Evaluation of Bayesian Linear Regression Models as a Fine Mapping tool

Merina Shrestha[1], Zhonghao Bai[1], Tahereh Gholipourshahraki[1], Astrid Johannesson Hjelholt[2,3,4], Mads Kjølby[2,3,4,5], Palle Duun Rohde[6], and Peter Sørensen[1]

[1] Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark, [2] Department of Biomedicine, Aarhus University, Aarhus, Denmark, [3] Department of Clinical Pharmacology, Aarhus University Hospital, Aarhus, Denmark, [4] Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus, Denmark, [5] Department of Population Health and Genomics, University of Dundee, Dundee, United Kingdom, [6] Genomic Medicine, Department of Health Science and Technology, Aalborg University, Aalborg, Denmark

Abstract

Bayesian linear regression (BLR) models consider the underlying genetic architecture of complex phenotypes by specifying different prior distributions for SNP effects allowing heterogenous distribution of the true genetic signals. Our goal is to evaluate BLR models with BayesC and BayesR prior distributions for fine mapping on simulated and real binary and quantitative phenotypes, and compare them to the state-of-the-art external models: FINEMAP, SuSIE-RSS, SuSIE-Inf and FINEMAP-Inf. Evaluation of models was based on the F_1 classification score for simulations, and predictive accuracy for the UK Biobank (UKB) phenotypes. We used over 533K genotyped SNPs (simulations) and 6.6 million imputed SNPs (UKB phenotypes), from over 335K White British Unrelated samples in the UK biobank. We simulated phenotypes from low (GA1) to moderate (GA2) polygenicity, heritability(h^2) of 30% and 10%, causal SNPs (π) of 0.1% and 1% sampled genome-wide, and prevalence (PV) of 5% and 15%. Summary statistics and in-sample linkage disequilibrium were used to fit models in regions defined by lead SNPs. BayesR improved the F_1 score, averaged across all simulations, by by 27.26%, 26.96%, 18.40%, 15.42%, and 13.32% relative to FINEMAP-Inf, BayesC, FINEMAP, SUSIE-RSS and SUSIE-Inf. Prediction R^2, averaged across all the UKB quantitative phenotypes, with BayesR decreased by 5.32% and 3.71% compared to SuSIE-Inf and FINEMAP-Inf, whereas increased by 7.93% and 8.3% compared to SuSIE-RSS and BayesC. Prediction AUC, averaged across all the UKB binary phenotypes, with BayesR increased by 0.40%, 0.16%, 0.08%., and 0.05% relative to SuSIE-RSS, BayesC, FINEMAP-Inf, and SuSIE-Inf. These findings suggest that the performance of the BLR models was comparable to the state-of-the-art external models. The performance of BayesR prior was closely aligned with SuSIE-Inf and FINEMAP-Inf models. Results from both simulations and application of the models in the UKB phenotypes suggest that the BLR models are efficient fine mapping tools.

Introduction

To better understand the genetic make-up of complex quantitative phenotypes and multifactorial diseases, it is essential to be able to identify the set of genetic variants that are most likely causal or linked with the causal genetic variants (effective genetic variants). Genome-wide association studies (GWAS) often identify too many genetic variants because long-range LD complicates statistical inference. Hence additional work such as statistical fine mapping is required to refine signal from GWAS to determine the genetic variant (or

variants) most likely responsible for complex phenotypes, given verified association of region (or regions) with a trait. This task of determining the genetics variants and quantifying the evidence of association is crucial as it is often followed by large scale replication studies, or laboratory functional studies to gain further biological insights for potential clinical application with drug discovery, drug-repositioning in humans. Fine mapping methods usually assume presence of potential causal genetic variants in the data (Hutchinson et al. 2020b). With the concept of existence of multiple causal genetic variants in a locus various Bayesian fine mapping methods were developed (Hutchinson et al. 2020a). Bayesian fine mapping methods are able to quantify uncertainty of a potential causal genetic variant through posterior inclusion probability (PIP) in a model. The PIP of a SNP refers to the mean of posterior probability that the SNP is included in a model with non-zero effect, which provides evidence that a SNP should be included as potentially causative (Schaid et al. 2018).These methods are also able to leverage knowledge about genetic architecture of complex phenotypes through prior distribution of the effects and number of the genetic variants, which helps to improve statistical power for identifying effective genetic variants (Cui et al. 2022).

In the quest for accurate identification of effects of potential causal genetic variants in complex phenotypes, various Bayesian fine mapping methods have been developed with different modeling assumptions. FINEMAP (Benner et al. 2016) for GWAS summary statistics, uses Shotgun Stochastic Search (SSS) algorithm to search for different possible causal configurations followed by focusing on the configurations with non-negligible probability. The algorithm conducts a pre-defined number of iterations within the space of causal configurations. In each iteration, the neighborhood of the current causal configuration is defined by configurations that result from deleting, changing, or adding a causal SNP from the current configuration. In the next iteration, a new causal configuration is sampled from the neighborhood configurations based on its posterior normalized within the neighborhood. The sum of single effects (SuSIE) (Wang et al. 2020), a new formulation of Bayesian Variable Selection Regression (BVSR) uses a procedure called as iterative Bayesian stepwise selection approach (IBSS) to fit a model assuming few sparse causal genetic variants in a locus. It estimates the vector of regression coefficients for sparse genetic variants by summing up multiple single-effect vectors that each have one non-zero entry for a potential causal variant. (Zou et al. 2022) extended SuSIE by integrating use of summary data and introduced it as SuSIE-RSS. SuSIE-Inf and FINEMAP-Inf (Cui et al. 2022), an infinitesimal model similar to linear mixed models, are extensions of SuSIE and FINEMAP respectively where infinitesimal effects for genetic variants in LD (estimated separately for locus) with those of the sparse components are modeled. As in FINEMAP, FINEMAP-Inf used SSS algorithm, where the posterior inference of the sparse genetic variants is marginalized over the infinitesimal effects and the residuals. As in SuSIE, SuSIE-Inf estimated the sparse causal effects by summing up multiple single-effect vectors, where for the posterior inference of sparse genetic effects the joint distribution of single-effect vectors is marginalized over the infinitesimal effects and residuals.

Following a similar pursuit of accurate fine mapping, we are interested in investigating the performance of Bayesian Linear Regression (BLR) or multiple regression models in fine mapping. The BLR models have been applied for mapping genetic variants, prediction (genomic scores), estimation of genetic parameters and genetic architectures (Moser et al. 2015). The BLR models allow joint estimation of marker effects accounting for linkage disequilibrium (LD) among SNPs capturing the amount of variation at a genetic locus whose extent is based on the extent of the colocalization of multiple causal genetic variants. Assumptions of number and effect sizes of genetic variants are based on different priors.

Here we focus on the performance of the BLR models with priors BayesC (Habier et al. 2011), and BayesR (Erbe et al. 2012). During joint estimation, depending on priors, BLR models either shrink non-causal genetic variants' effects or induce both variable selection method and shrinkage helping to obtain accurate estimates of effects of genetic variants. To our knowledge, the BLR models have been investigated numerous times in predictions (Lloyd-Jones et al. 2019; Privé et al. 2020; Zhang et al. 2021) but only few studies have investigated precision and power of these models in fine mapping approach (Zou et al. 2022; Cui et al. 2022).

In fine mapping, no single genetic variant is identified as causal due to complex correlation patterns existing between the genetic variants hence "credible sets" of potentially causal genetic variants are prioritized (Hutchinson et al. 2020b). Credible sets help in variable selection by narrowing down to a small set of most likely causal variants with certain probability, refining the fine mapping approach. Consortium et al. (2012) developed a standard Bayesian approach for fine mapping, assumed a single causal locus per genetic region, to prioritize for example 99% credible set (a set whose cumulative sum of PIPs exceeds 0.99 threshold) of potentially causal genetic variants providing a credible set for per region (Hutchinson et al. 2020a). With the main goal of improving fine mapping resolution, the credible sets are meant to contain as few genetic variants as possible while still capturing an effective genetic variant.

In our study, we aim to assess the efficiency of the BLR models (BayesC and BayesR) as a fine mapping tool using GWAS summary statistics. Using simulated data, we designed credible sets then investigated precision and power of inclusion of casual variants in the credible sets to calculate F_1 classification score (F_1 score). We also evaluated these models based on the features of credible sets, and the prediction accuracy using fine mapped regions for five binary and five quantitative UK Biobank (UKB) phenotypes (Bycroft et al. 2018). The results from BayesC and BayesR priors are compared to the state-of-the-art methods such as FINEMAP (Benner et al. 2016), SuSIE-RSS (Zou et al. 2022), SuSIE-Inf and FINEMAP-Inf (Cui et al. 2022).

Material and Methods

In our study, efficiency of different models was investigated on simulations and the UKB phenotypes. We explored efficiency of the models on complex nature of phenotypes by simulating phenotypes with low to high polygenic background and creating different architecture utilizing different values for heritability (h^2_{SNP}), proportion of causal markers (π) and their effect sizes (NR Wray et al. 2014). Efficiency of the models was also investigated using five quantitative and five binary UKB phenotypes available from the UKB. For the UKB phenotypes, imputed markers were used. We analysed quantitativeand binary phenotypes, five of each type from the UKB. We have also discussed the theory behind single marker-linear regression analysis and its extension to summary data followed by the prior assumptions of BayesC and BayesR, used in our study. Marginal marker effects obtained from the single SNP association analysis are adjusted at multiple designed fine-mapping regions using the BLR models and external fine-mapping models. We present the design of credible sets (CSs) and definition of precision and power in terms of CSs to estimate F_1 classification score (F_1) on simulations. We compare the predictive abilities (coefficient of determination: R^2 for quantitative phenotypes and Area under the curve(AUC) for binary phenotypes) and the features of CSs.

Data

UK biobank (UKB) genotyped and imputed data were used for simulations and analysis of the UKB phenotypes respectively. In our study, we had information about 488,377 participants. To obtain a genetic homogeneous study population we restricted our analyses to unrelated British Caucasians and excluded individuals with more than 5,000 missing markers or individuals with autosomal aneuploidy. Remaining (n=335,532) White British unrelated individuals (WBU) were used for analyses. Then, we excluded markers with minor allele frequency less than < 0.01, call rate less than < 0.95 and the markers deviating from Hardy-Weinberg equilibrium (P-value $< 1 \times 10^{-12}$). We excluded markers located within the major histocompatibility complex (MHC), having ambiguous allele (i.e., GC or AT), were multi-allelic or an indel (Marees et al. 2018). This resulted in a total of 533,679 single nucleotide polymorphism (SNP) markers in the simulated data. For the UKB imputed data, firstly the markers with the probability of 70% (–hard-call threshold 0.7) were converted to genotypes followed by retaining markers with imputation INFO score >= 0.8 using PLINK 2.0 (Chang et al. 2015). The same quality control criteria were applied to the imputed markers as for the genotyped data, except that we included MHC in the UKB phenotypes as this region contains many known disease-associated markers. After quality control we retained 6,627,732 SNPs and 335,532 WBU for downstream analysis in the UKB imputed data.

