N.R., Yengo, L., and Visscher, P.M. (2021). Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. Am. J. Hum. Genet. *108*, 786–798. https://doi.org/10.1016/j.ajhg.2021.02.014.
17. Palmer, D.S., Zhou, W., Abbott, L., Baya, N., Churchhouse, C., Seed, C., Poterba, T., King, D., Kanai, M., Bloemendal, A., et al. (2022). Analysis of genetic dominance in the UK Biobank. Preprint at bioRxiv. https://doi.org/10.1101/2021.08.15.456387.
18. Lambert, S.A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk scores. Hum. Mol. Genet. *28*, R133–R142. https://doi.org/10.1093/hmg/ddz187.
19. Hao, L., Kraft, P., Berriz, G.F., Hynes, E.D., Koch, C., Koratagere V Kumar, P., Parpattedar, S.S., Steeves, M., Yu, W., Antwi, A.A., et al. (2022). Development of a clinical polygenic risk score assay and reporting workflow. Nat. Med. *28*, 1006–1013. https://doi.org/10.1038/s41591-022-01767-6.

20. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224. https://doi.org/10.1038/s41588-018-0183-z.
21. Dai, J., Lv, J., Zhu, M., Wang, Y., Qin, N., Ma, H., He, Y.Q., Zhang, R., Tan, W., Fan, J., et al. (2019). Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. Lancet Respir. Med. *7*, 881–891. https://doi.org/10.1016/S2213-2600(19)30144-4.
22. Sugrue, L.P., and Desikan, R.S. (2019). What are polygenic scores and why are they important? JAMA, J. Am. Med. Assoc. *321*, 1820–1821. https://doi.org/10.1001/jama.2019.3893.
23. Natarajan, P., Young, R., Stitziel, N.O., Padmanabhan, S., Baber, U., Mehran, R., Sartori, S., Fuster, V., Reilly, D.F., Butterworth, A., et al. (2017). Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting". Circulation *135*, 2091–2101. https://doi.org/10.1161/CIRCULATIONAHA.116.024436.
24. Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics *157*, 1819–1829. https://doi.org/10.1093/genetics/157.4.1819.
25. Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the bayesian alphabet. Genetics *183*, 347–363. https://doi.org/10.1534/genetics.109.103952.
26. Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet. *11*, e1004969. https://doi.org/10.1371/journal.pgen.1004969.
27. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., and Sham, P.C. (2017). Polygenic scores via penalized regression on summary statistics. Genet. Epidemiol. *41*, 469–480. https://doi.org/10.1002/gepi.22050.
28. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776. https://doi.org/10.1038/s41467-019-09718-5.
29. Choi, S.W., and O'Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. GigaScience *8*, giz082. https://doi.org/10.1093/gigascience/giz082.
30. Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., Rivas, M.A., and Hastie, T. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. PLoS Genet. *16*. 110091411–30. https://doi.org/10.1371/journal.pgen.1009141.
31. Yang, S., and Zhou, X. (2020). Accurate and scalable construction of polygenic scores in large biobank data sets. Am. J. Hum. Genet. *106*, 679–693. https://doi.org/10.1016/j.ajhg.2020.03.013.
32. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2020). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029.
33. Zhou, G., and Zhao, H. (2021). A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. PLoS Genet. *17*, e1009697. https://doi.org/10.1371/journal.pgen.1009697.
34. Zhang, Q., Privé, F., Vilhjálmsson, B., and Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. Nat. Commun. *12*, 4192. https://doi.org/10.1038/s41467-021-24485-y.
35. Mitt, M., Kals, M., Pärn, K., Gabriel, S.B., Lander, E.S., Palotie, A., Ripatti, S., Morris, A.P., Metspalu T, A., Palta, P., and Mägi, R. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur. J. Hum. Genet. *25*, 869–876. https://doi.org/10.1038/ejhg.2017.51.
36. International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58. https://doi.org/10.1038/nature09298.
37. Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. *18*, 117–127. https://doi.org/10.1038/nrg.2016.142.
38. Ding, Y., Hou, K., Burch, K.S., Lapinska, S., Privé, F., Vilhjálmsson, B., Sankararaman, S., and Pasaniuc, B. (2022). Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. Nat. Genet. *54*, 30–39. https://doi.org/10.1038/s41588-021-00961-5.
39. Pain, O., Glanville, K.P., Hagenaars, S.P., Selzam, S., Fürtjes, A.E., Gaspar, H.A., Coleman, J.R.I., Rimfeld, K., Breen, G., Plomin, R., et al. (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. PLoS Genet. *17*. e1009021–e1009022. https://doi.org/10.1371/journal.pgen.1009021.
40. Yang, S., and Zhou, X. (2022). PGS-server: accuracy, robustness and transferability of polygenic score methods for biobank scale studies. Briefings Bioinf. *23*, 1477–4054. https://doi.org/10.1093/bib/bbac039.
41. Bishop, Christopher M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics) (Springer-Verlag).
42. Murphy, K.P. (2012). Probabilistic Machine Learning: An Introduction.
43. Halldorsson, B.V., Eggertsson, H.P., Moore, K.H.S., Hauswedell, H., Eiriksson, O., Ulfarsson, M.O., Palsson, G., Hardarson, M.T., Oddsson, A., Jensson, B.O., et al. (2022). The sequences of 150,119 genomes in the UK Biobank. Nature *607*, 732–740. https://doi.org/10.1038/s41586-022-04965-x.
44. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. (1999). An introduction to variational methods for graphical models. Mach. Learn. *37*, 183–233.
45. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational inference: a review for statisticians. J. Am. Stat. Assoc. *112*, 859–877. https://doi.org/10.1080/01621459.2017.1285773.
46. Hoffman, M.D., Blei, D.M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. J. Mach. Learn. Res. *14*.
47. Kingma, D.P., and Welling, M. (2014). Auto-encoding variational bayes. Preprint at arXiv. https://doi.org/10.48550/arXiv.1312.6114.
48. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian

mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290. https://doi.org/10.1038/ng.3190.

49. Logsdon, B.A., Hoffman, G.E., and Mezey, J.G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. BMC Bioinf. *11*, 58. https://doi.org/10.1186/1471-2105-11-58.
50. Demetci, P., Cheng, W., Darnell, G., Zhou, X., Ramachandran, S., and Crawford, L. (2021). Multi-scale inference of genetic trait architecture using biologically annotated neural networks. PLoS Genet. *17*. 110097544–53. https://doi.org/10.1371/journal.pgen.1009754.
51. Carbonetto, P., and Stephens, M. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. *7*. https://doi.org/10.1214/12-BA703.
52. Zhang, W., Najafabadi, H., and Yue, L. (2021). SparsePro: an efficient genome-wide fine-mapping method integrating summary statistics and functional annotations. Preprint at bioRxiv. https://doi.org/10.1101/2021.10.04.463133.
53. Carbonetto, P., and Stephens, M. (2013). Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in crohn's disease. PLoS Genet. *9*, e1003770. https://doi.org/10.1371/journal.pgen.1003770.
54. Zhu, X., and Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat. Commun. *9*, 4361. https://doi.org/10.1038/s41467-018-06805-x.
55. Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat. Commun. *8*, 456. https://doi.org/10.1038/s41467-017-00470-2.
56. Spence, J. (2020). Flexible mean field variational inference using mixtures of non-overlapping exponential families. In Adv. Neural Inf. Process. Syst., H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, eds. (Curran Associates, Inc.), pp. 19642–19654. https://proceedings.neurips.cc/paper/2020/file/e3a54649aeec04cf1c13907bc6c5c8aa-Paper.pdf.