Genetic architectures for simulations

To simulate genetic architectures from low to high polygenicity, we simulated quantitative phenotypes with heritability (h^2_{SNP}) of 30% and 10%, with two different proportions of causal SNPs (π), 0.1% and 1%, chosen randomly from the genome.

We generated two different types of genetic architectures under a multiple regression model. In the first genetic architecture (GA_1), causal SNPs (m_C) effects were sampled from the same normal distribution:

$$y_i = \sum_{j=1}^{m_C} w_{ij}\, b_j + e_i,$$

where y_i is the phenotype for i'th individual, b_j is the estimate of the j'th SNP effect (normally distributed with mean of 0 and variance given by σ_g^2/m_C). We assumed variance of a phenotype to be 1 such that σ_g^2 is equal to h^2_{snp}. w_{ij} represents the j'th centered and scaled genotype of the i'th individual:

where, x_{ij} is the effect allele count for i'th individual at the j'th SNP, p_j is the allele frequency of the j'th SNP. e_i is the residual that has a normal distribution with mean=0 and variance= $\sigma_g^2\left(1 \,/\, \left(h^2_{snp}\right) - 1\right)$. Residual variance was scaled in a way so that h^2_{snp} remained 30% (or 10%).

In the second genetic architecture scenario (GA_2), the effects of causal SNPs are sampled from a mixture of normal distributions.

$$y_i = \sum_{j=1}^{m_{C_1}} w_{ij}\, b_j + \sum_{k=1}^{m_{C_2}} w_{ik}\, b_k + \sum_{l=1}^{m_{C_3}} w_{il}\, b_l + e_i,$$

where, b_j , b_k , and b_l are the effect of causal SNPs sampled from normal distribution with mean=0 and variance = $(0.6\sigma_g^2)/(0.93m_C)$, $(0.2\sigma_g^2)/(0.05m_C)$, and $(0.2\sigma_g^2)/(0.02m_C)$ respectively. In this genetic model, the three normal distributions were designed such that 93% of the causal SNPs would have small effect sizes and the remaining 5% and 2% of the causal SNPs would have moderate and large effect sizes respectively. This genetic architecture was designed in a similar way as designed in the study by Lloyd-Jones et al. (2019).

All the other parameters in GA_2 are created in a similar way as for the GA_1.

We created ten replicates for each simulation scenario. The total sample of 335,532 individuals were divided into ten replicates. Each replicate contained 80% of the randomly sampled data from the total samples.

For quantitative phenotypes, in total we applied eight different simulation scenarios: two values of h^2_{snp}, two different proportions of causal SNPs π and two different genetic architecture scenarios.

To simulate binary phenotypes, in addition to the parameters: h^2_{SNP}, π and genetic architectures, we introduced another parameter "prevalence" `PV. Two different PV of 5% and 15% were used in our study. We simulated binary phenotypes from quantitative phenotypes. To simulate a binary phenotype, for example with PV 5% , we chose top 5% of individuals with highest simulated quantitative values as cases and the remaining as controls for the total sample in a replicate. Each scenario of a quantitative phenotype gave rise to two different scenarios for binary phenotype. In total we designed 16 different simulation scenarios for the binary phenotypes: two values of h^2_{snp}, two different proportions of causal SNPs π, two different genetic architecture scenarios, and two prevalence PV. Different scenarios for the quantitative and the binary phenotypes are presented in detail in Table S1. The flowchart of design of the simulations is presented in Figure 1.

Definition of phenotypes from the UKB data

We used the UKB phenotypes from the UKB for analysis. We chose five quantitative phenotypes: Body mass index (BMI), Hip circumference (HC), Standing height (Height), Waist circumference (WC) and Waist-to-hip ratio (WHR), and five binary phenotypes: Coronary artery disease (CAD), Hypertension (HTN), Psoriasis (PSO), Rheumatoid arthritis (RA) and Type 2 Diabetes mellitus (T2D). The quantitative phenotypes were identified using specific codes in the UKB data. To obtain WHR, we estimated ratio of the waist circumference to the hip circumference. A phenotype can have multiple instances in the UKB where we used only the first instance because of the least number of non-missing samples in that instance. We used specific codes to extract information of the quantitative phenotypes from the UKB data as mentioned in Table 1a. For the definition of the binary phenotypes, to define individuals as cases for a phenotype of interest we used codes from the data field "Diagnosis-main ICD10" along with codes from the self-reported information (Table 1b). All the individuals missing the appropriate codes for the phenotype of interest was reported as controls. We also extracted information: age at recruitment (p21022), sex (p31), and the UKB assessment center (p54). Detail information regarding number of samples, prevalence for the phenotypes are given in Table 1a and Table 1b.

Statistical model

In the multiple regression model the phenotype is related to the set of genetic markers:

$$y = Xb + e,$$

where y is the phenotype, X a matrix of genotyped markers, where values are standardized to give the ijth element as: $x_{ij} = (x_{ij} - 2p_j)/\sqrt{2p_j(1-p_j)}$, with x_{ij} the number of copies of the effect allele (e.g. 0, 1 or 2) for the ith individual at the jth marker and p_j the allele frequency of the effect allele. b are the genetic effects for each marker, and e the residual error. The dimensions of y, X, b and e are dependent upon the number of phenotypes, k, the number of markers, m, and the number of individuals, n. The residuals, e, are a priori assumed to be independently and identically distributed multivariate normal with null mean and covariance matrix $I\sigma_e^2$.

Extensions to summary data

The key parameter of interest in the multiple regression model are the marker effects. These can be obtained by solving an equation system like:

$$b \quad = \left(X'X + I\frac{\sigma_e^2}{\sigma_b^2}\right)^{-1} X'y.$$

To solve this equation system individual level data (genotypes/markers X and phenotypes y) is required. If these are not available, it is possible to reconstruct $X'y$ and $X'X$ from a LD correlation matrix B (from a population matched LD reference panel) and data:

$$\begin{aligned} X'X &= D^{0.5}BD^{0.5}, \\ X'y &= Db_m, \end{aligned}$$

where $D_i = \frac{1}{\sigma_{b_i}^2 + b_i^2/n_i}$ if the markers have been centered to mean 0 or $D_i = n_i$ if the markers have been centered to mean 0 and scaled to unit variance, b_i is the marker effect for the i'th marker, $\sigma_{b_i}^2$ is the variance of the marginal effects from GWAS. $b_m = D^{-1}X'y$ is the vector of marginal marker effects obtained from a standard GWAS. To construct LD correlation matrix, B, firstly LD among markers was computed using squared Pearson's correlation coefficient (BulikSullivan et al. 2015) and estimated as:

$$r_{adj}^2 = r^2 + (1 - r^2)/(N - 2).$$

Estimation of parameters using BLR models

The BLR models use an iterative algorithm, Markov Chain Monte Carlo (MCMC) techniques, to estimate joint marker effects which depends on additional model parameters such as a probability of being causal (π), an overall marker variance (σ_b^2), and residual variance (σ_e^2). The posterior density of the model parameters (b,σ_b^2,σ_e^2) depend on the likelihood of the data given the parameters and a prior probability for the model parameters which is discussed in detail by Rohde et al. (2022).

Ideally the choice of prior for the marker effect should reflect the genetic architecture of the phenotype. Most complex phenotypes and diseases are likely highly polygenic, with hundreds to thousands of causal genetic variants, most frequently of small effect sizes. Thus the prior distribution should account for many small and few large effects. Also, marker effects are a priori assumed to be uncorrelated but markers can be in strong linkage disequilibrium and therefore a high posterior correlation may exists. To accommodate

evolving ideas genetic architectures of phenotypes and diseases, many priors for marker effects have been proposed. Each prior gives rise to a method or family of methods, and two of them are described below:

BayesC

In the BayesC approach the marker effects, b, are a priori assumed to be sampled from a mixture with a point mass at zero and univariate normal distribution conditional on common marker effect variance σ_b^2. This reflect a very common thought that there were not many causal loci. This can be implemented by introducing additional variables δ_i which indicates if the i'th marker has an effect or not. In turn, these variables δ have a prior distribution called Bernoulli with a probability π of being 0. Therefore, the hierarchy of priors is:

$$p\left(b_j|\delta_i, \sigma_{b_i}^2, \pi\right) = \begin{cases} 0 & \text{with probability } \pi, \\ \sim N\left(0, \sigma_{b_i}^2\right) & \text{with probability } 1-\pi \end{cases}$$

$$p\left(\sigma_{b_i}^2|\upsilon_b, S_b^2\right) = S_b^2 \chi_{\upsilon_b}^{-1},$$

where $S_b^2 = \sigma_b^2 \upsilon_b$ with $\sigma_b^2 = \frac{\sigma_g^2}{(1-\pi)2\sum_i p_i(1-p_i)}$ because the variance of a t distribution is $\frac{\upsilon_b}{\upsilon_b - 2}$.

BayesR

In the BayesR (Erbe et al. 2012) approach the marker effects, b, are a priori assumed to be sampled from a mixture with a point mass at zero and univariate normal distributions conditional on common marker effect variance σ_b^2, and variance scaling factors, γ:

$$b_j|\pi, \sigma_b^2 = \begin{cases} 0 & \text{with probability } \pi_1, \\ \sim N(0, \gamma_2 \sigma_b^2) & \text{with probability } \pi_2, \\ \vdots & \\ \sim N(0, \gamma_C \sigma_b^2) & \text{with probability } 1 - \sum_{c=1}^{C-1} \pi_c \end{cases}$$

where $\pi = (\pi_1, \pi_2, \ldots, \pi_C)$ is a vector prior probabilities and $\gamma = (\gamma_1, \gamma_2, \ldots\ldots, \gamma_C)$ is a vector of variance scaling factors for each of C marker variance classes. The γ coefficients are prespecified and constrain how the common marker effect variance σ_b^2 scales within each mixture distribution. Typically $\gamma = (0,0.01,0.1,1.0)$. and $\pi = (0.95,0.02,0.02,0.01)$.