57. Spence, J.P., Sinnott-Armstrong, N., Assimes, T.L., and Pritchard, J.K. (2022). A flexible modeling and inference framework for estimating variant effect sizes from GWAS summary statistics. Preprint at bioRxiv. https://doi.org/10.1101/2022.04.18.488696.
58. Falconer, D.S. (1967). The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. Ann. Hum. Genet. *31*, 1–20. https://doi.org/10.1111/j.1469-1809.1967.tb01249.x.
59. Pirinen, M., Donnelly, P., and Spencer, C.C.A. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Ann. Appl. Stat. *1*, 369–390.
60. Gillett, A.C., Vassos, E., and Lewis, C. (2018). Transforming summary statistics from logistic regression to the liability scale: application to genetic and environmental risk scores. Hum. Hered. *83*, 210–224. https://www.jstor.org/stable/48506850.
61. Mitchell, T.J., and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. J. Am. Stat. Assoc. *83*, 1023–1032. https://doi.org/10.1080/01621459.1988.10478694.
62. George, E.I., and McCulloch, R.E. (1997). Approaches for bayesian variable selection. Stat. Sin. *7*.
63. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies". Nat. Genet. *47*, 291–295. https://doi.org/10.1038/ng.3211.
64. Titsias, M.K., and Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning.
65. Ročková, V., and George, E.I. (2014). The EM approach to Bayesian variable selection. J. Am. Stat. Assoc. *109*, 828–846. https://doi.org/10.1080/01621459.2013.869223.
66. Snoek, J., Larochelle, H., and Adams, Ryan P. (2012). Practical Bayesian optimization of machine learning algorithms *4*.
67. Agnihotri, A., and Batra, N. (2020). Exploring Bayesian Optimization. Distill *5*. https://doi.org/10.23915/distill.00026.
68. Carbonetto, P., Zhou, X., and Stephens, M. (2017). varbvs: Fast Variable Selection for Large-Scale Regression.
69. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*. https://doi.org/10.1186/s13742-015-0047-8.
70. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285. https://doi.org/10.1093/bioinformatics/btv546.
71. Wen, X., and Stephens, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. Ann. Appl. Stat. *4*, 1158–1182. https://doi.org/10.1214/10-AOAS338.
72. McCaw, Z.R., Lane, J.M., Saxena, R., Redline, S., and Lin, X. (2020). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. Biometrics *76*, 1262–1272.
73. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Bertrand, T., Grisel, O., Blondel, M., Peter, P., Weiss, R., Vincent, D., et al. (2011). Scikit-learn: machine learning in python. J. Mach. Learn. Res. *12*, 2825–2830.
74. Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 7317. https://doi.org/10.1038/nature09410.
75. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, Jian'An, Mägi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. *42*, 937–948. https://doi.org/10.1038/ng.686.
76. Graham, S.E., Clarke, S.L., Wu, K.-H.H., Kanoni, S., Zajac, G.J.M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T.W., et al. (2021). The power of genetic diversity in genome-wide association studies of lipids. Nature *600*, 675–679. https://doi.org/10.1038/s41586-021-04064-3.
77. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713. https://doi.org/10.1038/nature09270.

78. Shrine, N., Guyatt, A.L., Erzurumluoglu, A.M., Jackson, V.E., Hobbs, B.D., Melbourne, C.A., Batini, C., Fawcett, K.A., Song, K., Sakornsakolpat, P., et al. (2019). New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat. Genet. *51*, 481–493. https://doi.org/10.1038/s41588-018-0321-7.
79. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. *44*, 981–990. https://doi.org/10.1038/ng.2383.
80. Scott, R.A., Scott, L.J., Mägi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., Eicher, J.D., et al. (2017). An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes *66*, 2888–2902. https://doi.org/10.2337/db16-1253.
81. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381. https://doi.org/10.1038/nature12873.
82. Demenais, F., Margaritte-Jeannin, P., Barnes, K.C., Cookson, W.O.C., Altmüller, J., Ang, W., Barr, R.G., Beaty, T.H., Becker, A.B., Beilby, J., et al. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat. Genet. *50*, 42–53. https://doi.org/10.1038/s41588-017-0014-7.
83. Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P.F., and Vilhjálmsson, B.J. (2022). Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. Am. J. Hum. Genet. *109*, 12–23. https://doi.org/10.1016/j.ajhg.2021.11.008.
84. Ni, G., Zeng, J., Revez, J.A., Wang, Y., Zheng, Z., Ge, T., Restuadi, R., Kiewa, J., Nyholt, D.R., Coleman, J.R.I., et al. (2021). A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. Biol. Psychiatry *90*, 611–620. https://doi.org/10.1016/j.biopsych.2021.04.018.
85. Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D.S., and Smith, K. (2011). Cython: The best of both worlds. Comput. Sci. Eng. *13*, 31–39.
86. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., and Herrera, F. (2018). Learning from Imbalanced Data Sets. https://doi.org/10.1007/978-3-319-98074-4.
87. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591. https://doi.org/10.1038/s41588-019-0379-x.
88. Tzikas, D.G., Likas, A.C., and Galatsanos, N.P. (2008). The variational approximation for Bayesian inference. IEEE Signal Process. Mag. *25*, 131–146. https://doi.org/10.1109/msp.2008.929620.
89. Khan, M.E., Bouchard, G., Marlin, B.M., and Murphy, K.P. (2010). Variational bounds for mixed-data factor analysis.
90. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2019). Introduction to Statistical Learning with Applications in R*11*.
91. Galinsky, K.J., Reshef, Y.A., Finucane, H.K., Loh, P.-R., Zaitlen, N., Patterson, N.J., Brown, B.C., and Price, A.L. (2019). Estimating cross-population genetic correlations of causal effect sizes. Genet. Epidemiol. *43*, 180–188. https://doi.org/10.1002/gepi.22173.
92. Shi, H., Gazal, S., Kanai, M., Koch, E.M., Schoech, A.P., Siewert, K.M., Kim, S.S., Luo, Y., Amariuta, T., Huang, H., et al. (2021). Population-specific causal disease effect sizes in functionally important regions impacted by selection. Nat. Commun. *12*, 1098. https://doi.org/10.1038/s41467-021-21286-1.
93. Wainschtein, P., Jain, D., Zheng, Z., TOPMed Anthropometry Working Group; and NHLBI Trans-Omics for Precision Medicine TOPMed Consortium, Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat. Genet. *54*, 263–273. https://doi.org/10.1038/s41588-021-00997-7.
94. Privé, F., Arbel, J., Aschard, H., and Vilhjálmsson, B.J. (2022). Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. Human Genetics and Genomics Advances *3*, 100136. https://doi.org/10.1016/j.xhgg.2022.100136.
95. Chen, W., Wu, Y., Zheng, Z., Qi, T., Visscher, P.M., Zhu, Z., and Yang, J. (2021). Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. Nat. Commun. *12*, 7117. https://doi.org/10.1038/s41467-021-27438-7.
96. Kanai, M., Roy, E., Zhou, W., Zhou, W., Kanai, M., Wu, K.-H.H., Rasheed, H., Tsuo, K., Hirbo, J.B., Wang, Y., et al. (2022). Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. Cell Genomics *2*, 100210. https://doi.org/10.1016/j.xgen.2022.100210.
97. Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2022). Fine-mapping from summary data with the "Sum of Single Effects" model. PLoS Genet. *18*, 11010299–9–24. https://doi.org/10.1371/journal.pgen.1010299.
98. Wu, H., Forgetta, V., Zhou, S., Bhatnagar, S.R., Paré, G., and Brent Richards, J. (2021). Polygenic risk score for low-density lipoprotein cholesterol is associated with risk of ischemic heart disease and enriches for individuals with familial hypercholesterolemia. Circulation: Genomic and Precision Medicine *14*, e003106. https://doi.org/10.1161/CIRCGEN.120.003106.