The prior distribution for the marker variance σ_b^2 is assumed to be an inverse Chi-square prior distribution,$\chi^{-1}(S_b, \nu_b)$.

The proportion of markers in each mixture class π follows a Direchlet $(C, c + \alpha)$ distribution, where c is a vector of length C that contains the counts of the number of variants in each variance class and $\alpha = (1,1,1,1)'$.

Using the concept of data augmentation, an indicator variable $d = (d_1, d_2, .., d_{m-1}, d_m)$, is introduced, where d_j indicates whether the jth marker effect is zero or nonzero.

Genome-wide association study (GWAS)

For simulations, we had eight and sixteen simulation scenarios (10 replicates per scenario) for quantitative and binary phenotypes respectively. We performed GWAS by fitting a single

marker linear regression model using the R package "qgg" (Rohde et al. 2020). No co-variates were used in the model. For analysis of the UKB phenotypes, the total population (no missing phenotype) was divided into five replicates of training (80%) and validation (20%) populations. The design for the analysis of the UKB phenotypes is presented in Figure 2. GWAS was performed in the total population and the training population of the five replicates. We performed single marker linear regression using the R package "qgg" (Rohde et al. 2020), and logistic regression analysis using PLINK 1.9 (Purcell et al. 2007) for the quantitative and binary UKB phenotypes respectively. To account for any cryptic relatedness in the data, we used top ten principal components (PCS) along with age, sex and the UKB assessment center as co-variates in the analysis of the UKB imputed data. We computed PCs for WBU from 100K randomly sampled SNPs from the genotyped data after removing SNPs in the autosomal long-range LD regions (Price et al. 2008) with pairwise correlation (r^2)>0.1 in 500Kb region, using PLINK 2.0 (Chang et al. 2015).

Designing genomic region for fine mapping

For the simulated phenotypes, we designed fine-mapping regions based on the number of SNPs (at most 1000 SNPs in length). The regions were designed by defining a window of ~500 SNPs to the left and right of the causal SNPs. The number of the fine-mapping regions depended on the type of simulation scenario. We did not consider any overlaps across the regions. For the UKB phenotypes, we designed the fine-mapping regions based on the physical position around each lead SNP. Significant SNPs (p-value < 5 X 10-8) from GWAS were used as the lead SNPs to design the fine-mapping regions. We defined a genomic region of one mega base pair (1MB) (~1000kb on both sides) of the lead SNP. If the regions overlapped by more than 500kb then the regions were merged.

Methods for fine mapping using summary statistics

We implemented BayesC and BayesR and the following external models: FINEMAP (Benner et al. 2016), SuSIE-RSS (Zou et al. 2022), SuSIE-Inf and FINEMAP-Inf (Cui et al. 2022) for fine mapping.

BLR models

The BLR models BayesC and BayesR, differ based on their assumption of prior variance of the marker effects. Their assumptions have already been discussed in detail above in the section "BLR models". For the simulations, BayesC and BayesR were implemented region-wide and genome-wide, using the R package "qgg" (Rohde et al. 2020). This implementation is illustrated in Figure 1.

To apply these models region-wide, summary data from the GWAS for the SNPs in the fine-mapping regions along with the pair-wise linkage disequilibrium (LD) information among all the SNPs were used. The LD was estimated as the squared Pearson's correlation coefficient adjusted for sample size (Rohde et al. 2021) to avoid the upward bias of Pearson's correlation coefficient, and was computed with the R package "qgg" (Rohde et al. 2020).The region-wide analysis was performed in different ways (the following three options) depending on estimation of different model parameters as part of an iterative estimation procedure (Gibbs sampling technique) from fully conditional posterior distributions. For the first option, the parameter π was treated as random and estimated in each iteration along with the marker variance and the residual variance. For the second option, π was kept constant. For the third option, only the marker variance and residual variance were estimated. Option 1: σ_b^2, σ_g^2, σ_e^2 and π – update Option 2: σ_b^2, σ_g^2 and σ_e^2 – update Option 3: σ_b^2, and σ_e^2 – update

For genome-wide application, we used summary data from the GWAS and sparse LD matrix. We randomly sampled 50,000 out of n=335,532 WBU to estimate LD (sparse LD) for a group of SNPs in a sliding genomic window containing 2000 SNPs, which slid 1 SNP at a time. Due to computational challenge, for genome-wide analysis, only the "Option 1" was used. The PIPs for SNPs obtained from the genome-wide analysis were used further to design credible sets for the fine-mapped regions. A total of 3000 iterations were used in the analysis with the first 500 as burn-in.

External fine mapping tools

SuSIE-RSS model:

The model was applied using the R package susieR (Wang et al. 2020). We provided the summary statistics (beta estimates and standard error), the LD information and number of samples for the fine mapping regions. The residual variance was estimated as suggested by the model because in-sample LD was used. We used ten causal SNPs which is the default number in the R package SusieR. We used default parameters in the functions. No priors for the SNPs were provided.

SuSIE-Inf and FINEMAP-Inf models:

To apply these models, we downloaded python package "run_fine_mapping.py" from the link: https://github.com/FinucaneLab/fine-mapping-inf (Cui et al. 2022). We provided the summary statistics (SNP estimates and standard error) along with LD information and number of samples for the fine mapping regions. The number of causal SNPs was assumed to be ten to be consistent with the default number of causal SNPs in susieR. SuSIE-Inf and FINEMAP-Inf models were applied separately. No variance was shared and no priors for the SNPs were provided.

FINEMAP model:

We downloaded FINEMAP software from the link: http://www.christianbenner.com/finemap_v1.4_x86_64.tgz (v1.4) (Benner et al. 2016). We provided the summary statistics (SNP estimates and standard error) along with minor allele frequency (MAF), LD information, and the number of samples for the fine mapping regions. The number of causal SNPs was assumed to be ten. No priors for the SNPs were provided.

Quality control/convergence for models

The fine mapping tools such as FINEMAP, SuSIE-RSS, SuSIE-Inf and FINEMAP-Inf, explicitly mentioned convergence of the models in the output.

For the BLR models, we estimated the convergence of the key parameters: σ_b^2, σ_g^2, σ_e^2, and π. To assess the convergence, We used the metric "zscore". This involved calculating the difference between the average parameter values taken at the start and end of the iterations. This difference served as our metric to gauge the convergence of the desired parameter. The fine mapping regions with the absolute Value of the metric "zscore", for any of the parameters, greater than three was further investigated by thorough evaluation of the trace plots of the parameters, and scatter plots (SNP effect of the BLR models versus GWAS).

Assessment of fine mapping models in simulations

Efficiency of different models were investigated based on the F_1 score,a harmonic mean of precision and power estimated for the credible sets (CS).

Credible sets for simulations

Credible sets (CSs) help to refine association signals. The CS are defined as the minimum set of SNPs that contains all causal SNPs with probability α. When we assume only one causal SNP, α can be the sum of the PIPs for SNPs in a set. The CS in our study was designed according to Consortium et al. (2012); Schaid et al. (2018). To design the CS of SNPs with coverage probability (cut-off or threshold) of α, firstly SNPs were ranked according to descending order of their PIPs. A vector of cumulative sum of PIPs was created. We added each element of the vector until it crossed a specified coverage probability of α. All the sets exceeding the given threshold of α in the fine mapping regions were refered as the CSs. A CS can contain multiple SNPs if they cross the given PIP threshold. We use a strict coverage probability of 99% for the CSs.

In simulations, with an interest to compare only the core algorithms among different models in our study, we designed the CSs for all the models irrespective of potential of the models to output the CSs. FINEMAP-Inf doesn't give CSs, however we used the PIPs from FINEMAP-Inf to design CSs for the model. In the scenario where multiple SNPs have the same value of PIP, we investigated the list of SNPS, and if one of those SNPs is the simulated causal SNP then we included that SNP in the CS. The same procedure was applied to design CS for all the models. As we used only one causal SNP per fine-mapping regions without considering overlaps across the regions, the concept of one region harboring one causal SNP remained valid and supported our design of the CSs.

F_1 classification score for simulations

We assessed F_1 score for the fine-mapping regions based on the credible sets. All the fine-mapping regions harbored a simulated causal SNP (index SNP). The F_1 score takes a value between 0 and 1. The value close to 1 refers to the capability of a fine-mapping model to better identify true causal SNPs and reduce false positives.

$$F_1 = \frac{2pr}{p + r},$$

where, precision, $p = TP/TP + FP$ and recall $r = TP/TP + FN$.

F_1 score was calculated for each replicate of a simulation scenario. For one simulation replicate, True positive (TP) was the total number of CS (referred to as true positive CS; TP_CS) which contained the index SNP corresponding to that region. False positive (FP) was the total number of CS which crossed the threshold of alpha but did not contain the index SNP (referred to as false positives CS; FP_CS). False negative (FN) was the total number of genomic regions where the cumulative sum of PIPs did not cross the threshold of alpha, and no CS was detected. In addition to this criterion in our study for the FN, we also considered two additional criteria. We denoted "unconverged" fine-mapping regions for any methods as FN. We also considered "TP_CS" which contained more than ten SNPs as FN because large credible sets add little to no information in search of causal variants in fine-mapping procedure. We investigated the number of SNPs in true positive credible sets (TP_CS) to investigate the efficient model and tried to have the least number of SNPs in a CS as possible.

The design of CSs and estimation of F_1 score is represented in Figure 3.

Most efficient model

To investigate the most efficient model, we compared values of F_1 score, across all simulation scenarios (binary and quantitative phenotypes), among all the models. We performed the same procedure for precision and power.

Investigate influence of different factors in simulations

To investigate the influence of each parameters: h^2_{SNP}, π, genetic architectures (GA) and PV on the performance of the models, we performed TukeyHSD test in R.To quantify the factors with the greater influence in the simulations, for each model, We also performed ANOVA on the linear model where the F_1 score was regressed on h^2_{SNP}, π, and GA for the quantitative phenotypes (and PV for the binary phenotypes).

Similarities among model assumption in simulations

We also investigated similarities among the models based on their assumptions of genetic architectures for a complex phenotype. SuSIE, FINEMAP and BayesC assume contribution of sparse genetic variants in the genetic makeup of a complex phenotype. In addition to these sparse genetic variants, SuSIE-Inf and FINEMAP-Inf consider the influence of multiple genetic variants with small effect sizes (infinitesimal models). The BayesR model assume influence of sparse genetic variants with large effect sizes and non-sparse genetic variants with moderate to small effect sizes in the genetic makeup of a complex phenotype. We used total true positive credible sets (TP_CS) determined by each model for only the simulation scenarios for the quantitative phenotype. We investigated the number of overlaps of TP_CS of the BayesC model with SuSIE and FINEMAP, and the overlap of the BayesR with SusIE-Inf and FIENMAP-Inf.

Assessment of fine mapping models in the UKB phenotypes

Only the fine-mapped regions which converged across the models were used for downstream analysis to estimate predictive ability and features of the CSs.

Predictive Ability

For quantitative phenotypes, the predictive ability was determined by estimating the coefficient of determination, (R^2). For binary phenotypes, the predictive ability was determined by estimating Area under the receiver operating characteristic curve (AUC).

Firstly, genomic score (GS) (predicted phenotype) of an individual, also known as a predictive score for a phenotype was calculated for the validation population for each replicate. GS for an individual is the sum of the product of effect alleles weighted by their estimated effect size:

$$GS = \sum_{i=1}^{m} X_i \, \hat{b}_i.$$

where X_i refers to the genotype matrix that contains an allelic counts and $\hat{b}$ is the estimated marker effect for the i-th variant, m is the number of SNPs.

To quantify the accuracy of the GS for real quantitative phenotypes, co-variates adjusted scaled phenotypes for validation population was regressed on the predicted phenotypes. The coefficient of determination, R^2, from the regression was used as a metric to assess the predictive ability of the model. To quantify the accuracy of the GS for real binary phenotypes, AUC (N Wray et al. 2010) was reported:

$$AUC = \frac{1}{n_{d'}}\left(\bar{r}_d - \frac{n_d}{2} - \frac{1}{2}\right).$$

where, $n_{d'}$: number of controls n_d : number of cases $\bar{r}_d$: average rank of the cases.

To observe potential significant difference in estimates of R^2 and AUC (averaged across five replicates) among different methods, we performed TukeyHSD test.

Credible sets for the UKB phenotypes

The credible sets were designed using the algorithm by Consortium et al. (2012). From each converged fine-mapped region, the designed CS contained SNPs whose cumulative sum of PIPs exceeded a threshold of 0.99. For each model, we calculated the total number of CSs determined, median size of the CS (number of SNPs harbored in the CS), and the median value for the average correlations- (r^2) of the absolute correlations among SNPs in the CS. To estimate r^2, we excluded the sets with only one SNP as it is not informative.

Results

Here, we present the results of the application of the BLR models and the state-of-the-art models in fine mapping: FINEMAP, SuSIE-RSS, SuSIE-Inf and FINEMAP-Inf, in fine mapping using the simulations and the UKB phenotypes.

Application in simulations

For simulations, we presented the results of the F_1 score based on the credible sets to show the overall performance of the models across all simulation scenarios in Figure 4. Then to investigate the influence of each parameter considered while designing simulation scenarios, we present the results of the F_1 score in each simulation scenario for the quantitative(Figure S1) and the binary phenotypes (results not shown).

Most efficient model

We applied BayesR and BayesC genome-wide and region-wide, where region-wide analysis was performed using three different options (mentioned in Materials and method section). Among different options applied to simulations, the $bR3$ model (option 3) improved the $F_{1avg.sim}$ score (average across all the simulation scenarios) by 21.64%, 11% and 0.06% relative to the BayesR genome-wide analysis ($bRgw$), and the two other options, $bR1$ and $bR2$. We observed similar results for BayesC. The BayesC region-wide model (option 3, $bC3$) improved the $F_{1avg.sim}$ score by 23.41%, 11.72% and 0.10% relative to $bCgw$, $bC1$, and $bC2$.

Highest $F_{1avg.sim}$ score, averaged across all the twenty-four simulation (binary and quantitative) scenarios was observed for the BayesR region wide model ($bR3$) [$F_{1avg.sim}$score: 0.4] followed by SuSIE-Inf [$F_{1avg.sim}$ score: 0.35] and SuSIE-RSS

[$F_{1avg.sim}$ score: 0.34] (Figure 4). The $bR3$ improved the $F_{1avg.sim}$ score by 27.26%, 26.96%, 18.40%, 15.42%, and 13.32% relative to FINEMAP-Inf, $bC3$, FINEMAP, SUSIE-RSS and SUSIE-Inf. The precision ($Prec_{avg.sim}$) and power ($Pow_{avg.sim}$), averaged across all the simulation scenarios ranged between 0.29 to 0.39, and 0.25 to 0.98 respectively. The $bR3$ model also improved the $Pow_{avg.sim}$ by 58% to 72% relative to other models. However, this model decreased the $Prec_{avg.sim}$ by 26% to 33% relative to other models. Similar patterns were observed when the models were compared only within the quantitative phenotypes and within the binary phenotypes scenarios. For the $bR3$ model, we observed a higher F_1 score of 0.49, averaged across all scenarios for the quantitative phenotypes compared to 0.35 for the binary phenotypes (results not shown).

Here after, we compared only the BayesR model $bR3$ and the BayesC model $bC3$ with the external models.

Influence of parameters in simulations

We investigated the influence of the parameters: h^2_{SNP}, π, GA, and PV used in designing simulated phenotypes, on the efficiency of the BLR and the external models. The box plot of the F_1 scores for the model in different scenarios of the quantitative phenotypes are presented in Figure S1. To achieve this, we made a pairwise comparison of the $F_{1avg.rep}$ score (averaged across the replicates in a scenario) between all the scenarios for both the quantitative and the binary phenotypes. Then we compared specific scenarios to get more information regarding the parameter of interest. Also, we investigated for the parameter with a greater influence in efficiency of the models. To achieve this, for each model, we performed ANOVA on the results where F_1 scores were regressed on h^2_{SNP}, π, and GA (and PV for the binary phenotypes). To discuss our result in detail, we present the results for $bR3$ for the quantitativesimulated phenotypes in Figure S2. This figure shows the difference between the $F_{1avg.rep}$ scores along with the 95% confidence interval from different scenarios. The differences between all the scenarios are significant as none of the intervals harbored a value of zero. Such patterns were observed for all the models while investigating pairwise comparison of the values across all the scenarios. All the models performed the best (highest F_1 score) for the scenario with moderate h^2_{SNP} of 0.3, π of 0.001, and GA_1 for the quantitativephenotypes, and PV of 15% for the binary phenotypes (results not shown).

The largest value for the F-statistic (ratio of the "between-group" variance and "within-group" variance) was observed for the parameter π suggesting greater influence of this parameter in our simulated data (results not shown).

Similarities among methods assumptions in simulations

We investigated similarities among the models based on their assumptions of genetic architectures of a complex phenotype. We observed that at least 50% of the true positive credible sets (TP_CS) was shared among BayesC, SuSIE-RSS and FINEMAP (Figure S3). We observed similar results for the models BayesR, SuSIE-Inf and FINEMAP-Inf. BayesC genome-wide model ($bCgw$) determined the smallest number of total TP_CS summed across all the scenarios followed by FINEMAP-Inf. We observed that $bCgw$ shared 80% of the total TP_CS with SUSIE-RSS, FINEMAP and BayesC region-wide model ($bClw$ or $bC3$). The $bCgw$ shared ~91% of the total TP_CS with $bClw$. Similarly, FINEMAP-Inf shared 80% of the total TP_CS with SuSIE-Inf, $bRgw$, and $bRlw$ of $bR3$. $bRgw$ shared 85.1% of the total TP_CS with $bRlw$ or $bR3$.

Application in the UKB phenotypes

We applied the models in the quantitative and the binary phenotypes in the UKB data. We compared the efficiency on their predictive abilities dividing the main population into five replicates of training and validation populations. We also compared the models based on the number and length of the credible sets along with the median value for the average correlation among SNPs in the CSs. The number of participants in the main population and the replicates for all the phenotypes is described in detail in Table 1.

Predictive ability

The results of predictive ability are presented in Figure 6a and 6b. Prediction $AUC_{avg.bin}$ (averaged across all the binary phenotypes) with BayesR increased by 0.40%, 0.16%, 0.08%, 0.05% compared to SUSIE-RSS, BayesC, FINEMAP-Inf and SuSIE-Inf respectively. We didn't observe any significant differences between the $AUC_{avg.rep}$ (averaged across all the replicates) of models compared pairwise for any binary phenotypes except for HTN. For HTN, BayesR improved the $AUC_{avg.rep}$ significantly compared to SuSIE-RSS. The highest estimate of the $AUC_{avg.rep}$ was observed for T2D followed by HTN for all the models. The lowest estimate of the $AUC_{avg.rep}$ was observed for RA. Prediction $R^2_{avg.qt}$ (averaged across all the quantitative phenotypes) with BayesR decreased by 5.32% and 3.71% compared to SuSIE-Inf and FINEMAP-Inf, whereas increased by 7.93% and 8.3% compared to SuSIE-RSS and BayesC. BayesR model improved the $R^2_{avg.rep}$ (averaged across all the replicates) significantly compared to BayesC model for all the quantitative phenotypes except for WHR. We observed a significant decrease in the $R^2_{avg.rep}$ of BayesC and BayesR relative to SuSIE-Inf and FINEMAP-Inf for the phenotypes BMI, WC, HC and WHR. No significant difference in the $R^2_{avg.rep}$ was observed between BayesR compared to SuSIE-Inf and FINEMAP-Inf for Height, whereas a significant decrease was observed for BayesC compared to these models. We observed significant improvement in the $R^2_{avg.rep}$ of BayesR relative to SuSIE-RSS for Height. All the methods could predict Height better compared to other quantitative phenotypes.

Credible sets

We used the main population to investigate the efficiency of the BLR models and compare their efficiency with the external models comparing features of credible sets (CSs). The efficiency was assessed on the total number of the CSs determined, and the median size of the CS (number of SNPs harbored by the CS), presented in Table 2. We calculated the average correlation for a CS using absolute pair-wise correlations among SNPs in the CS. For comparison, we used the median of the distribution for average correlations of CS within a phenotype. We have presented the summary of the average correlations for the phenotypes in Table 2. The number of fine-mapped regions for the quantitative phenotypes ranged from 181 for WHR to 517 for Height (Table 2). For the binary phenotypes, the number of fine-mapped regions ranged from 6 for RA to 179 for HTN. Fine-mapped regions not converged in any of the models were excluded from all the downstream analysis for all the models. In our study, the highest non-converged regions were observed for RA (50%) followed by PSO (20%). For other phenotypes, the non-converged regions ranged from 0.8% to 7.5%. Within a phenotype, all the models determined almost the same number of the CSs except for FINEMAP-Inf which determined a smaller number of CSs. The BLR models showed the smallest median size for the CSs across all the phenotypes compared to the external fine-mapping models. BayesR showed the smallest median size of CS for seven out of the ten phenotypes in our

study, compared to BayesC. The median value for the average correlations of CS for the BLR models were smaller compared to the external models. However, BayesC showed the largest median for the average correlations of CSs compared to BayesR across all the phenotypes.