99. Hoffmann, T.J., and Witte, J.S. (2015). Strategies for Imputing and Analyzing Rare Variants in Association Studies. Trends Genet. *31*, 556–563. https://doi.org/10.1016/j.tig.2015.07.006. https://pubmed.ncbi.nlm.nih.gov/26450338.
100. Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., Wu, J., and Xiao, J. (2018). Comprehensive assessment of genotype imputation performance. Hum. Hered. *83*, 107–116. https://doi.org/10.1159/000489758.
101. O'Connor, T.D., Adam, K., Bamshad, M., Rich, S.S., Smith, J.D., Turner, E., NHLBIGO Exome Sequencing Project, Statistical Analysis Working Group ESP Population Genetics, M Leal, S., and Akey, J.M. (2013). Fine-scale patterns of population stratification confound rare variant association tests. PLoS One *8*, e65834. https://doi.org/10.1371/journal.pone.0065834. https://pubmed.ncbi.nlm.nih.gov/23861739.
102. Zaidi, A.A., and Mathieson, I. (2020). Demographic history mediates the effect of stratification on polygenic scores. G.H. Perry, M.C. Turchin, and A.R. Martin, eds. *9*, e61548. https://doi.org/10.7554/eLife.61548.

103. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246. https://doi.org/10.1038/ng.1074.
104. Fan, C., Mancuso, N., Charleston, W., and Chiang, K. (2022). A genealogical estimate of genetic relationships. Am. J. Hum. Genet. *109*, 812–824. https://doi.org/10.1016/j.ajhg.2022.03.016.
105. Márquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S.S., Furlotte, N., Auton, A., 23andMe Research Team, Price, A.L., Bell, Robert K., Bryc, K., et al. (2021). Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. Nat. Commun. *12*, 6052. https://doi.org/10.1038/s41467-021-25171-9.
106. Cai, M., Xiao, J., Zhang, S., Wan, X., Zhao, H., Chen, G., and Yang, C. (2021). A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. Am. J. Hum. Genet. *108*, 632–655. https://doi.org/10.1016/j.ajhg.2021.03.002.
107. Ruan, Y., Lin, Y.-F., Feng, Y.C.A., Chen, C.-Y., Lam, M., Guo, Z., Stanley Global Asia Initiatives, He, L., Sawa, A., Martin, A.R., et al. (2022). Improving polygenic prediction in ancestrally diverse populations. Nat. Genet. *54*, 573–580. https://doi.org/10.1038/s41588-022-01054-7.
108. Turner, R.E., and Sahani, M. (2011). Two problems with variational expectation maximisation for time series models. In Bayesian Time Series Models, A. David Barber, T. Cemgil, and S.E. Chiappa, eds. (Cambridge University Press), pp. 104–124. https://doi.org/10.1017/CBO9780511984679.006.
109. Giordano, R., Broderick, T., and Jordan, M.I. (2017). Covariances, robustness, and variational bayes. Preprint at arXiv. https://doi.org/10.48550/ARXIV.1709.02536.
110. Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. (2019). Advances in variational inference. IEEE Trans. Pattern Anal. Mach. Intell. *41*, 2008–2026. https://doi.org/10.1109/TPAMI.2018.2889774.
111. Miller, A.C., Foti, N.J., and Adams, R.P. (2017). Variational boosting: iteratively refining posterior approximations. In Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y.W. Teh, eds. (Proceedings of Machine Learning Research. PMLR), pp. 2420–2429. https://proceedings.mlr.press/v70/miller17a.html.
112. Salimans, T., Kingma, D.P., and Welling, M. (2014). Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. Preprint at arvix. https://doi.org/10.48550/ARXIV.1410.6460.