Discussion

In our study, we aimed to assess the efficiency of the BLR models, BayesC and BayesR, as a fine mapping tool. We applied these models in simulations and the real UKB data using summary statistics. In simulations, the efficiency was investigated based on the metric,F_1 score. As it was not possible to determine F_1 scores for the UKB phenotypes, we investigated the efficiency of the models based on the predictions of phenotypes and features of credible sets determined. BayesC and BayesR models' efficiency were compared to the state-of-the-art methods such as FINEMAP (Benner et al. 2016), SuSIE-RSS (Zou et al. 2022), SuSIE-Inf and FINEMAP-Inf (Cui et al. 2022). All the models used in our study serve the same purpose of identifying true effects of causal variants. However, they differ in the details in the algorithm and their implementation which applied together can have different impact on the overall performance.

BayesC and BayesR

BayesC and BayesR applied genome-wide and region-wide have the same assumptions of prior variance of marker effects, but they differed in their implementation in our study. To our knowledge this is the first study comparing implementation of BayesC and BayesR in such manner. We implemented the models, genome-wide where the posterior distributions of the model parameters were estimated based on taking SNPs genome-wide whereas region-wide implementation were limited to the fine-mapped regions designed based on the simulated causal SNPs. In simulations, our study showed better performance of both priors when implemented region-wide compared to genome-wide based on the $F_{1avg.sim}$ score for both BayesC and BayesR. However, the genome-wide models showed better $Prec_{avg.sim}$ but less $Pow_{avg.sim}$ (less number of true positive CSs) than region-wide models. High percentage of overlaps of the CSs for the genome-wide models with the region-wide models suggests that there is a potential in the genome-wide models. It would be interesting to investigate further the common CSs determined by the genome-wide and the region-wide models. The following discussion will focus only on the region-wide analysis of BayesC and BayesR and will be termed simply as BayesC and BayesR.

We compared BayesC and BayesR priors in simulations for fine mapping by investigating their precision and power. We applied these priors also in the UKB phenotypes and compared their efficiency based on prediction accuracy, and features of CS. BayesC assumes that the SNP effects follow a two-component normal mixture including zero and non-zero effects, whereas BayesR prior assumes four-components including one zero and three non-zero effects with mixture of normal distributions. BayesR showed significant improvement in prediction accuracy, $Prec_{avg.rep}$ for four out of the five quantitative phenotypes relative to BayesC. Our findings about prediction accuracy are consistent with previous studies (Zhu et al. 2019; Mollandin et al. 2023). Zhu et al. (2019) showed an increase of prediction ability, averaged across various economical phenotypes in cattle, using BayesR compared to BayesC. Mollandin et al. (2023) showed similar results across simulation scenarios for phenotypes with high heritability. Our results suggest that BayesR assumption about genetic architecture suits better for polygenic phenotypes predictions relative to BayesC.

Comparison of the BLR models to external models

To our knowledge this is the first study comparing BayesR model to the state-of-the-art models: FINEMAP, SuSIE-RSS, SuSIE-Inf and FINEMAP-Inf. Across different simulation scenarios, BayesR performed better with higher $F_{1avg.sim}$ score relative to the external models with high power but less precision. The average prediction accuracy ($R^2_{avg.rep}$) for the quantitative phenotypes was significantly lower for BayesR and BayesC compared to SuSIE-Inf and FINEMAP-Inf for the phenotypes: BMI, HC, WC and WHR. For Height though there was a decrease in the $R^2_{avg.rep}$ for BayesR compared to SuSIE-Inf and increase in the value compared to FINEMAP-Inf, however the difference was insignificant. In the fine mapping approach, SuSIE-Inf and FINEMAP-Inf assumes influence of majority of SNPs with infinitesimal effect sizes in addition to the few SNPs with large effect sizes (Cui et al. 2022). We observed highest estimates for the predictive ability of the infinitesimal models, which might have been because we used SNP effects from both sparse and infinitesimal components for SuSIE-Inf and FINEMAP-Inf for predictions, unlike in the study (Cui et al. 2022) itself where only sparse components were used to compare prediction accuracies between SuSIE, FINEMAP, SuSIE-Inf and FINEMAP-Inf. Since infinitesimal components do not neglect any SNP effects, this may also explain high prediction accuracy for the models including infinitesimal effects. Our results showed that the performance of BayesR is closer to the infinitesimal models.

Campos et al. (2023) compared the performance of BayesC to various methods including SuSIE and FINEMAP in fine mapping, where BayesC performed similar to SuSIE but better than FINEMAP in power and false discovery rate determination, for different simulation scenarios. In our study, BayesC improved $Pow_{avg.sim}$ by 2.27% relative to FINEMAP whereas the power decreased by 4.3% relative to SuSIE-RSS. This difference in results might be due to differences in implementation of these models as this study applied the models in whole-genome scale using local regression approach where we applied the models only in specific regions defined by simulated causal SNPs, that not necessarily included whole genome. Our study compared SuSIE-RSS (which is an extension of SuSIE that uses summary statistics) and BayesC, FINEMAP using summary statistics with in-sample LD among other models, whereas this study used individual levels data for BayesC and SuSIE, and summary statistics with in-sample LD for FINEMAP.

In our study, BayesC and BayesR showed a similar number of credible sets as the external models for the UKB phenotypes. The smallest median size of the credible set(CS) and the smallest median value for the average correlations in CS was observed for BayesC and BayesR. We constructed credible sets for all the models based on the algorithm from Consortium et al. (2012), where we applied the cut-off thresholds of 0.99 for a set to be a CS. We determined only one credible set per fine mapped region for the threshold. We are aware that FINEMAP, SuSIE-RSS and SuSIE-Inf also determine CS where multiple CS can be determined based on multiple causal variants in the fine mapped region. Such procedure of determining CSs might alter the features of the CSs. However, the main objective of our study was to compare the efficiency of the algorithms of these models. Introducing a comparison based on the CSs they determine would introduce additional complexity and divert us far from our objective. Hence, we determined CSs using the same algorithm for all the models. However, it would be interesting to determine CSs similar to those of the external models (assuming multiple causal genetic variants) and evaluate their efficiency.

Influence of parameters in simulations

In our study, we observed a significant effect between components of h^2_{SNP}: 30% and 10%), π (0.001 and 0.1) and GA (GA_1 and GA_2), and PV (5% and 15%). The causal SNPs were sampled randomly genome wide. The region for fine mapping was designed around these causal SNPs. The pairwise comparison of $F_{1avg.rep}$ score, within a scenario, among different simulation scenarios for each model showed significant differences among scenarios and significant contribution of each parameter. However, the large value of the F-statistic obtained from ANOVA on the results of the regression ($F_1 \sim h^2_{SNP} + \pi + GA$) was seen for the parameter π suggesting greater influence of this parameter in performance of the model. In our simulation, a smaller number of causal SNPs for a given genetic variance would be sampled from a larger marker effect variance compared to a higher number of causal SNPs. This large effect SNPs must have high PIPs such that the credible sets determined by the models harbored the true causal SNPs. The F_1 score was based on the detection of a true simulated causal variant in a credible set. In addition to the threshold for a cumulative sum of PIPs [0.99] that a set needs to cross to be a credible set, we also set a limit on the size of CS (not more than 10). The main motive of the CS is to refine the resolution of the fine mapping region and a CS with large number of SNPs even if it harbored a true causal variant would not be informative. We used a strict cut-off threshold of 0.99 for cumulative sum of PIPs and maximum size of 10 for CS. The results might differ with lenient thresholds for the cumulative sum, and size of CS. We have considered only common variance (MAF >= 0.01) for our study to construct polygenic phenotypes. It will also be interesting to introduce the contribution of rare variants for simulations to investigate the performance of models. As per our expectation, all the models performed significantly better (high $F_{1avg.rep}$ score) for the phenotypes with moderate h^2_{SNP} [30%], a smaller number of causal SNPs [π: 0.001], and the phenotypes simulated with few SNPs with large effect size [GA_1] for the quantitative phenotypes, and the worst performance was observed for the phenotypes with low h^2_{SNP} [10%] and a larger number of causal SNPs [π: 0.01].

F_1 classification score – power and precision of the BLR models

F1 score is a harmonic mean of precision and power/recall and is a well-known performance metric used for model comparison especially under class imbalance. It penalizes the performance even when only one of precision or power is low. In our study, both precision and power are given equal importance for the performance of a model. In our study, we observed higher $Pow_{avg.sim}$ for BayesR compared to other models.The power in our study, refers to the ability of a model to determine CSs that harbor true simulated causal SNP. Along with the threshold, size of CS is also used. Highest power of BayesR referred to the scenario where majority of CSs obtained from BayesR had small size CSs. We used in-sample LD and this might not always be the case while using external summary statistics as also mentioned by Campos et al. (2023). Hence, the power may decrease while using an external reference LD panel. We observed low $Prec_{avg.sim}$ of BayesR. This referred to as substantial amount of CSs were false positives. The range of $Prec_{avg.sim}$ across all the models is not vast suggesting that all the models showed similar performance for precision.

The UKB phenotypes, accuracy and fine mapping, credible sets

The predictive accuracies for the phenotypes determined in our study are smaller compared to other studies. $R^2_{avg.rep}$ for BMI, Height, HC, WHR, and $AUC_{avg.rep}$ for T2D for the UKB data presented by Lloyd-Jones et al. (2019) using SBayesR and around 1.1 million SNPs

were larger compared to the values estimated using BayesR in our study. In our study, we used imputed SNPs (around 6.6 million), but our accuracies were limited only to the fine mapped regions. For polygenic phenotypes for example in Height and BMI, Wainschtein et al. (2022) suggested enrichment of heritability from rare genetic variants (MAF < 0.01). In our study, we discarded rare SNPs with MAF < 0.01 and focused only on common SNP effects,. In our study, non-converged regions for a model were excluded from analysis for prediction accuracies (also credible sets) which might also have impacted the estimated accuracies.

Conclusions

We observed that the performance of the BLR models was comparable to the state-of-the-art external models. The performance of BayesR prior was closely aligned with SuSIE-Inf and FINEMAP-Inf models. Results from both simulations and application of the models in the UKB phenotypes suggest that the BLR models are efficient fine mapping tools.

Data availability statement

The genetic and phenotypic data utilized in our study were obtained from the UK Biobank Resource (ID 96479).

Ethics statement

Human studies in the UK Biobank project have received approval from the Ethics and Governance Framework (EGF), which ensures data and sample usage adheres to scientific and ethical standards. The consent to participation will apply throughout the lifetime of the UK Biobank, unless participants withdraw, and involves the collection and storage of biological samples (blood, saliva, urine) and electronic health records (GP, hospitals, dental, prescriptions). Individual data is anonymized, with each research project receiving its own anonymized dataset. The ethics committee waived the need for written informed consent.

Funding

Our project was funded by Novo Nordisk Foundation through the drug discovery platform, Open Discovery Innovation Network (ODIN) under grant number “NNF20SA0061466”. This funding aims to foster collaboration between universities and companies promoting long-term benefits of innovation.

Author Approvals

All authors have seen and approved the manuscript, and it hasn’t been accepted or published elsewhere.

Competing interests

There are no competing interests.

References

Benner, Christian, Chris C. A. Spencer, Aki S. Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. 2016. "FINEMAP: Efficient Variable Selection Using Summary Data from Genome-Wide Association Studies." *Bioinformatics* 32 (10): 1493–1501. https://doi.org/10.1093/bioinformatics/btw018.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2018. "The UK Biobank Resource with Deep Phenotyping and Genomic Data." *Nature* 562 (7726): 203–9. https://doi.org/10.1038/s41586-018-0579-z.

Campos, Gustavo de los, Alexander Grueneberg, Scott Funkhouser, Paulino Perez-Rodriguez, and Anirban Samaddar. 2023. "Fine Mapping and Accurate Prediction of Complex Traits Using Bayesian Variable Selection Models Applied to Biobank-Size Data." *European Journal of Human Genetics* 31: 313–20. https://doi.org/10.1038/s41431-022-01135-5.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (1): 7. https://doi.org/10.1186/s13742-015-0047-8.

Consortium, The Wellcome Trust Case Control, Julian B Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, et al. 2012. "Bayesian Refinement of Association Signals for 14 Loci in 3 Common Diseases." *Nat Genet* 44 (12): 1294–1301. https://doi.org/10.1038/ng.2435.

Cui, Ran, Roy A Elzur, Masahiro Kanai, Jacob C Ulirsch, Omer Weissbrod, Mark J Daly, Benjamin M Neale, Zhou Fan, Hilary K Finucane, et al. 2022. "Improving Fine-Mapping by Modeling Infinitesimal Effects." *bioRxiv*. https://doi.org/10.1101/2022.10.21.513123.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. "Improving Accuracy of Genomic Predictions Within and Between Dairy Cattle Breeds with Imputed High-Density Single Nucleotide Polymorphism Panels." *J. Dairy Sci.* 95 (7): 4114–29. https://doi.org/http://dx.doi.org/10.3168/jds.2011-5019.

Habier, David, Rohan L Fernando, Kadir Kizilkaya, and Dorian J Garrick. 2011. "Extension of the Bayesian Alphabet for Genomic Selection." *BMC Bioinformatics* 12: 186. http://www.biomedcentral.com/1471-2105/12/186.

Hutchinson, Anna, Jennifer Asimit, and Chris Wallace. 2020b. "Fine-Mapping Genetic Associations." *Human Molecular Genetics* 29 (R1): R82–89. https://doi.org/10.1093/hmg/ddaa148.

Hutchinson, A., H. Watson, and C. Wallace. 2020a. "Improving the Coverage of Credible Sets in Bayesian Genetic Fine-Mapping." *PLoS Comput Biol* 16 (4): e1007829. https://doi.org/10.1371/journal.pcbi.1007829.

Lloyd-Jones, Luke R., Jian Zeng, Julia Sidorenko, Loc Yengo, Gerhard Moser, Kathryn E. Kemper, Huanwei Wang, et al. 2019. "Improved Polygenic Prediction by Bayesian Multiple Regression on Summary Statistics." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-12653-0.

Marees, Andries T., Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M. Derks. 2018. "A Tutorial on Conducting Genome-Wide Association Studies: Quality Control and Statistical Analysis." *International Journal of Methods in Psychiatric Research* 27: 1–10. https://doi.org/10.1002/mpr.1608.

Mollandin, Fanny, Andrea Rau, and Pascal Croiseau. 2023. "An Evaluation of the Predictive Performance and Mapping Power of the BayesR Model for Genomic Prediction." *G3*. https://doi.org/10.1093/g3journal/jkab225.

Moser, Gerhard, Sang Hong Lee, Ben J. Hayes, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. 2015. "Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model." Edited by Chris Haley. *PLOS Genetics* 11 (4): e1004969. https://doi.org/10.1371/journal.pgen.1004969.

Price, Alkes L., Michael E. Weale, Nick Patterson, Simon R. Myers, Anna C. Need, Kevin V. Shianna, Dongliang Ge, et al. 2008. "Long-Range LD Can Confound Genome Scans in Admixed Populations." *The American Journal of Human Genetics* 83 (1): 127–47.

Privé, Florian, Julyan Arbel, and Bjarni J Vilhjálmsson. 2020. "LDpred2: Better, Faster, Stronger." Edited by Russell Schwartz. *Bioinformatics* 36 (22-23): 5424–31. https://doi.org/10.1093/bioinformatics/btaa1029.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Luciah Thomas, Manuel AR Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81 (3): 559–75.

Rohde, Palle Duun, Mette Nyegaard, Mads Kjolby, and Peter Sørensen. 2021. "Multi-Trait Genomic Risk Stratification for Type 2 Diabetes." *Frontiers in Medicine*. https://doi.org/10.3389/fmed.2021.711208.

Rohde, Palle Duun, Izel Fourie Sørensen, and Peter Sørensen. 2020. "Qgg: An r Package for Large-Scale Quantitative Genetic Analyses." *Bioinformatics* 36 (April): 2614–15. https://doi.org/10.1093/bioinformatics/btz955.

Rohde, Palle Duun, Izel Fourie Sørensen, and Peter Sørensen. 2022. "Expanded Utility of the r Package Qgg with Applications Within Genomic Medicine." *bioRxiv*. https://doi.org/10.1101/2022.09.03.506466v1.

Schaid, Daniel J., Wenan Chen, and Nicholas B. Larson. 2018. "From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping." *Nature Reviews Genetics* 19 (8): 491–504. https://doi.org/10.1038/s41576-018-0016-z.

Wainschtein, Pierrick, Deepti Jain, Zhili Zheng, TOPMed Anthropometry Working Group, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, L. Adrienne Cupples, Aladdin H. Shadyab, et al. 2022. "Assessing the Contribution of Rare Variants to Complex Trait Heritability from Whole-Genome Sequence Data." *Nature Genetics* 54: 263–73. https://doi.org/https://doi.org/10.1038/s41588-021-00997-7.

Wang, Gao, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. 2020. "A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping." *J. R. Statist. Soc. B* 82 (5): 1273–1300.

Wray, Naomi R., Sang Hong Lee, Divya Mehta, Anna A. E. Vinkhuyzen, Frank Dudbridge, and Christel M. Middeldorp. 2014. "Research Review: Polygenic Methods and Their Application to Psychiatric Traits." *Journal of Child Psychology and Psychiatry* 55 (10): 1068–87. https://doi.org/10.1111/jcpp.12295.

Wray, NR, J Yang, ME Goddard, and PM Visscher. 2010. "The Genetic Interpretation of Area Under the ROC Curve in Genomic Profiling." *PLoS Genet* 6 (2): e1000864. https://doi.org/10.1371/journal.pgen.1000864.

Zhang, Qianqian, Florian Privé, Bjarni Vilhjálmsson, and Doug Speed. 2021. "Improved Genetic Prediction of Complex Traits from Individual-Level Data or Summary Statistics." *Nature Communications* 12 (1). https://doi.org/10.1038/s41467-021-24485-y.

Zhu, B, P Guo, Z Wang, W Zhang, Y Chen, et al. 2019. "Accuracies of Genomic Prediction for Twenty Economically Important Traits in Chinese Simmental Beef Cattle." *Anim Genet* 50: 634–43. https://doi.org/10.1111/age.12853.

Zou, Y, P Carbonetto, G Wang, and M Stephens. 2022. "Fine-Mapping from Summary Data with the Sum of Single Effects Model." *PLoS Genet* 18 (7): e1010299. https://doi.org/10.1371/journal.pgen.1010299.

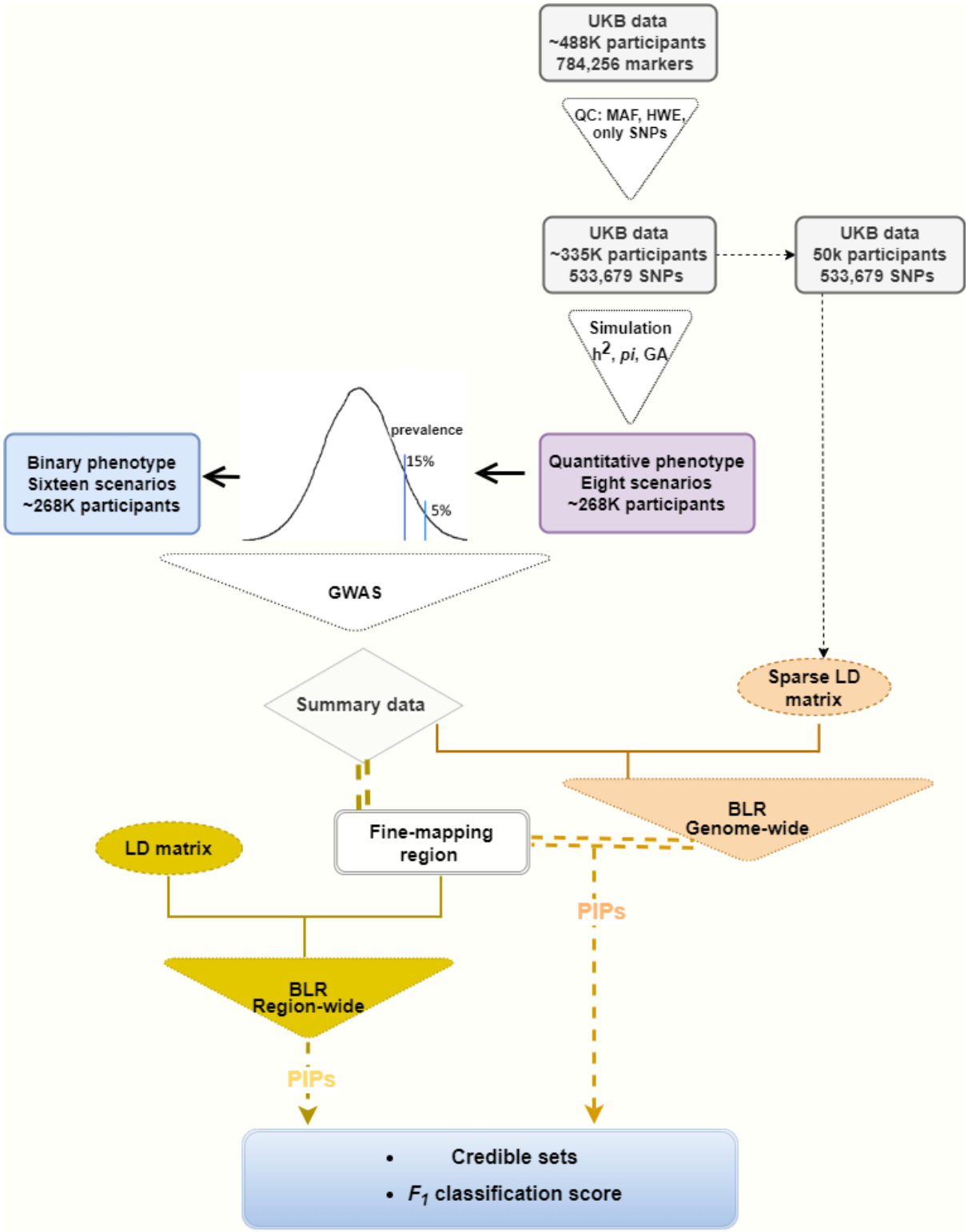


Figure 1. Flowchart illustrating the design of the simulation scenarios for both quantitative and the phenotypes, followed by fine mapping using Bayesian Linear Regression models. The BLR models were implemented in different ways, and the resulting posterior inclusion probability (PIPs) for SNPs were used to estimate the F_1 classification score based on the credible sets.

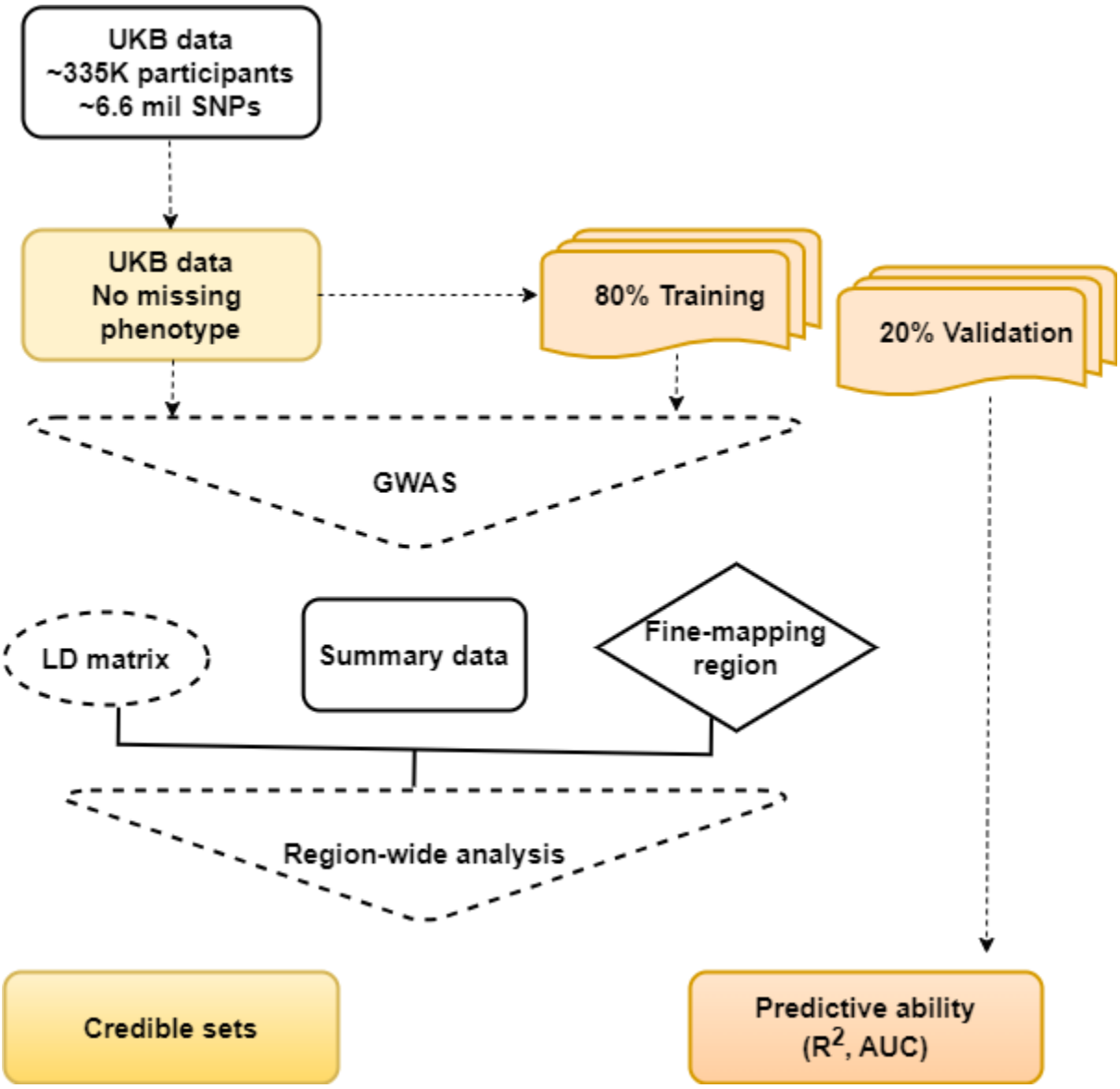
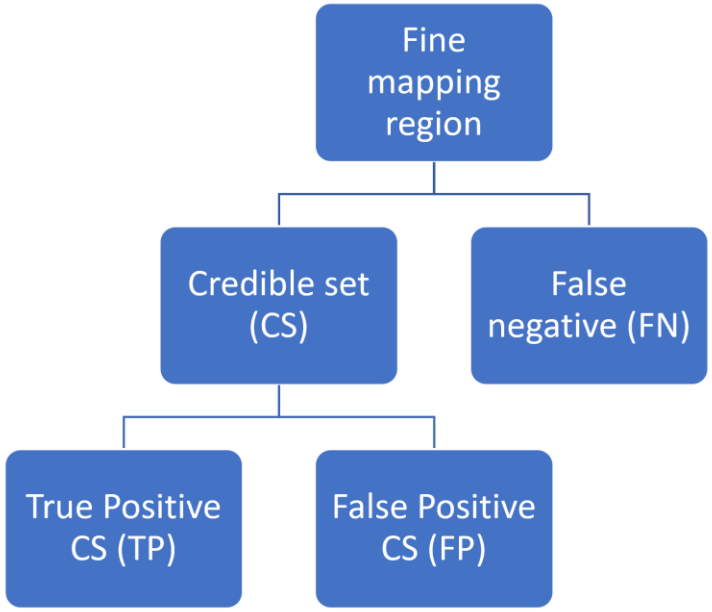


Figure 2. Flowchart illustrating the design of populations for the analysis of the UK Biobank phenotypes to determine the predictive abilities and features of credible sets across different models.



1. **Design Credible set (CS)**:
CS: cumulative sum of PIPs >= 0.99
FN: cumulative sum of PIPs < 0.99
+ "non converged " + "> 10 SNPs"

2. **Identify true and false positives**
True positive CS: harbors the index SNP
False positive CS: doesn't harbor the index SNP

3. **Estimate F_1 classification score**:
Precision: $p = TP/(TP + FP)$
Power: $r = TP/(TP + FN)$

$$\mathbf{F_1\ classification\ score} = \frac{\boldsymbol{2pr}}{\mathbf{p} + \boldsymbol{r}}$$

Figure 3. Design of credible sets with a 0.99 threshold for the cumulative sum of Posterior Inclusion Probabilities (PIPs), and estimation of the F_1 classification score based on the credible sets.

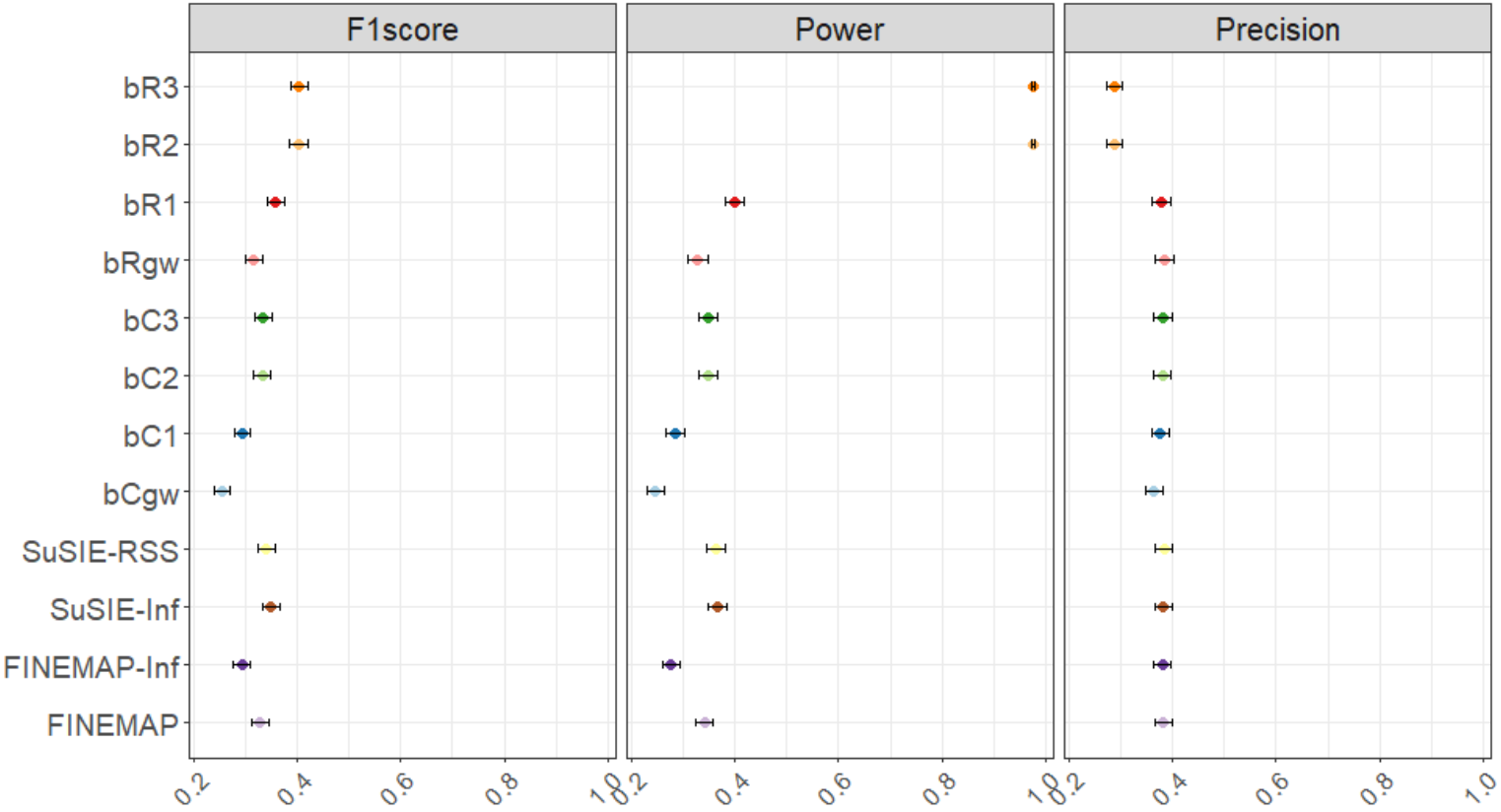


Figure 4. F_1 classification score (F1score), power and precision, averaged across all twenty-four simulation scenarios, for the BLR fine mapping models: BayesR region-wide models (bR3, bR2 and bR1), BayesR genome-wide model (bRgw), BayesC region-wide models (bC3, bC2 and bC1), BayesC genome-wide model (bCgw) and external models. The black solid line represents standard error for the average estimate.

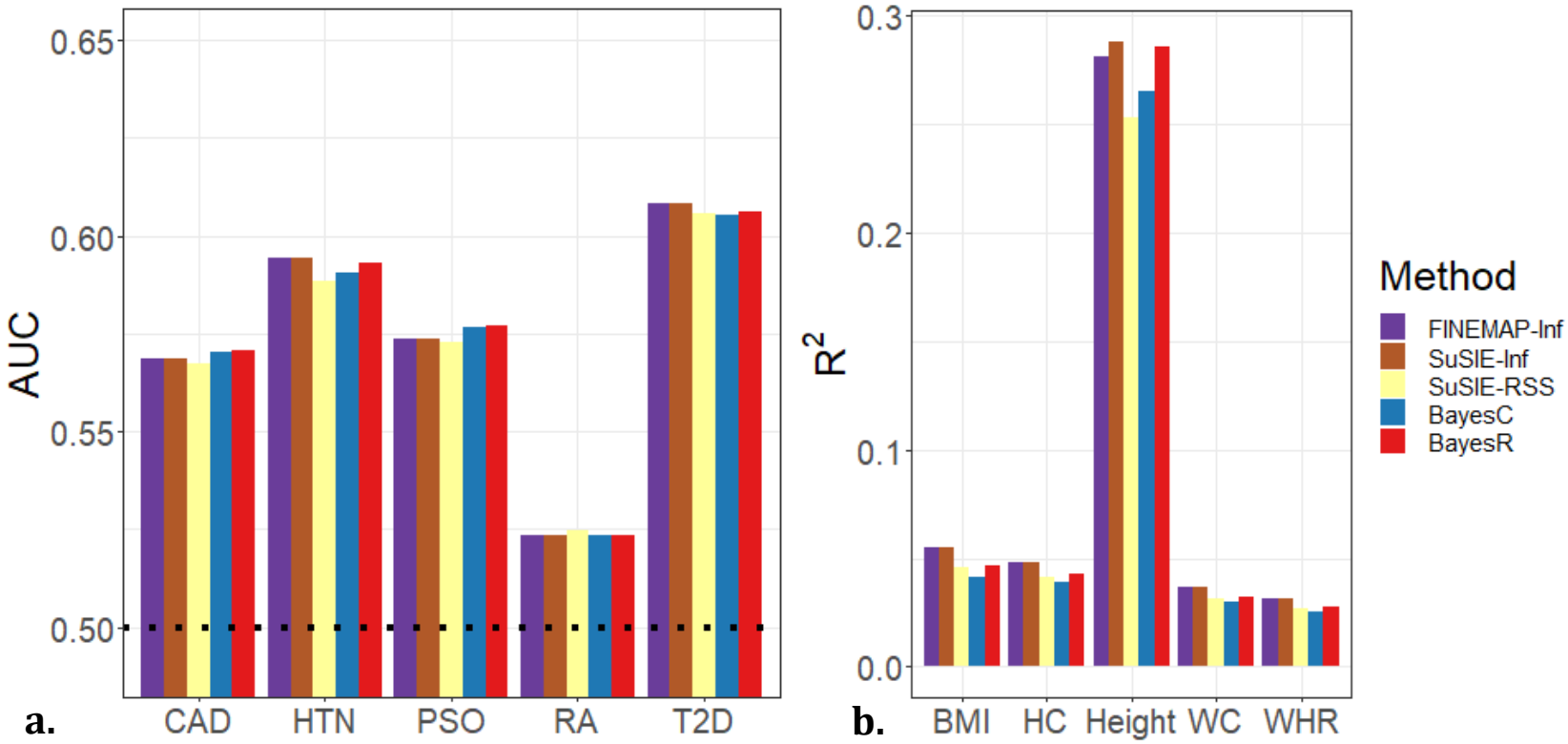


Figure 6. Prediction accuracies estimated from fine mapped regions. **a.** Bar plot of prediction accuracy, represented by the Area under the Curve (AUC), averaged across five replicates for the UKB binary phenotypes: coronary artery disease (CAD), hypertension (HTN), psoriasis (PSO), rheumatoid arthritis (RA), and type 2 diabetes (T2D). **b.** Bar plot of prediction accuracy, represented by the coefficient of determination (R^2)**,** averaged across five replicates for the UKB quantitative phenotypes: body mass index (BMI), hip circumference (HC), standing height (Height), waist circumference (WC), and waist-hip ratio (WHR). The models used in the fine mapping can be identified by the colors in the legend associated with each model.

Tabel 1a. Details of the data fields along with the total number of non-missing samples, age (mean and standard deviation), number of females (n_female) and average value for the UKB quantitative phenotypes.

UKB Quantitative phenotypes	**UKB data field**	**Total**	**Age (Mean [sd])**	**Sex (n_female)**	**Phenotype (Mean [sd])**
Body Mass Index (BMI)	p21001	334464	56.87 [7.98]	179309	27.4 [4.76]
Hip Cirumference (HC)	p49	334949	56.87 [7.98]	179517	103.44 [9.15]
Standing Height (Height)	p50	334828	56.87 [7.98]	179492	168.86 [9.25]
Waist Circumference (WC)	p48	334983	56.87 [7.98]	179532	90.37 [13.49]
Waist-Hip Ratio (WHR)	NA*	334917	56.87 [7.98]	179503	0.87 [0.09]

*NA because the phenotype was calculated in our study.

Tabel 1b. Details of the ICD10 and self-reported code used for diagnosis of cases for the UKB binary phenotypes, total number of cases, controls along with the distribution of age (mean and standard deviation) and number of females (n_female) within cases and controls.

UKB Binary phenotypes	**Definition of cases**		**Cases**	**Controls**	**Age (Mean [sd])**		**Sex (n_female)**	
	ICD10 code [p41270]	**Self-reported code**			**Cases**	**Controls**	**Cases**	**Controls**
Coronary Artery Disease (CAD)	I21; I22; I23; I24; I25	1075	34726	300806	61.24 [6.38]	56.37 [7.99]	10845	168989
Hypertension (HTN)	I10	1065	129580	205952	59.75 [6.98]	55.07 [8.04]	60859	118975
Psorias (PSO)	L40	1453	6628	328904	57.15 [7.88]	56.87 [7.98]	3090	176744
Rheumatoid Arthritis (RA)	M06	1464	7955	327577	59.6 [7.04]	56.81 [7.99]	5251	174583
Type 2 Diabetes (T2D)	E11	1220;1223	25828	309704	60.11 [6.9]	56.6 [8.01]	10072	169762

Tabel 2. Total number of fine mapped regions and non-converged regions (in percentage) for the UKB phenotypes along with the total number of credible sets (CSs), median size of the CSs, and median of the average correlations (r^2) of the CSs for all the models.

UKB Phenotypes	Total regions	Non-converged regions (%)	FINEMAP			FINEMAP-Inf			SUSIE-RSS			SUSIE-Inf			BayesC			BayesR		
			Credible sets	Median size	Median r^2	Credible sets	Median size	Median r^2	Creible sets	Median size	Median r^2	Credible sets	Median size	Median r^2	Credible sets	Median size	Median r^2	Credible sets	Median size	Median r^2
Body Mass Index (BMI)	293	2.39	286	7	0.46	90	13.5	0.58	286	8	0.43	286	56.5	0.18	286	3	0.39	286	2	0.13
Hip Cirumference (HC)	278	2.52	271	7	0.47	103	12	0.56	271	8	0.44	271	37	0.20	270	3	0.47	271	2	0.16
Standing Height (Height)	517	7.54	479	2	0.30	311	3	0.39	479	2	0.30	479	5	0.25	478	1	0.25	479	1	0.04
Waist Circumference (WC)	240	0.83	238	9	0.50	82	16	0.59	238	9	0.44	238	45.5	0.22	237	5	0.47	238	2	0.27
Waist-Hip Ratio (WHR)	181	2.76	176	8.5	0.58	80	13.5	0.62	176	9	0.48	176	27	0.28	176	4	0.50	176	2	0.30
Coronary Artery Disease (CAD)	43	2.33	42	9	0.65	31	11	0.79	42	8	0.54	42	16	0.64	42	2.5	0.37	42	2.5	0.37
Hypertension (HPN)	179	3.91	172	7	0.54	105	7	0.59	172	7	0.46	172	12.5	0.43	172	2	0.39	172	2	0.15
Psoriasis (PSO)	15	20.00	15	12	0.69	14	13	0.67	15	9.5	0.64	15	11.5	0.67	15	2.5	0.43	15	2	0.11

Rheumatoid Arthritis (RA)	6	50.00	3	2	0.957	3	2	0.957	3	2	0.957	3	2	0.957	3	2	0.492	3	2	0.776
Type 2 Diabetes (T2D)	69	2.90	67	11	0.72	52	14	0.69	67	9	0.60	67	18	0.55	67	3	0.44	67	2	0.28